Defining Data Science and Data Scientist

by

Dana M. Dedge Parks

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctorate of Business Administration
Muma College of Business
University of South Florida

Major Professors: Richard Plank, PhD and Richard Will, PhD

Date of Approval:
10/12/17

Keywords: data analyst, data driven decisions, business analysis, descriptive analytics

Dedication

To my mother who told me to remove the words "I can't" from my vocabulary when I was 4 years old. Thank you, Mom!

To my father who taught me how to simplify a complex problem into small chunks in order to achieve the goal. Thank you Dad!

To my husband, John Parks; my rock and forever best friend. He is a constant believer in my dreams.

To our daughters, Tristen, Carroll and Korri; it is never too late to achieve your dreams.

Acknowledgements

This paper would not have been possible without the thoughtful guidance from Dr. Richard Plank and Dr. Richard Will. Their patience and unending support are extremely appreciated during this twelve month journey.

Wayne Willoughby, deserves a special thank you for being a constant advocate and supporter for my continued focus on the topic of data science. Your perspective, understanding of the subject and agreement that data science is the future have been appreciated.

Thank you to Dr. Grandon Gill and Dr. Moez Limayen, whose leadership and vision made this academic goal possible for so many of us. Thank you for your passion for educating both practitioners and academics.

Table of Contents

Abstract

The world's data sets are growing exponentially every day due to the large number of devices generating data residue across the multitude of global data centers. What to do with the massive data stores, how to manage them and defining who are performing these tasks has not been adequately defined and agreed upon by academics and practitioners. Data science is a cross disciplinary, amalgam of skills, techniques and tools which allow business organizations to identify trends and build assumptions which lead to key decisions. It is in an evolutionary state as new technologies with capabilities are still being developed and deployed. The data science tasks and the data scientist skills needed in order to be successful with the analytics across the data stores are defined in this document. The research conducted across twenty-two academic articles, one book, eleven interviews and seventy-eight surveys are combined to articulate the convergence on the terms data science. In addition, the research identified that there are five key skill categories (themes) which have fifty-five competencies that are used globally by data scientists to successfully perform the art and science activities of data science.

Unspecified portions of statistics, technology programming, development of models and calculations are combined to determine outcomes which lead global organizations to make strategic decisions every day.

This research is intended to provide a constructive summary about the topics data science and data scientist in order to spark the dialogue for us to formally finalize the definitions and ultimately change the world by establishing set guidelines on how data science is performed and measured.

Summary

The world will make decisions by either guessing or using their gut.

They will be lucky or wrong. - Suhail Doshi (CEO Mixpanel)

Massive data centers and data stores are used to securely archive today's data and digital footprint that every computer, cell phone, laptop or tablet user leaves behind as they perform daily functions on devices. The science and management disciplines required to develop data analytics capabilities within an organization is a developing field. Developing an organization's cultural mindset to be a productive functioning unit capable of leveraging data requires a fundamental understanding of data science - *what it is and what it is not.*

The terms data science and data scientist are not formally defined and adopted by academics and practitioners. There are misconceptions and misunderstandings about what constitutes data science and data scientist. Some industries have incorporated a basic level of data science into their practice. In other fields, it is a completely new topic with neither integration nor adoption into the business.

Data science is a methodology by which a data scientist takes unspecified portions of statistics, scientific rigor and systemic capabilities to ensure that an answer to a data question is accurate. During one of the greatest periods in the information age, the challenge is to define roles for data experts and develop strategies for spreading strong data solutions across the business industry.

Maytal Saar-Tsechansky (2015) frames the relevance of data science as existing "at the core of a host of ongoing business transformations, and disruptive technologies. The application of data

science methods to new and old business problems presents a wealth of research opportunities that the information systems (IS) data science community is uniquely positioned to focus their efforts."

The goal of this document is to build on the baseline definition and cornerstone foundations of data science and the role of the data scientist. Both practitioners and scholars work to build theory, implementations, and to develop a data focused culture as a business organization matures.

The results gathered from a literary review, interviews, and surveys are utilized to define the terms data science and data scientist in order to establish a foundation that will chronicle the skills, tasks, and types of systems needed in order to capitalize on data science. Per Thomas Davenport in Big Data @ Work, "big data refers to data that is too big to fit on a single server, too unstructured to fit in a row/column database or too continuously flowing to fit into a static data warehouse." Data science deals both with big data and small datasets.

This document is organized in the following manner:

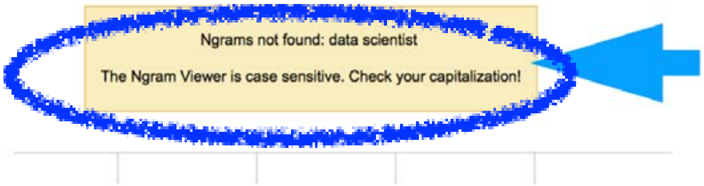
- Foundation of Research - describes the approach taken with the research
    - Literature Review - summarizes of articles used to develop the baseline of data science and data scientist
    - Interviews - summarizes the demographics of interviewees
    - Surveys - summarizes the demographics of the respondents to the surveys
- Key Findings Summary - provides a short synopsis of the results of the research performed across the literature review, interviews and surveys.
    - includes a table of skills and competencies collected directly from the analysis

- Defining Data Science - provides supporting evidence from the literature review, interviews and surveys (in that order)
    - documents the similarities and differences between business and the academic community
- Defining Data Scientist - provides supporting evidence from the literature review, interviews and surveys (in that order)
    - documents the similarities and differences between business and the academic community
- Related Skills & Topics - documents specific details on
    - Machine Learning
    - Data Management
    - Strategic Decisions
    - Data Challenges
- Conclusion - summarizes the findings and discusses next steps

Foundation of Research

Summary of Literature Review

A detailed review of the published literature on big data, data science and the data scientist has been performed. There is very little written on the topics as they relate to the business industry. This Google NGRAM result visually displays the progression across the past 50 years of where data science, data trends, big data, data tools and data strategy terms have been used in books written in the English language. Note that the term data scientist has not appeared in English

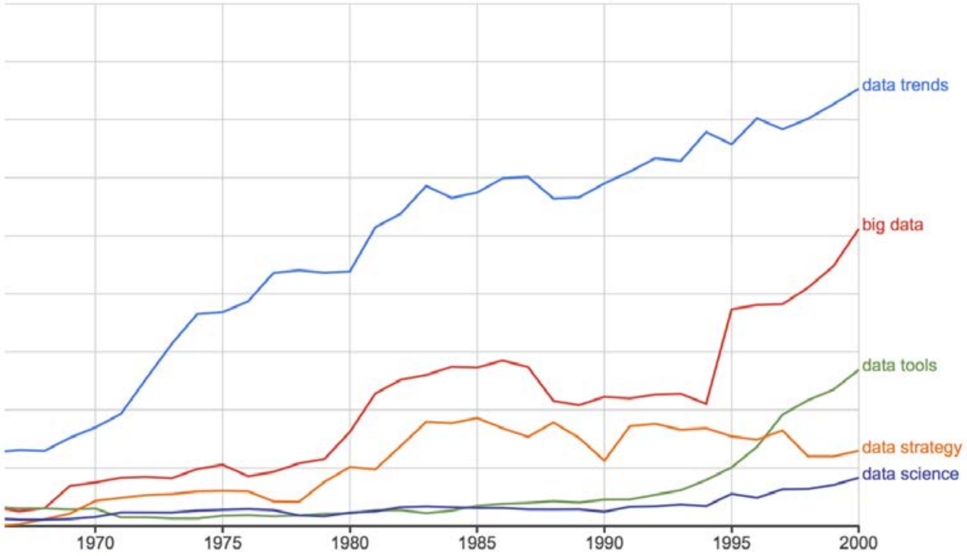


Ngrams not found: data scientist

The Ngram Viewer is case sensitive. Check your capitalization!

NOTE: Data Scientist Not Published in any books since 1800 - 2008

EXPANDED VIEW OF NGRAM



language texts ever between 1800 - 2008. (Full view is in Appendix A).

The literature research for this dissertation was performed by leveraging Google Scholar and the University of South Florida online libraries to provide a baseline understanding and definition for the terms data science and data scientist. It is difficult to research these two topics without touching on the data driven decision making culture and the ways that organizations are adopting data science and their master technicians who are called data analysts, business analysts, statisticians, or data scientists.

As a result, twenty-two articles and one book by academic authors were identified that could be leveraged to begin the aggregation of consensus and opposing views on the actual terms. What appears to be lacking is a definition or agreement across academics on what data science means, how to perform the tasks, the skills the technicians need, the systems architecture, and tools required for the job. The table on the next page shows the publications and year of each article.

ROW #	YEAR	JOURNAL / SOURCE	TOTAL OF ARTICLES
1	2014	ASA DATA SCIENCE JOURNAL	2
2	2001	CALIFORNIA MANAGEMENT REVIEW	1
3	2015	COMMUNICATIONS OF THE ACM	1
4	2012	HARVARD BUSINESS REVIEW	1
5	2015	INTERNAL AUDITOR	1
6	2015	JOURNAL OF ACADEMIC LIBRARIANSHIP	1
7	2016	JOURNAL OF CHAOS, SOLITONS & FRACTALS	1
8	2016	JOURNAL OF INFORMATION SYSTEMS	1
9	2015	JOURNAL OF INFORMATION SYSTEMS EDUCATION	1
10	2012	JOURNAL OF MANAGEMENT DECISION	1
11	2015	JOURNAL OF SOCIAL AND BEHAVIORAL SCIENCES	1
12	2016	MCKINSEY & COMPANY	1
13	2013	MCKINSEY QUARTERLY	1
14	2015	MIS QUARTERLY	1
15	2014	NOTICES OF AMERICAN MATHEMATICAL SOCIETY	1
16	2016	PEOPLE STRATEGY JOURNAL	1
17	2013	PHYSICIAN EDUCATION JOURNAL	1
18	2015	ROYAL STATISTICAL SOCIETY	1
19	2014	UDACITY	1
20	2013	US DEPARTMENT OF LABOR	1
21	2014	WILEY PUBLICATIONS, INC	1

YEAR PUBLISHED	BOOK TITLE	AUTHOR
2014	BIG DATA @ WORK	THOMAS DAVENPORT

Methodology

Summary of Interview Technique

A total of eleven individuals were interviewed. Their interviews were recorded and transcribed verbatim, then coded to identify the key phrases and terminology used to describe data science. This allowed for qualitative analysis to identify repeat phrases, unique phrases, or terms used to describe the functions. In addition, the skills, tools, and system capabilities needed to perform the work were captured and summarized. The individuals interviewed are across different domains but each person has multiple years of experience working with data. Interviewees are from the following fields: data analyst, data architect, data scientist, chief financial officer, academic professor using data, or business executive leveraging data for daily decisions.   The interviewees span a wide range of tenure typically found in the workplace; less than 5 years, 6 to 20 years and greater than 20 years.

