RESEARCH PAPER

Topic: Data Science

SUBJECT: Software Project Management

CLASS: BSSE-3RD YEAR

SECTION: A

Abstract:

### ***Data science education is well into its formative stages of development; it is evolving into a self-supporting discipline and producing professionals with distinct and complementary skills relative to professionals in the computer, information, and statistical sciences.***

GROUP MEMBERS:

Areeba Shahzad(B16158013)

Musfira Zia (B16158065)

musfira zia

#### **A Brief History of Data Science**

The term data science has existed for the better part of the last 30 years and was originally used as a substitute for "computer science" in 1960. Approximately 15 years later, the term was used to define the survey of data processing methods used in different applications. In 2001, data science was introduced as an independent discipline. The Harvard Business Review published an [article](https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century) in 2012 describing the role of the data scientist as the “sexiest job of the 21st century.”

***DATA SCIENCE:***

Data science is a [multi-disciplinary](https://en.wikipedia.org/wiki/Multi-disciplinary) field that uses scientific methods, processes, algorithms and systems to extract [knowledge](https://en.wikipedia.org/wiki/Knowledge) and insights from structured and [unstructured data](https://en.wikipedia.org/wiki/Unstructured_data). Data science is the same concept as [data mining](https://en.wikipedia.org/wiki/Data_mining) and [big data](https://en.wikipedia.org/wiki/Big_data): "use the most powerful hardware, the most powerful programming systems, and the most efficient algorithms to solve problems".

Data science is a "concept to unify [statistics](https://en.wikipedia.org/wiki/Statistics), [data analysis](https://en.wikipedia.org/wiki/Data_analysis), [machine learning](https://en.wikipedia.org/wiki/Machine_learning) and their related methods" in order to "understand and analyze actual phenomena" with data. It employs techniques and theories drawn from many fields within the context of [mathematics](https://en.wikipedia.org/wiki/Mathematics), [statistics](https://en.wikipedia.org/wiki/Statistics), [computer science](https://en.wikipedia.org/wiki/Computer_science), and [information science](https://en.wikipedia.org/wiki/Information_science).

Data science is the study of [data](https://techterms.com/definition/data). It involves developing methods of recording, storing, and analyzing data to effectively extract useful information. The goal of data science is to gain insights and knowledge from any type of data both structured and unstructured.

Data science is related to [computer science](https://techterms.com/definition/computer_science), but is a separate field. Computer science involves creating [programs](https://techterms.com/definition/program) and [algorithms](https://techterms.com/definition/algorithm) to record and process data, while data science covers any type of data analysis, which may or may not use computers. Data science is more closely related to the mathematics field of Statistics, which includes the collection, organization, analysis, and presentation of data.

Because of the large amounts of data modern companies and organizations maintain, data science has become an integral part of [IT](https://techterms.com/definition/it). For example, a company that has [petabytes](https://techterms.com/definition/petabyte) of user data may use data science to develop effective ways to store, manage, and analyze the data. The company may use the scientific method to run tests and extract results that can provide meaningful insights about their users.

Data science is the study of where information comes from, what it represents and how it can be turned into a valuable resource in the creation of business and [IT strategies](https://searchcio.techtarget.com/definition/IT-strategy-information-technology-strategy). It identify patterns can help an organization rein in costs, increase efficiencies, recognize new market opportunities and increase the organization's competitive advantage.

The data science field employs mathematics, statistics and  computer science disciplines, and incorporates techniques like [machine learning](https://searchenterpriseai.techtarget.com/definition/machine-learning-ML), cluster analysis, [data mining](https://searchsqlserver.techtarget.com/definition/data-mining) and visualization.

### Understanding Data Science is like Data is drawn from different sectors, channels, and platforms including cell phones, social media, e-commerce sites, healthcare surveys, and Internet searches. The increase in the amount of data available opened the door to a new field of study based on big data. The massive data sets that contribute to the creation of better operational tools in all [sectors](https://www.investopedia.com/terms/s/sector.asp).