INTERVIEWEE	ROLE	INDUSTRY	TENURE
1	Chief Financial Officer	Healthcare	20+
2	Academic Professor / Board of Directors	Education	20+
3	Senior Operations Executive	Financial Services	20+
4	Senior Operations Executive	Financial Services	20+
5	Senior Consultant on Data Analytics	Consulting	20+
6	Data Architect	Technology	20+
7	Senior Data Analyst	Marketing	5
8	Senior Business Analyst	Financial Services	5
9	Data Scientist	Consulting	5

INTERVIEWEE	ROLE	INDUSTRY	TENURE
10	Senior Business Analyst	Financial Services	4
11	Data Analyst	Marketing	4

Summary of Surveys - Free Form Text & Multiple Choice (s) Approach

Working individuals and students were requested to complete a survey of eighteen questions. Seventy-eight people responded and provided answers. Seven of the eighteen questions are open-ended; eleven questions are multiple choice (s) and allow for users to select “other” in order to enter anything that they use for data science that was not listed. The open-ended questions allow for the individuals to contribute their free-form text of what data science and data scientist means to them, their organization and who are the people performing the work. In addition, several questions were asked about the functions, programming and systems used to do the work by selecting multiple options. The surveys were sent to people in the researchers’ professional network while others were randomly selected and included data scientists from companies across the world. Questions on data management, data hygiene, machine learning, frameworks, and methods were included to assess an individual’s knowledge of other data related topics.

The survey respondents are from the following fields: college student, business executive, technologist, marketing, data analyst, data scientist, academic professor, real estate, government, healthcare, education and construction. Both qualitative and quantitative analysis is used to describe the results. The respondents span a wide range of tenures typically found in the workplace.

INDUSTRY	RESPONSES
Financial Services / Banking	22
Technology	18
Education	8
Healthcare	8
Real Estate	5
Marketing	5
Consulting	4
Communications	2
Human Resources	2
Other - Non-Profit, Government, Religion, Writer, Construction, etc	12
Total	86
*12 Respondents Checked multiple industries **6 Respondents Checked Other but did not indicate an industry	

TENURE	# OF RESPONSES
0 - 5 Years	26
5 - 20 Years	25
Greater than 20 Years	29
Total	80
*2 Respondents had multiple responses selected	

Introduction

What is Data Science and who are Data Scientists?

According to the US Department of Labor publication by Royster (2013):

> Today’s datasets are so big, they are measured in exabytes—one quintillion (1 followed by 18 zeroes) bytes. By comparison, an mp3 song is typically less than 10 megabytes (1 followed by 6 zeroes).

The algorithms and commands data scientist are creating to enable machines to sort through huge stores of data can be complex. An approach that supports technical advancements while allowing business executives the freedom to compete for business contracts is strongly encouraged. Although it was not intentional, all academic sources used are less than five years old, with the exception of one source. This is a direct result of the relative new usage of the terms investigated.

Professors and scientists describe big data and the techniques for managing, analyzing and systematically generating results that allows multiple industries to consider these techniques in their own the decision making process.   Brown, Court and Willmott (2013) writes:

> Without sufficient senior leadership, it is difficult to catalyze the widespread organizational change needed to capture data analytics opportunities. Capturing data related opportunities to improve revenue, boost productivity, and create new businesses puts demand on companies requiring not only new talent and investments in information infrastructure, but also significant changes in mind-sets and frontline training.

Data analysts, statisticians, mathematicians, technologists and data scientists, together with top senior business executives, are working to formalize the definitions, techniques, labels and systemic configurations for this relatively new field in order to improve revenue and increase productivity while providing opportunities for data scientists to advance the field of data science.

Key Findings

There is a commonality between the academic resources and practitioner's perspective shared in the interview and survey. Looking across all findings, trends develop that allow for the ability to group the attributes required for successful data science execution.

Regarding data science, many sources summarized the terms similarly. Key phrases included a combination of statistics, math, and programming in order to draw meaningful insights from the data. In addition, the authors offered perspectives related primarily to science, structures, controls and processes used to perform the functions.

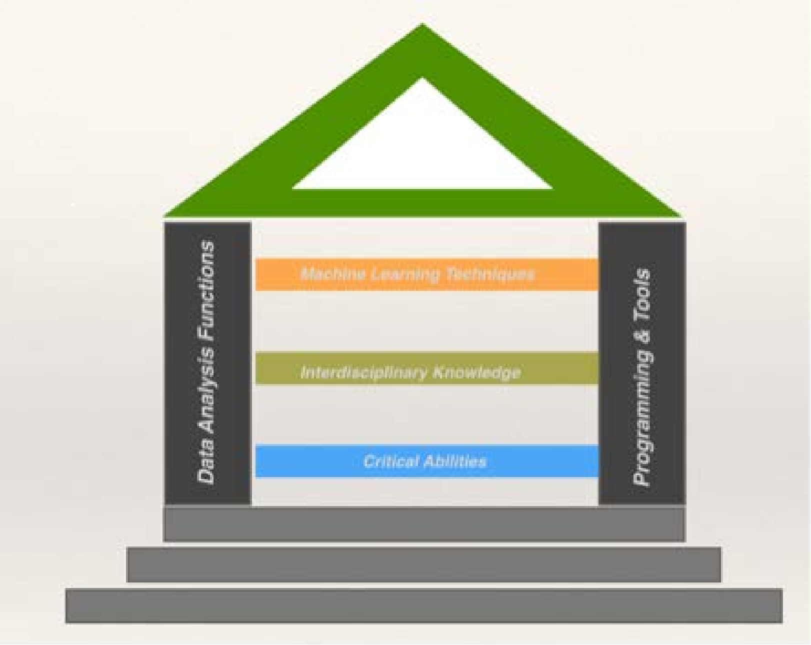
An example of this is provided by the Cleveland and Hafer (2014) commentary to respond to concerns raised by the Kary Myers and Scott Vander Wiel (2014) document. Myers and Vander Wiel compliments Cleveland's Action Plan (2014) and "how much of it has been integrated into the fabric of the statistics community and especially as it relates to multidisciplinary projects and computing with data." They state that "just as Cleveland said, our involvement blurs exactly who is a statistical and who is not. In addition, it blurs who is and is not an astronomer, physicist or chemist." Myers and Vander Wiel close their document by declaring that Cleveland's action plan has been used to determine who should be hired. A person with the ability to solve problems across disciplines and analyzes data that crosses outside of the boundaries of a statistician is a valuable asset for a business.

For data scientists, the skills can be grouped into five key areas:

- **Data Analysis Functions:** These capabilities are the types of math and analytical methods and modeling skills needed to solve data problems.

- **Programming & Tools:** These are the technical abilities that are required for data scientists to take data and either use a tool or write code to customize a tool.
- **Machine Learning Techniques and Solutions:** These are methods used to review data and build supervised and unsupervised machine learning proficiency.
- **Interdisciplinary Knowledge:** These represent the functions that must exist across academic disciplines such as accounting, economics and information systems.
- **Critical Abilities:** The capabilities listed here are the softer skills that are required for data results to be summarized and communicated.

Overlap across the three populations, literature, surveys and interviews, determined that Data Analysis Functions include statistics, basic math, algebra, calculus, programming and interdisciplinary knowledge that are important to be successful. There is notable division seen in the Programming and Tools that are in the practitioner's mainstream focus. Tools such as Spark, Radiant, Tensorflow, Tableau, Qlikview and SAS are not documented in the academic literature. It is possible that these tools are new to the workplace and have not been fully embedded into the academic culture. These tools are highlighted in the interviews, and even more by the surveys. Ability to write code using Java, C, C++ and HTML are common among practitioners. The academic community did not consider this skill as necessary for data scientists.



Regarding Machine Learning Techniques and Solutions, minimal capabilities are highlighted by academics. Practitioner's use of decision trees, neurals, vectors, and others listed in the table below as important for a data scientist. When machine learning was brought up during the interviews, the respondents were quick to opine that we are in a pioneering stage in the business world.

The Interdisciplinary Knowledge category results emphasize that the field of data science and skills required are across many academic disciplines in order to develop solutions and answer data problems.

What was unexpected is the lack of commentary in the academic literature on the capabilities captured under the Critical Abilities category. These skills are often called soft skills; communication, data visualization and subject matter expertise mentioned by the academic experts. None of them mentions key attributes such as the ability to ask intelligent questions when defining a data problem, interpersonal skills, critical thinking or curiosity. These specific capabilities are mentioned numerous times in the interviews.

One area that evoked conversation during the interviews is data intuition. Respondents would either ask for clarification on the topic or begin to explain how this is important. This capability is not considered as a sixth sense, but it distinguishes data scientists who have the ability to quickly make sense of data and determine the important aspects while dispelling the irrelevant information.

The table below captures the results from the three research methods supporting this dissertation. It can be used as a reference when reading the literature review results and qualitative results from the surveys and interviews. It can also be used as a hiring guide for managers looking to develop a data science team in their own organizations.

5 Key Skill Categories / 55 Competencies

Skill	Competencies	Literature Review	Interviews	Surveys
Data Analysis Function	9	9	5	9
Programming & Tools	13	6	9	13
Machine Learning Techniques Solutions	10	0	0	10
Interdisciplinary Knowledge	5	3	3	3
Critical Abilities	18	9	13	3
Total	55	27	30	38

Table of all Competencies

Row #	Skill Categories	Competencies	Literature Review	Interviews	Surveys
1	Data Analysis Function	Statistics	YES	YES	YES
2	Data Analysis Function	Basic Math	YES	YES	YES
3	Data Analysis Function	Calculus	YES	YES	YES
4	Data Analysis Function	Multi-variate	YES		YES
5	Data Analysis Function	Algebra, Linear and Quadratic models	YES	YES	YES
6	Data Analysis Function	Integer Programming	YES		YES
7	Data Analysis Function	PearsonR, MLIB, Lambda Functions or Chi-Square	YES		YES
8	Data Analysis Function	Tests for Significance	YES		YES
9	Data Analysis Function	Standard Deviation	YES	YES	YES
10	Programming & Tools	R Programming	YES	YES	YES
11	Programming & Tools	Python Programming	YES	YES	YES
12	Programming & Tools	Excel	YES	YES	YES
13	Programming & Tools	VBA		YES	YES
14	Programming & Tools	Java, C, C++ and HTML			YES
15	Programming & Tools	SQL		YES	YES
16	Programming & Tools	Spark			YES

Table of all Competencies

Row #	Skill Categories	Competencies	Literature Review	Interviews	Surveys
17	Programming & Tools	Tensorflow			YES
18	Programming & Tools	Radiant			YES
19	Programming & Tools	Tableau / Qlikview	YES	YES	YES
20	Programming & Tools	SAS		YES	YES
21	Programming & Tools	Data Wrangling	YES	YES	YES
22	Programming & Tools	Hadoop / Tessera	YES	YES	YES
23	Machine Learning Techniques	Decision Trees			YES
24	Machine Learning Techniques	Ordinary			YES
25	Machine Learning Techniques	Neural			YES
26	Machine Learning Techniques	Vectors			YES
27	Machine Learning Techniques	Clustering			YES
28	Machine Learning Techniques	Independent Component Analysis			YES
29	Machine Learning Techniques	Natural Language Processing			YES
30	Machine Learning Techniques	Apache			YES
31	Machine Learning Techniques	Amazon Machine Learning			YES
32	Machine Learning Techniques	Azure ML, Caffe, H2O, Massive, MLIB mlPack, Pattern, Shogun, Torch, Tensorflow			YES

Table of all Competencies

Row #	Skill Categories	Competencies	Literature Review	Interviews	Surveys
33	Interdisciplinary Knowledge	Accounting	YES	YES	YES
34	Interdisciplinary Knowledge	Economics	YES	YES	YES
35	Interdisciplinary Knowledge	Computer Programming / Information Systems	YES	YES	YES
36	Interdisciplinary Knowledge	Marketing	YES	YES	YES
37	Interdisciplinary Knowledge	Decision Science	YES		
38	Critical Abilities	Data Management / Governance	YES	YES	YES
39	Critical Abilities	Database Management (technical)	YES	YES	YES
40	Critical Abilities	Strategic Thinking		YES	
41	Critical Abilities	Ability to ask intelligent questions		YES	
42	Critical Abilities	Organization (of data, of concepts, of priorities)	YES	YES	
43	Critical Abilities	Data Visualization	YES	YES	YES
44	Critical Abilities	Communication - Written		YES	YES
45	Critical Abilities	Communication - Verbal		YES	YES
46	Critical Abilities	Interpersonal Relationships		YES	
47	Critical Abilities	Data Intuition	YES	YES	
48	Critical Abilities	Critical Thinking / Logic		YES	
49	Critical Abilities	Curiosity		YES	YES
50	Critical Abilities	Hacking Skills	YES		YES
51	Critical Abilities	Scientist Skills	YES		
52	Critical Abilities	Quantitative Analysis Skills	YES	YES	
53	Critical Abilities	Trusted Advisor	YES		
54	Critical Abilities	Business Expert / Subject Matter Expert	YES	YES	
55	Critical Abilities	Focus On Precision	YES	YES	YES

Defining Data Science

Several articles cite that data science is a multidisciplinary field; Cleveland (2014), Myers and Vander Wiel (2014), Aasheim, et al. (2015), Coderre (2015), Davenport (2001) to name a few. The academics and scientists have determined that the tasks performed in data science require a combination of statistics, programming, math, special tools, technical architecture understanding and business expertise.   There is agreement and disagreement by academic scholars and practitioners on the definition of data science. Consensus documented, through the findings, that data science is multidisciplinary requiring portions of information science, statistics and database management skills. The disagreements in the academic and scientific communities emanate from the lack of uniformity in the process of taking un-specified portions from each discipline to perform data science. In addition, there is controversy between the data science and statistics communities; mostly because data science is not measurable whereas statistics has well-formed theories that have been used for generations. Each data problem is unique and data science is part art and part science, requiring both soft skills and technical abilities.

Literature Review

William S Cleveland's (2014) *Data Science: An Action Plan for Expanding the Areas of the Field of Statistics*. This action plan is most cited and points to needed improvements in order to drive data science forward. He is purposeful and clearly states that data science is *the altered field of statistics* (Cleveland, 2014). Cleveland's (2014) plan focuses on the data analyst and describes six areas of focus: Multidiscipinary Investigations (25%), Models and Methods for Data (20%), Computing with Data (15%), Pedagogy (15%), Tool Evaluation (5%), and Theory

(20%). Cleveland (2014) also outlines specific steps in analyzing data by explaining that data must be cleaned before it can be utilized for analytical purposes. Data management should be strict, with no exceptions, in order to be effective.

In Cleveland (2014), he states that data science consists of all technical areas that come into play in the analysis of data and deep analysis of large complex data challenges.   All of the technical areas, from statistical theory to the architecture of clusters designed specifically for data, need to be tightly integrated. He states that statistics should be considered a synonym for data science. The plan outlined by Cleveland (2014) claims that a substantial change is required in the major areas of technical work and statistics in order to create optimal results for the data analyst (data scientist).

Cleveland (2014) documents his perspective on how the analyst's time is spent on multidiscipinary functions to solve problems, and his comments are not unique. Data science requires mathematics, statistics and programming tools for the data analyst (data scientist) to be successful. However, Cleveland's commentary on models and methods is distinctive in that his recommendation is to apply two aspects: specification - the building of a model for the data, and estimation and distribution - formal, mathematical-probabilistic inferences based on the model. He intentionally emphasizes that data science is a process made complete by precision. Cleveland (2014) states, "the model must be balanced by information from the data, information from sources external to the data and desirability of parsimony." His emphasis is on the tools for the data scientist and ensuring that their requirements are clearly understood.

Regarding the importance of computing tools, Cleveland (2014) shares that hardware and software available in today's business world are powerful and statisticians should look to computing for knowledge, just as data science looked to mathematics in the past. Although pedagogy and tool evaluation are part of Cleveland's formula for the data analyst (data scientist), he focuses more commentary on the theory. He states "another provocative perspective that theories on mathematical and non-mathematical are vital for data science. And the tools of data science - models, methods, and computational systems link data and theory."

In addition to the previous reference on precision, Cleveland and Hafen's (2014) commentary on *Divide and Recombine*, state that that data must be divided into subsets by a statistical division and the subsets are required to be stored in the same data structure, either on disk or in memory. Cleveland and Hafen (2014) outline a system design including a detailed profile of Tessera and Hadoop solutions. This optimizes the computational capabilities for data analyst teams performing big data mining and data analytics. Typically data mining refers to smaller datasets. They are the only scholars who discreetly describe the controlled process and systems requirements in detail in order to educate the proper process and logical divide and recombine process. Although this may seem to be a logical process, it is obviously meant to be emphasized; in their own words "as an example to demonstrate work in all the areas and their tight integration." The details of their architecture design will be shared later in this document.

Aasheim, Williams, Rutner, Gardiner (2015), outline the specifics of data science as a set of fundamental principles that support and guide the extraction of information and knowledge from data. Their document separates data analytics from data science as it relates to undergraduate programs. The information they share is vital to the data science conversation because it expresses a sense of urgency to solve the problem and prepare skilled workers for the future due to the rapid, exponential growth of data stores.

Their perspective also includes the defining characteristics of big data as the three V's; Volume, Velocity and Variety. Aasheim, et al. uniquely states that "what data scientists do is make discoveries while swimming in data." This is giving reference to the vast expanse of data today's world generates by the millions of mobile and personal computing devices.

When summarizing their perspective on the data scientist's primary skills, Aasheim et al. write data scientists require traditional relational database management systems as well as the ability to extract, transform, load, and data mine. Although their document is primarily focused on comparing the education courses for undergraduates on data analytics and data science, it is informative to view their perspective and understand the skills required as well as the gap in society's working class who have the skills necessary to perform the data analytics (data science) tasks needed by businesses. Unlike Cleveland, who focuses on systemic process and technical solutions, Aasheim et al. focus on skills, functions and knowledge required by the person performing the tasks for data science.

Maytal Saar-Tsechansky (2015), author of The Business of Business Data Science in IS Journals, has written for MIS Quarterly and focuses on the specific business data problem nicely. His writing bridges the gap between old and new business problems. His specific guidance to data scientists is to start by outlining what is meaningful and significant information to share about the data science findings. He states explicitly that "data science is a design science field of research and a broad, interdisciplinary field but it has uncertainty on how to apply the guidelines of data science to information science problems. He fully admits that data science has important contributions and can impact science in novel and meaningful ways."

What is interesting is that Saar-Tsechansky (2015) focuses on clear articulation of findings from his data science colleagues.   For instance, Brown, Court, Willmott (2013) state in their article, *Mobilizing the C Suite for Data Analytics*, the power of data and analytics is profoundly altering the business landscape.   Companies need to clearly articulate data related opportunities to improve revenue, boost productivity, and new business opportunities. Establishing new mindset is an imperative to successfully articulate a data analytics strategy. Key partnerships are required by the business unit leader and the data analytics expert in order to pioneer the needed frontline changes.

Saar-Tsechansky (2015) writes that data science is the extraction of informative patterns from data, and that it differs from that of other streams of IS research and thus calls for different assessment guidelines. Since data science is a design science field of research, the guidelines outlined can be applied to produce and assess data science contributions. Data contributions are fairly recent in IS (Information Systems) and there remains uncertainty on how to apply the guidelines to data science.

Kate Matsudaira (2015) highlights the significant communication gap and strategic planning challenges she faced when she took a role to manage a data science team. It was clear the team was proposing ideas, investigating hunches and testing hypotheses, but it was difficult to estimate the work effort and guarantee the results would provide significant business value. To overcome this challenge she develops a new process that would account for the uncertainty and keep stakeholders in the loop on the progression of the team's work. Developing a communication plan to bring transparency was step one. The team develops a process to answer stakeholder questions. The challenge is to answer the question, but use the right vocabulary. Her team develops a model that would allow stakeholders to understand the definition of "done".

The data science team measuring the results and completing an experiment need to determine that the model was strong enough to be integrated into a product.

In order to close this gap, the team develops a precision measure for each algorithm and communicated the results. The team began using terms that would explain the customer experience and business metrics in language that stakeholders could relate to easily. In addition to these changes, the team develops ways to show improvements, clearly express the level of complexity surrounding a specific question, add deadlines to the research, and develop agile demonstrations in order to showcase the work. The best practices she develops are adopted and results replicated across data science teams. This adds valuable adjustments to the overall communication strategy and data visualization effectiveness of data science solutions.

Mellin (2013) details the stages of an organization and adoption of analytical capabilities and progress towards full maturity. Mellin specifically highlights five key questions that all executives should be asking to press adoption of data practices and decision making into their organizations. Mellin's summary document is heavily centered on Davenport's (2001 and 2014) work but adds data governance and data reliability topics into the conversation that are not present in all of the prior works discussed in this dissertation. Data governance must be important at all times. A NO EXCEPTIONS mandate is essential for data quality and effective data governance. Data reliability assessment is vital to the success of an organization's data adoption and oversight of the accuracy of the data.

Brian Hayes (2014) is helpful in visualization of defining the complexity of data science. Hayes (2014) not only describes data science as the sexiest job of the 21st century, but he also includes the controversial aspects that underlie the overlap between statistics and data science. Hayes

(2014) describes the role of data scientist as someone who knows how to extract meaning from and interpret data using tools and methods from statistics and machine learning.