The continually increasing access to data is possible due to advancements in technology and collection techniques. Individuals buying patterns and behavior can be monitored and predictions made based on the information gathered.

However, the ever-increasing data is unstructured and requires parsing for effective decision making. This process is complex and time-consuming for companies hence, the emergence of data science.

Data science combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data. Data science practitioners apply [machine learning](https://www.datarobot.com/wiki/machine-learning/) [algorithms](https://www.datarobot.com/wiki/algorithm/) to numbers, text, images, video, audio, and more to produce [artificial intelligence (AI)](https://www.datarobot.com/wiki/artificial-intelligence/) systems to perform tasks that ordinarily require human intelligence. In turn, these systems generate [insights](https://www.datarobot.com/wiki/insights/) which analysts and business users can translate into tangible business value.

#### **How Data Science Is Applied**

Data science incorporates tools from multiple disciplines to gather a data set, process, and derive insights from the data set, extract meaningful data from the set, and interpret it for decision-making purposes. The disciplinary areas that make up the data science field include mining, statistics, machine learning, analytics, and programming.

[Data mining](https://www.investopedia.com/terms/d/datamining.asp) applies algorithms to the complex data set to reveal patterns that are then used to extract useful and relevant data from the set. Statistical measures or [predictive analytics](https://www.investopedia.com/terms/p/predictive-analytics.asp) use this extracted data to gauge events that are likely to happen in the future based on what the data shows happened in the past.

Machine learning is an artificial intelligence tool that processes mass quantities of data that a human would be unable to process in a lifetime. Machine learning perfects the decision model presented under predictive analytics by matching the likelihood of an event happening to what actually happened at a predicted time.

Using analytics, the data analyst collects and processes the structured data from the machine learning stage using [algorithms](https://www.investopedia.com/terms/a/algorithm.asp). The analyst interprets, converts, and summarizes the data into a cohesive language that the decision-making team can understand. Data science is applied to practically all contexts and, as the data scientist's role evolves, the field will expand to encompass data architecture, data engineering, and data administration.

*“The ability to take data to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it that’s going to be a hugely important skills in the next decades.”*

* Hal Varian, chief economist at Google and UC Berkeley professor of information sciences, business and economics.

Effective data scientists are able to identify relevant questions, collect data from a multiple of different data sources, organize the information, translate results into solutions and communicate their findings in a way that positively affects business decision. These skills are required in almost all industries causing skilled data scientists to be increasingly valuable to companies.

Data science continues to evolve as one of the most promising and in demand career paths for skilled professionals. Today successful data professionals understand that they must advance the traditional skills of analyzing large amounts of data, data mining and programming skills. In order to uncover useful intelligence for their organizations, data scientists must master the full spectrum of the data science life cycle and possess a level of flexibility and understanding to maximize returns at each phase of the process

*“Data is everywhere and expansive. A variety of terms related to mining, cleaning, analyzing and interpreting data are often used interchangeably but they can actually involve different skill sets and complexity of data”.*

As the amount of data generated by the typical modern business increases, so does the prominence of data scientists hired by organizations to help them turn [raw data](https://searchdatamanagement.techtarget.com/definition/raw-data) into valuable business [information](https://searchsqlserver.techtarget.com/definition/information). Data extraction is the act of retrieving specific data from unstructured or poorly structured data sources for further processing and investigation. Data scientists must possess a combination of analytic, machine learning, data mining and statistical skills, as well as experience with algorithms and coding. Along with managing and interpreting large amounts of data, many data scientists are also tasked with creating data visualization models that help illustrate the business value of digital information. To be effective, however, data scientists must possess [emotional intelligence](https://searchcio.techtarget.com/definition/emotional-intelligence) in addition to education and experience in [data analytics](https://searchdatamanagement.techtarget.com/definition/data-analytics). Perhaps the most important skill a data scientist must possess is the ability to present the data insights to others, including C-suite executives, and explain the significance of the data in a way that can be easily understood.