Peter Diggle's (2015) answer to the "what is data science" question is not new. Diggle (2015) shares that data science is "the extraction of knowledge from data….it employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, and information technology. Information science is an interdisciplinary field primarily concerned with the analysis, collection, classification, manipulation, storage, retrieval, movement, dissemination, and protection of information. Statistics is the study of the collection, analysis, interpretation, presentation and organization of data." Diggle (2015) admits that the definitions show considerable overlap and states clearly that the definition of data science and statistics are close. Data science in itself is not just another name for statistics as data science includes informatics (hardware and software engineering). In summary, Diggle (2015) writes that data science, information science and statistics has considerable overlap and he has argued in the past that data science was just a new name for statistics.

The Hayes (2014) and Diggle (2015) summaries are versions of the same explanation Cleveland (2014), Saar-Tsechansky (2015) and Aasheim (2015) describe. They all also acknowledge this seems to still be a field that is in an evolutionary state to have full consensus by scholars on what data science is and who data scientists are. They admit this is multidisciplinary and requires algorithms, systems and applications to solve problems effectively.

One area where the authors diverge is the level of comfort that is interpreted from their writings on the topics of model, methods and measures. Because statistics has a foundation that has been baselined for generations and leveraged by all scholars to some degree as they performs their research, it is uncomfortable for the statisticians to agree that the looseness with data science measures needs to be evolved. The "king of the hill" measure described by Hayes (2014) is not empirically accepted and anyone would be hard pressed to use it as the foundational cornerstone to data science. Hayes writes that what is empirical with data science is that we still need a central limit theorem that we can all feel comfortable in measuring the success of a model and results.

King of the Hill Measure: the primary measure used today to say we are performing data science activities well.

Examples of our inability to actually measure the success or failure of data science are presented below. The data stores are massive and results have never been delivered to the consumer before so there is nothing to compare against.

- Google search and how quickly responses for any topic are returned within milliseconds on any browser.
- Netflix business model conceiving videos as data and the video store as a data library. They deliver movies to subscribers via internet streaming and suggest similar movies to those previously viewed before.

Two additional articles provide unique perspective are by Bayrak (2015) and Mayhew, Saleh and Williams (2016). In the article *A Review of Business Analytics*, Bayrak writes many of the same summary comments of previous authors. What is different about his findings are the terms used to describe analytics. Specifically, he defines what businesses are doing today when they use metrics to measure productivity, accuracy and forecast staffing. In addition, he describes the type of analytics that can be used for prediction and heavy computations.

Types of analytics described by Bayrak (2015):

- Descriptive Analytics: Data mining to provide trending information on past and current events.
- Predictive Analytics: Models and techniques to predict future outcomes based on historical and current data.
- Prescriptive Analytics: Set of mathematical techniques that computationally determine a set of high-value alternative actions or decisions given a complex set of objectives, requirements and constraints.

All three types are used to improve the overall performance of a business, but are not interchangeable and must be performed sequentially for an organization to advance in their adoption of a data driven decision making culture. This breakdown of analytics is referenced in two interviews specifically and not included in any of the surveys.

In *Making Data Analytics Work for You,* Mayhew, Salem and Williams (2016) write about the data science revolution is here and transforming organizations. The primary purpose of their document focuses on making improvements in the data science processes used in organizations today. What is unique about their document, is that they focus on communication in a section titled “Make your Output Usable - and Beautiful”. They recommend that the appearance of interfaces which must have elegance and intuitive usability. Further, they state that the quantitative calculations driving the facts must be easily interpreted and understood in order for them to be effective and people to take action or respond to the facts.

In a short article on *Challenges in Data Science*, Carbone, Jensen and Sato (2016) write about the four V’s of data science. This work compliments the work of Aasheim, et al (2015) in their three V’s perspective. Carbone and team list the four V’s as Velocity, Volume, Variety and Veracity. Veracity calls out the uncertainty tied to data science and to highlight that analytics are only as good as the accuracy and validity of the factors used to assess the data. They want to ensure that this measure is controlled and focused in order to ensure data science success. The rest of their document substantiates other definitions of data science with the standard science, math, and computing background that we have seen from Cleveland (2014), Hayes (2014), Diggle (2015) and others.

This review exposes data science as a term used to define multiple methods used across businesses in order to determine an answer to a business problem by using data. The definitions are often vague and lack specificity which can cause conflict with scholars and business teams trying to develop the tools to allow organizations to solve data challenges.

Practitioner's Perspective

In summary, the practitioners say that in today's business world, there is a limited amount of knowledge on how to manage today's massive amounts of data. Simply labeling the data stores as big data is not enough. The science driving a data strategy to implement the optimal methodology for data management, data mining, data analytics, data governance and data decisioning is not widely known nor understood by today's business experts. Large corporations are developing data strategies in order to gain key insights, optimize revenue, hire and retain top talent, effectively market, place products, reduce expenses.…the list is endless. In addition, both academics and practitioners are exploring automation and machine learning capabilities that will result in evolving today's human manual processing tasks.

In laymen's terms, data science is the scientific approach used to perform technical analysis with surgical, robot-like precision on data. The data is large, has complex relationships, and has a variety of orders and structures.

In reality, companies are using established tools, and a fixed mindset to solve gigantic business problems. What they are missing by staying in the same mental confines and by using historical methods are the potential opportunities that may offer new insights. Some practitioners cannot conceive what it means to have unsupervised machine learning providing results for critical business decisions. Since they cannot conceive the concepts, they mistrust the results, and stay chained to the old regimens.

In some organizations, data science is separate from statistics, algorithms or the technical approach to analysis. Their versions of data science revolves around answering the question of the moment using a methodology that is easy to follow and explain. It can seem that being able to explain the results is more important than allowing people to explore the data and determine what hidden gems can be found. Often today’s data are too complex, too large and too unstructured to effectively analyze using old tools. In addition, the skills needed to be effective have not been fostered or developed by the organizations.

Booz Allen Hamilton (2015) summarizes data science as the “catalyzing force behind our next evolutionary lead. Our own evolution is now inextricably linked to that of computers. Data is our new currency and data science is the mechanism utilized to tap into it. Data science is the art of turning data into actions.”

Interviews

The interviewees responded with a definition that data science is not just one thing; it is a combination of techniques, knowledge and skills that are applied to data in order to identify insights that can be used to guide teams to make key decisions. The phrases most used include gain insight, reveal business trends, and make sense of data that requires multidisciplinary techniques across science, math, and computer programming. A small number of comments refer to the cloud as a prerequisite for big data or that managing data can only be done in excel, access or on mainframes. With comments like this, it is evident that there are misconceptions about data science and how it is applicable in the business workplace. In some cases, terms like data mining, data analysis, data research are used to explain data science. This is likely driven by the fact that no formal definitions are in place to explain data science.

Key words that are used include statistics, organizing, analyzing and researching to describe the activities, complexity of the topic and the methodical approach that is required in order to truly garner benefits from the huge volumes of data that are generated each day. To perform the tasks, an understanding of multiple concepts including statistics, math, and data management techniques are necessary to scrub data.

Summary of Responses Identified from the Transcripts:

- the meeting of statistics, math and science to use raw data to make decisions; using analytics, studying it and extrapolating the most useful pieces
- the ability to harvest the data, organize it, analyze it and draw meaningful insights
- taking information from various sources, research and making connections from the data
- ability to understand the data with statistical analysis; performing the analysis and higher level methodologies to gain insights
- to glean some sort of insight or knowledge out of data; get new insight and knowledge from data
- an empirically driven approach with a scientific method towards solving business problems drive strictly off of data items; using multiple approaches and coming to a conclusion that best fits the stated question
- interdisciplinary field that combines several other fields of business and social science, particularly mathematics, statistics and management of information systems to get data from systems, and organize it in a meaningful way using a scientific method; to reveal business insights that would not normally be revealed through elementary business practice

- data mining is discovering information that may never even be thought about; there is rationale for picking those questions out
- ability to make sense or logic of undefined information
- to understand patterns for example prices of aluminum parts coming from Asia tend to be more expensive between October and December when it is raining and it is an odd year
- DS is a new interdisciplinary field that is combining different techniques and skills

Surveys - Free Form Text

A survey with eighteen questions was sent; seven open ended / free form text questions and eleven multiple choice questions are included in the survey. A total of seventy-eight responses were received with 100 percent completion on all questions. The respondents are diverse and come from a variety of backgrounds, industries, roles, tenure and income. The responses provide a colorful array of comments that demonstrate what is known and unknown about the terms data science and data scientists. There are a lot of people who have a good perspective on data science, data scientists and their role in modern day business, and there are several who have no idea how to describe data science and the role data scientists perform. Qualitative and quantitative results are derived based on the answers to seventeen of the eighteen questions. One question is removed due to ambiguity and feedback from several participants.

The survey results, answering the question in free-form text “What is data science?”, indicate similar trends as the interviews. Twenty-six of the seventy-eight respondent’s comments are vague or uninformed that they are excluded for this particular question. Comments such as "science based on data" or "science that has been studied". The remaining fifty-two respondents are consistent with the interviews. The comments provide additional supporting evidence that data science is not just one thing, it is an aggregation of multiple disciplines.

Examples of Free-form Text answers are listed below:

- using existing data sets to test hypothesis, employing sophisticated statistical methods to identify relationships and understand trends
- the art and science of extracting valuable information from existing data for the decision making process
- data science is a field focused on understanding data through tools, models and analysis, presumably to use that data to make informed and better decisions
- data science is another word for applied statistics
- software engineering is data science if a project that is mostly software programming has some type of statistical application embedded in its code
- data science is the art of combining mathematics, programming and visualization to discover knowledge
- field of study where math, statistics, information technology and analysis is used on large volumes of data to gain knowledge and insight
- it can include data mining, pattern recognition, and data clustering analysis
- data science is the ability to extract information algorithmically from large volumes of structured and unstructured data
- data science is the use of statistics, economics, and programming languages to extrapolate and visualize data

The terms used the most to describe data scientists are listed below:

- statistics / models occurred in twelve responses
- math / algorithms and methods occurred in six responses
- trends / programming occurred in five responses

As explained, there is significant consensus that highlights the various scholastic disciplines required for data science to be performed. Overall, it is clear that data science requires variable amounts of mathematics, statistics, programming, methods, economics, and accounting to have output that is usable in the business world. The use of data science can translate into bottom line impacts for businesses that are service providers, buyers / sellers, marketing, government, health and education professionals.

Defining Data Scientist

In 2012, the United States Department of Labor wrote an article titled Working with Big Data. An excerpt is captured below:

> Most workers who deal with big data are known as data scientists, although they may be called data analysts or have some other designation. The term data scientist is so new, we do not yet have it in our job descriptions at Fermilab, says physicist Robert Roser head of the Scientific Computing Division at this national laboratory in Batavia, Illinois. The US Bureau of Labor and Statistics classifies these workers as statisticians, computer programmers, or in other occupations depending on their tasks. Whatever their title, these workers study big data using conventional and newly developed statistical methods.

Although the article was written five years ago, when a recent search was performed on the US Dept of Labor website, the job of data scientist is not found. The only other reference that was any way relatable to big data was a second article discussing the STEM crisis or surplus question. Nonetheless, the role of a data scientist has made its way into corporate America and even into government, marketing agencies, service providers such as Price Waterhouse-Cooper, Accenture, and Cognizant.

Literature Review

Cleveland and Hafen (2014), in *Divide and Recombine: Data Science for Large Complex Data*, articulate a systemic solution and process that divides the data into subsets by a statistical division method where the subsets are stored in memory, computations are performed and then the results are recombined via a statistical method. This output requires that deep analysis goes to the most granular level. The output, should not only be summary statistics, an automated data reduction algorithm, or random samples of data as the outcome will likely be missing large chunks of critical information.

In addition, Cleveland and Hafen (2014) outline a system that has a Hadoop back end running on a Linux server cluster, Tessera middle software layer and R front end. This solution allows for the parallel processing capability, computing power and optimizes the statistical analysis required for large data sets. Hadoop has the ability to manage the subsets described in the D&R function and manage the processing capability with speed. In addition, Cleveland states the applicability of D&R with Tessera is wide and allows for deep analysis and performing the complex computations for massively large datasets. Cleveland and Hafen (2014) recommend that departments of data science should contain faculty members, who devote their careers to advances in computing with data and who form partnerships with computer scientists.

In Cleveland's (2014) article *Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics*, the people performing the tasks of data science are called data analysts. As mentioned in the data science definition, Cleveland outlined the data scientist functions containing various percentages of time for multidisciplinary, models and methods, computing with data, pedagogy, tool evaluation, and theory. What he is not explicit in stating, but one could infer from the article, is that data scientists require the ability to perform computer programming, database skills, modeling, tool assessment, and statistics. He focuses on the fact that analysts (data scientists) will need optimized tools and systems in order to develop precise and functioning models to solve business problems.

David Coderre (2015) wrote *Gauge Your Analytics* and mentions that the solution should include an audit solution that will log repeatable tests and document the test design so that the analytics are tracking as an integral part of the audit process. This may seem pedantic, but this idea is not specified in the other solutions. He specifically highlights that the audit team should include a data analytics function in order to ensure the audit functions can move along the analytics maturity curve to be better aligned when the analytics function becomes complex. Per Coderre's guidance, the audit data analysts' function should be able to perform some level of programming and have the technical skills needed to ensure the models are verifiable and reproducible. This perspective is important since this topic has not been addressed by other authors. Specifically, he writes the audit data analyst should have:

- understanding of data concepts
- understanding of database structures (both logical and physical)
- ability to communicate with IT and related functions to achieve optimal results
- ability to perform ad hoc data analysis
- ability to design, build and maintain well documented, ongoing automated data analysis routines
- ability to provide consultative assistance to others who are involved in applying analytics

In the Mondore, Spell and Douthitt (2016) article, *From the Boardroom to the Front Line*, the authors write that many leaders will agree that correlation does not equal causation and then accept research that is based on correlations and group comparisons. Correlation and simple regression identify the strength and direction of relationship between two items. In addition, they state that statistical modeling methods called structural equations modeling (SEM) allow for various factors or causes to be assessed in relation to multiple outcomes concurrently. This is important because events do not occur in a vacuum, but rather with multiple influencers. Specifically, they cite four advantages to SEM:

- multiple inputs or causes can be tested along with multiple outcomes concurrently
- an accurate assessment of ROI can be calculated
- it provides the ability to correct for measurement error
- causation can be inferred

Mondore, et al. have outlined a handbook type article that provides specific guidance to leaders for ensuring that statistical rigor, correlation, regression, and advanced analytics can be leveraged to achieve the SEM output. Data analytics can have a positive impact on the bottom-line and ultimately increase dividends.

Davenport, Harris, DeLong and Jacobson wrote *Data to Knowledge to Results: Building an Analytical Capability* in 2001 and describe the competencies needed for a data driven culture. What is important to include here is their description of the data team members needed to develop strong analytical capabilities; database administrator, business analyst, data modeler, decision maker, and outcome manager. The team members will need strong technology skills to leverage software for manipulating and analyzing data. In addition, Davenport et al. state that statistical modeling and analytic skills are required to run models and assess results. They also include, knowledge of the data fields, business, and strong communication skills as necessary to apply the right models to the business problems that are relevant to today's business leaders. In *Data Scientist: The Sexiest Job of the 21st Century*, Davenport and Patel (2012) writes data scientist is a high ranking professional with the training and curiosity to make discoveries in the world of big data.

In *The General Theory of Decisions*, Aliev, Pedrycz, Keinovich, Huseynov (2015), summarize a new theory for decision models. Although their article is focused on the existing theories, practices, and approaches, there are concepts that are relevant to the topic of data scientists. They recommend a shift in the foundation of analysis to fuzzy logic versus the traditional binary logic. In addition, they focus heavily on the topics of models and the constructs required to have effective, accurate models for decision making. What this implies is that for data science to be accurate, the resources performing the tasks must have significant technical capabilities with tools, calculus and statistical models. Specifically, a mental state of a decision model is a complex system of factors; evolution of modeling decision-relevant information based on numerical information, interval-valued information, fuzzy information, and information with uncertainty.

In Davenport's (2014) book entitled Big Data @ Work, there is a full chapter titled, "The Human Side of Big Data". In this chapter, he describes the data scientist having five key traits:

- hacker - able to code and understand big data architecture
- scientist - can support evidence based decision making, improvise, and action orientation
- quantitative analyst - performs statistical analysis, visual analytics, machine learning and analysis of unstructured data

- trusted adviser - strong communication and relationship skills, able to frame decisions and decision processes

- business expert - know how business works to make money and has a good sense of where to apply big data analytics

What is unique about Davenport's perspective is that he focuses attention on the questions being asked to the data scientist rather than the data. He specifically highlights that between 70 - 80% of business intelligence projects fail in corporations because of poor communications between IT and the business managers. He is aware of the importance of solution communication skills. Davenport (2015) draws the difference between two types of data scientist; vertical (having deep knowledge on a narrow topic) and horizontal (having a combination of skills). He includes his perspective on hiring and retaining top data scientists by using the typical factors; salary and solid relationships with peers and the business manager.

Interviews

The interviewees ten out of eleven responded to the question "Who is doing this today in your organization" with a consistent answer — business analysts / data analysts perform these tasks today. In addition, it was alluded to that the analysts are junior members of the organizations and not responsible for driving strategy or innovation in this space. In reality, the questions today's business executives are asking require a true data scientist to review large volumes of big data and provide a scientific approach to performing an assessment on the data. The answers executives are looking for are not always lying on the surface, but require a multilayered approach in order to make sense of the vast data stores.

The best summary is a “pure data scientist is the one in possession of the best theoretical or practical applications of math or other sciences to leveraging statistics, use of advanced math, leveraging machine learning, or greater statistical functions to be more accurate.” In simple terms, they are the “person uncovering relationships in the data that were not known or possible to perceive with current tooling.”

Definition of Data Scientists - Summary of Responses from the transcripts

- team of data analysts who perform reporting focus on cutting edge options and look to industry trends on a larger scale rather than looking at data after an event has already passed
- data analysts use tableau and micro strategy and run reports \
- informatics and data analyst perform the analytics; informatics have subject matter knowledge about a particular area (clinical knowledge) and can perform the data analytic whereas a data analyst is pulling data and running reports
- the data analyst has accounting or finance degrees with minor statistics background but not engineering or higher math
- informatics are looking for trends in populations to improve care across certain populations of patients
- the quantitative analyst is usually a PHD, advanced math level trained in order to get the knowledge or insights from data

- pure data scientist is in possession of the best theoretical or practical application of math or other sciences to data
- must be good at understanding technology infrastructure to get access to large amount of data
- teams consist of data engineers who understand the business context and data analysts who use tools to interpret the data
- person uncovering relationships in data that were not possible to perceive with current tooling

Data Scientist Skills

Interviews

According to the interviewees, ten out of eleven responded that data scientists need are multidisciplinary knowledge and capabilities across math, statistics, accounting, economics, and programming. In addition, the data scientist must have the ability to perform critical thinking, apply logical reasoning, and have a curious nature, good personal skills and the ability to develop summaries that articulate their findings.

A constant theme is that a person performing data science must be able to perform calculations leveraging multiple techniques. Because advanced analytical approaches are required when looking for trends and significant relevance of particular components within the data, the consensus is that a typical data analyst is not going to provide the same level of insight because they have not been taught these techniques. A senior marketing analyst reflects the importance of data science results to their client's bottom line; "math, statistics, economics and accounting are all required for a data scientist to be effective as it helps companies understand their impacts to revenue, market share, margins, taxes, etc."

Math - Summary of Responses from the transcripts:

- deal with more statistics and high level thinking the typical analyst
- pulling larger information across customers to make decisions
- scientific qualifications in science and math

- must use statistical, math tools and code to apply to data sets
- ability to leverage well known modules and apply statistics techniques and math
- math, statistics, economics and accounting
- basic math skills
- engineering and higher math skills
- statistics deep diving more into regressive models and any kind of algorithms that may be more math or coding
- advanced statistical analytics
- a PhD in math
- ability to look at data in multiple ways
- there are traditional pillars of analytics that are part predictive and part prescriptive
- statistics, calculus, logic and critical thinking
- math training and high analytical training

The comments to the question “What programming capabilities does a data scientist need to perform their job?” ranged from the standard VBA, Excel Macros, and Sequel (SQL) to more complex statistics programs. Most respondents agreed that data scientists do not need to know how to write all of the code they use, however, they need to know about tools such as R, SAS, or Python in order to get the needed results. One respondent who works with data scientists on a daily basis commented “they know one to three programming languages, half a dozen statistical packages and modeling techniques such as R or Python. They typically know mathematical modeling and they know the business….they know what they are looking for.” The example shared was related to a team of data scientists working at General Motors. This team understands patterns for prices of aluminum parts coming from Asia —they tend to be more expensive between October and December when it is raining and it is an odd year.

Two interesting misnomers called out by the respondents are that you must know Mainframe programming to be a data scientist and to use Hadoop. It seems logical that experienced data scientists would have the ability to write code for these two technology platforms, but it is definitely not a requirement.