Data scientists draw the digital information they are studying from a growing list of channels and sources, including smartphones, internet of things [(IoT) devices](https://internetofthingsagenda.techtarget.com/definition/IoT-device), social media, surveys, purchases, and internet searches and behavior. By sorting through these large data sets, data scientists can identify patterns to solve problems through data analysis -- a process known as data mining.

***Skills required to become data science***

Programing skills (SAS,R,Python), statistical and mathematical skills, storytelling and data visualization, Hadoop,SQL, Machine Learning

***What does a Data Scientist do?***

In the past decades, data scientist have become necessary assets and are present in almost all organizations. These professionals are well-rounded data driven individuals with high level technical skills who are capable of buildings complex quantitative algorithms to organize and synthesize large amounts of information used to answer questions and drive strategy in their organization. This is coupled with the experience in communication and leadership needed to deliver tangible results to various stakeholders across an organization or business.

Data scientists need to be curious and result oriented with exceptional industry specific knowledge and communication skills that allow them to explain highly technical results to their non-technical counterparts. They possess a strong quantitative background in statistics and linear algebra as well as programming knowledge with focuses in data warehousing, mining and modeling to build and analyze algorithms. They must also be able to utilize key technical tools and skills including

* R
* Python
* Apache Hadoop
* MapReduce
* Apache Spark
* NoSQL databases
* Cloud Computing
* D3
* Apache Pig
* Tableau
* IPython Notebooks
* Github

***Data Analyst***

Data Analyst bridge the gap between Data Scientists and Business Analyst. They are provided with the questions that need answering from an organization and then organize and analyze data to find results that align with high level business strategy. Data Analysts are responsible for translating technical analysis to quantitative action items and effectively communicating their findings to diverse stakeholders.

***Data Engineer***

Data Engineers manage exponential amounts of rapidly changing data. They focus on the development, management and optimization of data pipelines and infrastructure to transform and transfer data to Data Scientists for querying.

***The Data Science Life Cycle***

The Data Science Life Cycle depend on five stages

* Capture (data acquisition, data entry, single reception, data extraction)
* Maintain (data warehousing, data cleansing, data staging, data processing, data architecture)
* Process (data mining, clustering / classification, data modeling, data summarization)
* Analyze (exploratory / confirmatory, predictive analysis, regression, text mining, qualitative analysis)
* Communicate (data reporting, data visualization, business intelligence, decision making)

## ***Why Data Science is Important?***

More and more companies are coming to realize the importance of data science, AI, and machine learning. Regardless of industry or size, organizations that wish to remain competitive in the age of [big data](https://www.datarobot.com/wiki/big-data/) need to efficiently develop and implement data science capabilities or risk being left behind.

### *Benefits of Data Science*

The main advantage of enlisting data science in an organization is the empowerment and facilitation of decision-making. Organizations with data scientists can factor in quantifiable, data-based evidence into their business decisions. These [data-driven decisions](https://whatis.techtarget.com/definition/data-driven-decision-management-DDDM) can ultimately lead to increased profitability and improved operational efficiency, business performance and workflows. In customer-facing organizations, data science helps identify and refine target audiences. Data science can also assist recruitment: Internal processing of applications and data-driven aptitude tests and games can help an organization's human resources team make quicker and more accurate selections during the hiring process.

The specific benefits of data science vary depending on the company's goal and the industry. Sales and marketing departments, for example, can mine customer data to improve conversion rates or create one-to-one marketing campaigns. Banking institutions are mining data to enhance fraud detection. Streaming services like Netflix mine data to determine what its users are interested in, and use that data to determine what TV shows or films to produce. Data-based algorithms are also used at Netflix to create personalized recommendations based on a user's viewing history. Shipment companies like DHL, FedEx and UPS use data science to find the best delivery routes and times, as well as the best modes of transport for their shipments.