Programming - Summary of Responses from the transcripts:

- some technical coding abilities
- ability to use a suite of problem solving tools through pre-built modules such as python and Hadoop
- know one to three programming languages, half a dozen statistical packages modeling techniques, mathematical modeling and business too
- specifically python, R and if the information is coming from a database of some kind, they will need SQL
- SQL - grab a bunch of data and a lot of excel.
- VBA or excel macros
- ability to code the applications (programmer from another team would do the coding)
- statistical evaluations
- able to code or know what is capable of being code and ability to write computational algorithm using theoretical quantum mechanics

- python because it allows you to import what others create
- ability to use R and python; should be involved in developing software and provide requirement and test at completion
- leveraging off the shelf tools and should be able to code
- R is a very popular statistical purpose language
- Macro ability and mainframe writing, excel, access to and ability to know computers, logic, and combine that into business sense

Surveys - Free Form Text

When reviewing the summary data as it relates to "Who are data scientists?", the respondents are not widely different than the literature review and interviews. Nine responses were too vague and unusable out of seventy-eight. The remaining sixty-nine responses agree that data scientists can apply a variety of statistical, mathematical methods and tools to raw data in order to gain insights from the data for decision making. One respondent describes a data scientist as "someone who sits between the business and IT with the task of proposing and executing analytical strategies to identify hidden but actionable insights from business data for key business initiatives." What is unique about the surveys versus the interviews is the concise descriptions used to describe a data scientist. In the interviews, the respondents would expand on their definition and add commentary and justification for their definition or share how the data analysis is performed inside their organization. The survey respondents just focused on the specific question without adding background context. Their strength is their knowledge about the topic and ability to discern the relevant facts related to answering the specific question at hand.

In the Free-form text answers, terms used the most to describe data scientists are listed below:

- analyzes / evaluates occur in twenty responses
- statistics / mathematics occur in sixteen responses
- programming / uses tools occur eight times
- studies data occur six times
- categorizes and compiles occur six times
- methods and models occur three times

When compared to the responses the same participants submitted for the multiple-choice (s) question asking what data functions are used to perform data analytics, the answers had a lot more contrast. Fifty-two of seventy-eight (or 67 percent) responsed for statistics. Basic math has sixty-four out of seventy-eight (or 82 percent) while algebra has twenty-seven responses, multivariate functions receive twenty-two responses. Seven respondents choose Other and list lambda functions, predictive models, linear and quadratic models, PearsonR, Grubbs (outliers), integer programming and text analytics.

Free-form text sample responses are listed below:

- someone who tries to solve various problems via computer model
- data scientist uses different techniques to acquire, cleanse, curate, and standardize data to derive intelligence from the data
- data scientist is someone who applies methods to raw data to extract knowledge and insights

- data scientists are individuals with mathematical, statistical and programming acumen who extract information, insight and trends from structured as well as unstructured data
- data scientist uses statistics, economics, and programming tools to analyze raw data and create interpretations that can be used in a variety of ways for different audiences
- person who analyzes data using new and existing scientific qualitative and quantitative methods
- half tech, half business person who fully understands the collected data and its implications to business performance
- an individual who applies analytics to diverse data and obtain insights
- person trained/or has knowledge about extracting /evaluating information received by analyzing data (i.e.: google searches, metadata, client behavior, etc.)
- data scientist is an individual trained in advanced statistics but also has a specific domain expertise

Surveys - Multiple Choice (s)

When the same survey respondents were asked to describe in their own words the skills that a data scientist must have, the responses are consistent. The descriptions from nine respondents were removed due to vague or not applicable responses. The remaining sixty-nine responses provide a good aggregate of the overall skills a data scientist needs; math, statistics, computer knowledge, research techniques, ability to perform analysis, gather data, and perform quant skills. What was unexpectedly included are the terms used to describe soft skills, specifically that data scientists must have curiosity, executive communication, presentation skills, data visualization capabilities, and critical thinking to the data problem. Some responses contained reference to different concepts such as econometrics, data curation, and stochastic modeling. Some comments state that a data scientist needs to be able to perform linear and non-linear regressions, program in R, Python and SAS, and univariate or multivariate advanced math skills. This is great news because it shows that the respondents have a deeper awareness of specific techniques used by statisticians, mathematicians and technologists to perform data analytics and that the respondents understand statistics and math at a more granular level than the two terms that have been used times in the interviews, surveys and literature review. What is surprising is the consistent themes from literature to interviews to surveys.

Math Skills - Multiple Choice (s)	**# of Responses**	**% of Responses**
Statistics	52	66.7%
Algebra	27	34.6%
Calculus	9	11.5%
Multivariate Functions	22	28.2%
Basic Math	64	82.1%
Other - Lambda Functions, Predictive models, linear and quadratic models, PearsonR, MLIB, Grubbs (outliers)	7	9.0%

Programming Skills - Multiple Choice (s)	**# of Responses**	**% of Responses**
R Programming	17	21.8%
Python Programming	19	24.4%
Excel	63	80.8%
VBA	16	20.5%
Java	14	17.9%
SQL	30	38.5%
Other - Spark, Databricks, MsR, Tensorflow, Decision Insight, Businessbridge, CPLEX, Stata, SAS, Radiant, JASP, Tableau	8	10.3%

Related Skills & Topics

Without much influence, the interviews always led to comments about the additional skills that data scientists must know how to do. Critical thinking was a top skill that eight out of eleven respondents made reference to and for the scientist to be able to use their knowledge, curiosity, critical thinking, and intuition in order to approach a big data problem with scientific focus. The ability to communicate well, have good interpersonal skills, and most importantly, data visualization abilities were other favorite skills mentioned. Most respondents brought up Tableau or Qlik/Qlikview since these are two favorite software dash-boarding packages used across the business industry. Consistently, the interviewees felt that having the ability to ask intelligent questions is a requirement for a data scientist to be effective. Mixed comments are made when asked about data visualization as the respondents are split on if the data scientist should be the one performing the presentation and dashboard materials or if they need a colleague to partner with to perform these tasks.

When asked about their preferred technique to handling a big data problem, the respondents have a variety of answers but the common themes are the ability to take the problem and break it into manageable pieces. “Do not recreate the wheel by writing code, instead fetch code from a community - then drop and drag it into your tool….do not start from scratch,” one expert commented. Another respondent shares that they consider trend analysis immediately while others comment that a data scientist must understand the data canonical and be able to know the meaning of the data in question. As a standard part of analysis, reviewing the process and results with a business expert is a prerequisite in order to fully understand the problem, cut the data and determine the best way to present the results.

Machine Learning

Interviews

When asked about machine learning, the interviewees took a variety of approaches to explain what they thought it was and the stage of their organization relative to the use of machine learning. *A full consensus is that we are in the pioneer phase of this capability.* A couple of firms are using “bots” to perform scripted, repeatable tasks, but are quick to add that these capabilities are in the pilot phases and that there is so much more to understand about the concept of machine learning. The simplest, most concise comment made is that “machine learning is the ability to create models that allows machine learning software to discover things."

Surveys - Multiple Choice (s)

The survey results for the multiple choice (s) question asking what machine learning capabilities the respondents have used are varied. Decision trees (twenty-two), logistics regression (twenty-one), least squares regression (fifteen) and Clustering (fourteen) are the four leading techniques in use by the seventy-eight total responders. The second question on machine learning is about the framework currently used by the organizations. For this question there are clear leaders with Apache receiving sixteen responses, MLIB/ Spark receiving nine and Amazon Machine Learning receiving eight. All other responses were four or less that falls in line with the interview responses that this evolution of computer programming / data science is still in the pioneering phase.

Data Management

Interviews

When asked what does data management mean or how does it fit into data science and the tasks of the data scientist, the responses were from both sides of the spectrum. Some respondents immediately talk about their organization’s data requirements for data ingestion from their clients and the stringent mapping used to ensure the client data can be modeled. They were quick to share that the data quality is not the focus and they have to trust the client to send good data. Several respondents comment on the architecture, logical and physical data structures where data management is the practice of defining a standard, unifying model and language in order for the data to be defined, controlled and quality maintained. When asked; what is the role of data scientist as it relates to data management, it was consistent that it should not be their primary concern. They can influence, be a stakeholder, and offer opinions on the practice, but that it is a separate group responsible for administering the data management function in an organization. When asked about the maturity of their organization in the practice of data management, most folks replied that their organization is in the very early stages of defining this construct inside their business.

Surveys - Free Form Text

Regarding data management and what it means to the respondents and their organization, the simplest answers are “no idea, we don’t, not very salient, new program, not relevant, a bunch of buzzwords, overlooked and underused, and voodoo.”

Comments include:

- data management is tracing and managing the data used for different initiatives to help determine decisions
- prerequisite for survival and success in our competitive world; having information available for decision making stored in a secure but quickly accessible, proper format
- data management is key to good decisions on a tactical and strategic basis
- data management is the strategic and high quality way to move forward and it will evolve in the next three years across the organization…we are already exploring blockchain technology and hope to take more in the years to come about digital leader within the banking industry
- it is an untapped opportunity for better customer experience, enhanced operational efficiency, better product decisions, risk and regulatory compliance
- data management is the process of gathering, storing and retrieving information in timely, secured, precise and accurate manner to efficiently operate and cater to business needs
- data management is the process of taking care of data
- data management includes the security, safeguarding, and quality of the data
- making sure that information is captured in a way that is always stored consistently (fields always contain the same context/normalized); ensuring that the data is quickly available/extractable; right tools are available to consume/analyze the data
- while most data is free and openly available, knowledge of when it is acquired, if it has been ETL'd, and its location on the network is vital to future projects
- data that is acquired through subscription has an additional layer of work to stay aware of cost and when the subscription expires and whether it should be renewed

Out of the seventy-eight responses, the one that is most clear and concise is “managing the whole life cycle of data from the point of data creation until the data is purged, along with data governance, and data security.” What is interesting is that the phrase data governance is only mentioned four times in the surveys but is highlighted and sometimes used synonymously with data management in the interviews.

The last open ended, free form text question is to ask about the role of data management and data hygiene in their organizations. Twenty-three out of seventy-eight responses are

“we don’t, no idea, unknown, nonexistent - we only have dirty data and not much of a strategy.” Out of the remaining fifty-five other responses there is clear evidence that data management and data hygiene are concepts recognized in multiple industries.

Sample responses:

- we have a multilevel complex system of checks & balances to ensure data integrity
- data is being moved from various sites to a centralized warehouse for quick access to data to facilitate decision making process
- we are moving toward a golden source model but it is challenging given legacy systems
- our organization is actively planning or creating strategies / plans for handing the data created, stored, managed, and processed by our systems
- the data management strategy at my company manages terabytes of data and involves many data integration processes enabling business insights
- there is increased attention being given to more expansive data quality and data governance processes
- data hygiene is a critical component of a data management platform and strategy and must be adopted on load and post load

Strategic Decisions

Interviews

When asked about strategic decisions being made based on data results, the interviewees are able to proudly share that their organization in some capacity does rely on data for some key decisions. One interviewee commented that the firm they work at still has a population of people using their gut to make decisions. When asked why, the response is that they have not been exposed to data science or the decision being made does not have the data readily available, so the senior manager uses their expertise when making strategic decisions. Some types of decisions made by the firms represented are related to product improvements, sales, social demographics, risk reduction, major deals or purchases, and member surveys. Both qualitative and quantitative analysis techniques are referenced by the interviewee when describing their approach for the data analysis.

Surveys - Free Form Text

When asked how their organizations uses data as part of the decision making process, the respondents are very forthright. Five responses stated “not much, we don’t, don’t know or I heard that we use it but not exactly for what or how.” The other respondents are able to articulate their uses that range from determining output goals, to see if we are doing a good job, to benchmarking, budgets, sales reports and quantity of clients served. This indicates that an organization is likely to be new or a novice with data science practices. Other responses include uses such as evidenced based medicine (assuming medical decisions), detect fraud, and improve products and services, technology adoption rates, predictive return on investment, marketing information, student test scores, and sales projections. This is done in order to improve and focus attention on specific topics ranging from educating on concepts and skills, improving sales and services, effectiveness, and overall output goals.

Some of the key decisions made by the organizations represented by the respondents are related to new equipment purchases, staffing and hiring projections, recruiting students, budgets, bonuses, responding to customer surveys and investing or divesting products and services. Some respondents state that risk based decisions are made as it relates to fraud reducing risk and prevention of issues by identifying manual processes to improve.

Surveys - Multiple Choice (s)

In addition to the free form text survey questions, we ask; what types of decisions has your organization made using data?   The responses to this question are exciting to see as there are so many ways which organizations are actively using data for decisions. Marketing, promotion, risk/controls, and fraud prevention are the top areas organizations are making data driven decisions. Other answers are strategy, budgets, education/instructional decisions, and product enhancements. These responses are encouraging because it means we are moving in the right direction and using data for providing a foundation for logic versus instinct and intuition.

Data Challenges

Literature Review

The experts have a different position on the topic of pain-points in data science. Mellin (2013) states that embracing analytics is the key to excellence. He is a huge supporter of analysis and documents that unlock the knowledge from the data, depends on many factors; analytical tools and technologies are at the heart of this. The leader must both understand the current state of the organization and serve as a champion to advance internal analytical capabilities as simply demanding advanced analytics in an organization that is at an early stage of maturity will only lead to frustration. In addition, he states that methods to ensure consistent approach to applying models and repeating analysis across multiple locations requires strict governance.

Thomas Davenport's (2014) summary includes significant evidence exists that shows that companies have difficulties turning data into knowledge. Both systematic study and casual observation shows lack of data-derived knowledge and action across a number of different situations and environments. Many companies have enterprise resource planning systems, customer relationship management tools, point of sale scanner data, and web / e-commerce transaction data. Analytical capabilities are not used or, in many cases, available for use.

Another perspective is Saar-Tsechansky (2015) who states that what is needed from the Data Science IS community is two components:

- The review teams’ assessment of contribution needs to explicitly state with clarity allowing the review team to focus on establishing whether the claimed contributions are significant and on how convincingly the research establishes these contributions.

- In addition, the authors of data science contributions discuss the rationale for the choice of prior work against either empirical or analytical evaluations.

Booth and Hendrix (2015) wrote *Libraries and Institutional Data Analytics: Challenges and Opportunities*. In their article, they eloquently state the issues with devising a data driven culture. Their extensive thoughts have been synthesized here to focus on the relevant aspects as it relates to data science.

The authors list eight key challenge areas:

- culture - data driven cultures are not prevalent and are often full of mistrust of data, measures, analysis, and reporting

- talent - data scientists are lacking because skills and knowledge to be contained in one person is hard to find and the need for this type of talent is not recognized

- cost - data tools and capabilities are not cost savings but should be strategic investment

- data ownership - recommendation is for the data to be centralized versus remaining in the individual silos

- data quality - requirements to ensure data accuracy should be mandated; without this there are significant barriers to successful data analytics projects

- data standardization - organizations should be creating data dictionaries in order to make the analysis process more easily understood and performed

- data access - recommendation is that this is centralized and silos are removed

The last key topic related to challenges is that measures are not standardized. Cleveland (2014) does not empirically state anything specific to measure, however he is emphatic that we need a solution that will be based on statistics methods and models.

Matsudaira (2015) writes that there is a science to managing data science. She writes that as a VP of Engineering for a startup firm her job is to make sure the analytics are focused on the right things and that they are performing valuable work. What she has to do is make sure that they define simple concepts like the definition of done. In addition, the perception was that the data analytics team was busy but not communicating results effectively. Matsudaira put in place an approach to document findings in a consistent format. The teams receiving the work must shift their expectations and understand that done is not always elegant. Deadlines are applied to research, a backlog has been created track any uncompleted work.

Mondore, Spell and Douthitt (2016) document a heat map to deliver manager results and organize the business results among key drivers. They claim this allows managers to strategically focus on only the categories that they have areas for improvement and utilize resources in an effective manner in order to see return on investment from their actions.

Four quadrants verbatim from their document are:

- focus – any category that falls into this quadrant is scoring below the organizational average and is a significant driver of business outcomes

- promote – any category the leader is scoring well on and they are important drivers of business outcomes; this can be used to identify areas to congratulate teams, promote people, and celebrate wins

- monitor – any category the leader is scoring low on and is not a significant driver for the business.

- maintain – any category the leader is doing a great job and not impactful on business outcomes

What is evident are the limitations today’s business executives face are too numerous to count. Since data access, data tools and data techniques are deficient for today’s executive, they are forced to accept answers that are partly correct and leverage their instincts and intuition to answer critical questions that will have global implications.

Interviews

The big data challenges most prevalent are access to the data, manual processes required at this stage of maturity in an organization, age of the data, and senior buy-in on the findings. A bureaucracy that limits access to data is the biggest hurdle mentioned by the respondents to preventing forward movement. One candidate gave an excellent descriptive example - “consider if you are building a model or software solution that requires data from twelve data stores. The data is silo’d and if you want to develop machine learning capabilities, you need the data to be 1) accessible and 2) consistent across those data stores. Very few times will data experts find the data to be consistent and normalized so that a machine can be trained to leverage the data it is given." One respondent phrased it nicely; "what is lacking is precision in definition of the problem, definition of the data - there is no precision in data right now."

Oddly, what was not highlighted but is a well-known constraint, is that today's tools used in business are not set up for data analyst to perform data science tasks easily.

Surveys - Multiple Choice (s)

The responses to the same question in multiple choice (s) format are outlined in the table below. Because these issues are so prevalent and represent true pain points in organizations that we need to resolve. What this shows us is the same symptoms with an additional factor that highlights the number of responses across the seventy-eight responders. Unorganized data, too much data, and inaccessible data are crippling exploration and future innovation across global businesses. It is unfortunate that these pain points are not easily remedied and will require specific strategies, dedication, discipline and time in order to resolve.

Multiple Choice (s) Option	# of Responses	% of Responses
Too Much Data	29	37.2%
Unorganized / Not Normalized	42	53.8%
Not accessible	31	39.7%
Tools are not robust enough	23	29.5%
System crashes often	8	10.3%
Takes too much time to extract the data	27	34.6%
Other - Examples are Insufficient staff, too many databases, too normalized, limited applications	9	21.8%

Conclusion

Data Science is a multidisciplinary amalgam of techniques and tools that allow business organizations to work with data with robot-like precision, identify hidden patterns and build assumptions which lead to key decisions. It is in an evolutionary state with daily shifts as new technologies with inconceivable capabilities are developed and deployed. The exciting news is that it is being adopted and has a cohort of scientific-minded people with veracity who are dedicated to uncovering the hidden gems of knowledge captured in the massive data stores in data centers across the globe.

Ask any practitioner in the data field and they will wholeheartedly agree….likely they will say it is like pushing a boulder up a vertical mountain. From a practitioner's perspective, the measure of success for a data system and data science organization is that teams are able to effectively reply to data questions with speed and agility. They are able to translate complex results into actionable information and generate information that allows for critical business decisions to be made with accuracy and effectiveness. They are judged on their accuracy, reliability, and for having consistent results that strategic decisions are built upon….decisions that can in fact change the world on a daily basis.

The next research topic will be related to adoption of data science and data scientists into business organizations; primarily focusing on organization science and cultural adoption of data driven decision making and what it takes to deploy machine learning solutions.

The key information learned is that data science is both art and science. It is an aggregate of skills, techniques, and capabilities that are surprisingly agreed upon by academics and practitioners alike. The science, programming, and math components of data science are unanimously agreed upon as required by all respondents and research results. The art components are mentioned sparingly by the scientific community, but well understood and documented by the practitioners in the interviews and survey results.

Ultimately, what is easily determined by the findings, is that there are a consortium of people who are able to traverse the rough data landscape in today’s business world. They execute their projects and solve complex data problems with finesse even without a formal definition for the terms, data science and data scientists.

These are the people we should herald as modern-day explorers. They are reaching beyond common confines and pushing the boundaries of technical solutions and data constructs in order to transform the data science world. They are today’s pioneers and are taking data science forward to build a better tomorrow.

References

Aasheim, Cheryl L, Williams, Susan, Rutner, Paige, & Gardiner, Adrian. (2015). Data Analytics vs Data Science: A Study of Similarities and Differences in UnderGraduate Programs Based on Course Descriptions. *Journal of Information Systems Education, 26*(2).

Aliev, R. A., Pedrycz, Witold, Kreinovich, V., & Huseynov, O. H. (2016). The general theory of decisions. *Information Sciences, 327*, 125-148. doi: 10.1016/j.ins.2015.07.055

Bayrak, Tuncay. (2015). A Review of Business Analytics: A Business Enabler or Another Passing Fad. *Procedia - Social and Behavioral Sciences, 195*, 230-239. doi: 10.1016/j.sbspro.2015.06.354

Booth, Austin H., & Hendrix, Dean. (2015). Libraries and Institutional Data Analytics: Challenges and Opportunities. The Journal of Academic Librarianship, 41(5), 695-699. doi: 10.1016/j.acalib.2015.08.001

Brown, Brad, Court, David, & Willmott, Paul. (2013). Mobilizing the C Suite for Data Analytics. *McKinsey Quarterly*(4), 12.

Carbone, Anna, Jensen, Meiko, & Sato, Aki-Hiro. (2016). Challenges in data science: a complex systems perspective. *Chaos, Solitons & Fractals, 90*, 1-7. doi: 10.1016/j.chaos.2016.04.020

Cleveland, William S. (2014). Data science: An action plan for expanding the technical areas of the field of statistics. *Statistical Analysis and Data Mining: The ASA Data Science Journal, 7*(6), 414-417. doi: 10.1002/sam.11239

Cleveland, William S., & Hafen, Ryan. (2014). Divide and recombine (D&R): Data science for large complex data. *Statistical Analysis and Data Mining: The ASA Data Science Journal, 7*(6), 425-433. doi: 10.1002/sam.11242

Coderre, David. (2015). Gauge your Analytics. *Internal Auditor: Data Analytics*.

Davenport, Thomas H. (2014). *Big Data @ Work*. Massachusetts: Harvard Business School

Davenport, Thomas H., Harris, Jeanne G, DeLong, David W, & Jacobson, Alvin L. (2001). Data Driven to Knowledge: Building an Analytic Capability. *California Management Review, 43*(2).

Davenport, Thomas H. & Patel, D.J. (2012). *Data Scientist: The Sexiest Job of the 21st Century*. Harvard Business Review. Oct 2012.

Diggle, Peter J. (2015). Statistics: A Data Science for the 21st Century. *Royal Statistical Society, 178*(Part 4), 20.

Hayes, Brian. (2014). Doing Data Science, A Book Review. *Notices of the American Mathematical Society, 61*(09), 1068. doi: 10.1090/noti1167

Matsudaira, Kate. (2015). The Science Of Managing Data Science. *Communications of the ACM, 58*(6).

Mayhew, Helen., Saleh, Tamim., Williams, Simon. (2016). McKinsey & Company. Making Data Analytics Work for You Instead of the Other Way Around.

Mellin, Andrew. (2013). Embracing Analytics as the Key to Excellence. Physician Education Journal.

Mondore, Scott., Spell, Hannah., Douthitt, Shane. (2016). From the Boardroom to the Frontline: Prioritization and Practicality with Advanced Analytics. People-Strategy Journal Volume 39, Issue 2.

Myers, Kary and Vander Wiel, Scott. (2014). Discussion of Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics. Wiley Periodicals.

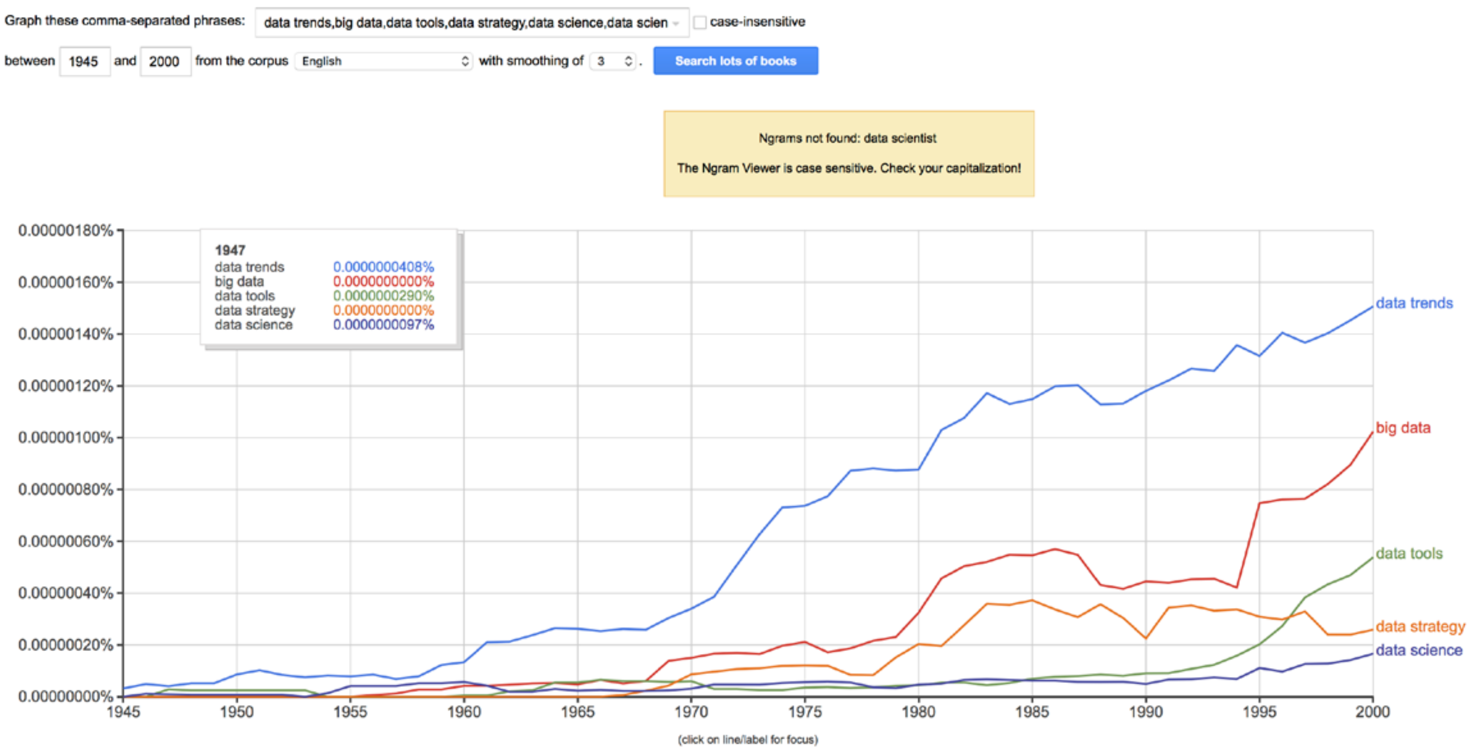
Royster, Sara. (2013). Working with Big Data. US Department of Labor

Saar-Tsechansky, Maytal. (2015). The Business of Business Data Science in IS Journals. *MIS Quarterly, 39*(4), iii-vi.

Sharma, Nitin. (2014). udacity.com. Ultimate Skills Checklist for your First Data Analytics Job.

Appendix A

Google NGRAM Results Graph



Graph these comma-separated phrases: data trends,big data,data tools,data strategy,data science,data scien □ case-insensitive

between 1945 and 2000 from the corpus English with smoothing of 3 .

Ngrams not found: data scientist

The Ngram Viewer is case sensitive. Check your capitalization!

1947	
data trends	0.0000000408%
big data	0.0000000000%
data tools	0.0000000290%
data strategy	0.0000000000%
data science	0.0000000097%

Appendix B

IRB Approval Letters

USF
UNIVERSITY OF
SOUTH FLORIDA

RESEARCH INTEGRITY AND COMPLIANCE
Institutional Review Boards, FWA No. 00001669
12901 Bruce B. Downs Blvd., MDC035 • Tampa, FL 33612-4799
(813) 974-5638 • FAX (813) 974-7091

June 8, 2017

Dana Parks
College of Business Administration
Tampa, FL 33612

RE: **Expedited Approval for Initial Review**
IRB#: Pro00030443
Title: Data Driven Decisions – Organization Attributes Required to Adopt Data Science and Effect Change (Interviews)

Study Approval Period: 6/7/2017 to 6/7/2018

Dear Ms. Parks:

On 6/7/2017, the Institutional Review Board (IRB) reviewed and **APPROVED** the above application and all documents contained within, including those outlined below.

Approved Item(s):
Protocol Document(s):
Research Questionnaire Protocol in word Interview.docx

Consent/Assent Document(s)*:
Adult Consent, #1.pdf.pdf

*Please use only the official IRB stamped informed consent/assent document(s) found under the "Attachments" tab. Please note, these consent/assent documents are valid until the consent document is amended and approved.

It was the determination of the IRB that your study qualified for expedited review which includes activities that (1) present no more than minimal risk to human subjects, and (2) involve only procedures listed in one or more of the categories outlined below. The IRB may review research through the expedited review procedure authorized by 45CFR46.110. The research proposed in this study is categorized under the following expedited review category:

(6) Collection of data from voice, video, digital, or image recordings made for research purposes.

(7) Research on individual or group characteristics or behavior (including, but not limited to, research on perception, cognition, motivation, identity, language, communication, cultural beliefs or practices, and social behavior) or research employing survey, interview, oral history, focus group, program evaluation, human factors evaluation, or quality assurance methodologies.

As the principal investigator of this study, it is your responsibility to conduct this study in accordance with IRB policies and procedures and as approved by the IRB. Any changes to the approved research must be submitted to the IRB for review and approval via an amendment. Additionally, all unanticipated problems must be reported to the USF IRB within five (5) calendar days.

We appreciate your dedication to the ethical conduct of human subject research at the University of South Florida and your continued commitment to human research protections. If you have any questions regarding this matter, please call 813-974-5638.

Sincerely,

John A Schinka, Ph.D.

John Schinka, Ph.D.,
Chairperson USF
Institutional Review Board

USF
UNIVERSITY OF
SOUTH FLORIDA

RESEARCH INTEGRITY AND COMPLIANCE
Institutional Review Boards, FWA No. 00001669
12901 Bruce B. Downs Blvd., MDC035 • Tampa, FL 33612-4799
(813) 974-5638 • FAX (813) 974-7091

6/16/2017

Dana Parks
College of Business Administration
5405 Coral Shell Way
Apollo Beach, FL 33572

RE: **Exempt Certification**
IRB#: Pro00030537
Title: SURVEY Data Driven Decisions - Data Scientist Skills, Taxonomy and Competencies

Dear Ms. Parks:

On 6/14/2017, the Institutional Review Board (IRB) determined that your research meets criteria for exemption from the federal regulations as outlined by 45CFR46.101(b):

(2) Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures or observation of public behavior, unless:
(i) information obtained is recorded in such a manner that human subjects can be identified, directly or through identifiers linked to the subjects; and (ii) any disclosure of the human subjects' responses outside the research could reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, or reputation.

As the principal investigator for this study, it is your responsibility to ensure that this research is conducted as outlined in your application and consistent with the ethical principles outlined in the Belmont Report and with USF HRPP policies and procedures.

Please note, as per USF HRPP Policy, once the Exempt determination is made, the application is closed in ARC. Any proposed or anticipated changes to the study design that was previously declared exempt from IRB review must be submitted to the IRB as a new study prior to initiation of the change. However, administrative changes, including changes in research personnel, do not warrant an amendment or new application.

Given the determination of exemption, this application is being closed in ARC. This does not limit your ability to conduct your research project.

We appreciate your dedication to the ethical conduct of human subject research at the University of South Florida and your continued commitment to human research protections. If you have any questions regarding this matter, please call 813-974-5638.

Sincerely,

John A Schinka, Ph.D.

John Schinka, Ph.D.,
Chairperson USF Institutional
Review Board

Vita

Dana Parks is a full time employee at a global corporate bank leading strategic programs. She also works on developing data solutions and leading a team of analysts to perform analytics on large datasets. She is working closely with Information Architects to build a cutting edge platform to deliver full analysis and analytics capabilities to a large population of users. These efforts include reducing user tools in existence today and develop new, specific tools for big data analytics focused on fraud controls, risk controls and detecting anomalies in client behaviors. In her career, Dana has performed a variety of roles in financial services from retail banking to banking operations and technology.

Most recently, Dana has led multiple technology organizations building sets of user tools and manages large portfolios of projects and maintaining large applications. She is a graduate of Saint Leo University where she obtained her Bachelor of Science degree in Computer Information Systems with a minor in Business Management in 2004 and her Master's in Business Administration in 2009.

Her passion is driving positive change that will improve global operations and technology teams with new processes in order to optimize controls and develop solutions to streamline functions.