Data science is still an emerging field within the enterprise because the identification and analysis of vast amounts of [unstructured data](https://searchbusinessanalytics.techtarget.com/definition/unstructured-data) can prove too complex, expensive and time-consuming for companies.

### *Data Science and Machine Learning*

Machine learning is often incorporated in data science. Machine learning is an [artificial intelligence](https://searchenterpriseai.techtarget.com/definition/AI-Artificial-Intelligence) (AI) tool that essentially automates the data-processing portion of data science. Machine learning integrates advanced algorithms that learn on their own and can process massive amounts of data in a fraction of the time it would take a human. After collecting and processing the structured data from the machine learning tools, data scientists interpret, convert and summarize the data so it is useful for the company's decision-makers.

Machine learning applications used in the data science field include image recognition and [speech recognition](https://searchcrm.techtarget.com/definition/speech-recognition). Machine learning algorithms are also being integrated into self-driving vehicles.

## ***Data Science and Data Robot***

Ramping up data science efforts is difficult even for companies with near-unlimited resources. The Data Robot [automated machine learning](https://www.datarobot.com/wiki/automated-machine-learning/) platform democratizes data science and AI, enabling analysts, business users, and other technical professionals to become [Citizen Data Scientists](https://www.datarobot.com/wiki/citizen-data-scientist/) and [AI Engineers](https://www.datarobot.com/wiki/ai-engineer/), in addition to making data scientists more productive. It automates repetitive [modeling](https://www.datarobot.com/wiki/model/) tasks that once occupied the vast majority of data scientists’ time and brainpower. Data Robot bridges the gap between data scientists and the rest of the organization, making enterprise machine learning more accessible than ever.

### *Data Science vs Data Mining*

Data science is often confused with [data mining](https://techterms.com/definition/data_mining). However, data mining is a subset of data science. It involves analyzing large amounts of data (such as [big data](https://techterms.com/definition/big_data)) in order to discover patterns and other useful information. Data science covers the entire scope of data collection and processing.

Data mining is the Raw data that can be unstructured and messy, with information coming from disparate data sources, mismatched or missing records, and a slew of other tricky issues. Data munging is a term to describe the data wrangling to bring together data into cohesive views, as well as the janitorial work of cleaning up data so that it is polished and ready for downstream usage. This requires good pattern-recognition sense and clever hacking skills to merge and transform masses of database-level information. If not properly done, dirty data can obfuscate the 'truth' hidden in the data set and completely mislead results. Thus, any data scientist must be skillful and nimble at data munging in order to have accurate, usable data before applying more sophisticated analytical tactics.

# ***Big Data vs Smart Data***

Big Data describes massive amounts of data, both unstructured and structured, that is collected by organizations on a daily basis. This [Big Data](http://www.dataversity.net/what-is-big-data/) can then be filtered, and turned into Smart Data before being analyzed for insights, in turn, leading to more efficient decision-making. Smart Data can be described as Big Data that has been cleansed, filtered, and prepared for context.

There are two primary kinds of Smart Data often discussed by experts in the industry. One form is information picked up by a sensor, and then sent to a nearby collection point, and acted upon, before being sent to an [Analytics](http://www.dataversity.net/what-is-analytics/) platform. This data is sourced from Smart Sensors, especially within the Industrial Internet of Things (IIoT) systems. The other kind of Smart Data is Big Data that has been processed and is waiting to be turned into actionable information. For purposes of this article, data going to, and from, a Smart Sensor is “sensor data.” The term, [Smart Data](https://www.business2community.com/big-data/forget-big-data-think-smart-data-01868818), will refer to Big Data that has been screened for useful information.

Smart Data is a new tool for business. Big Data gets turned into Smart Data when it is collected and optimized, using the specific needs of the industry and the individual organization. The following areas are some of the use cases for Smart Data:

* [Customer Journey Analytics](https://www.pointillist.com/blog/what-is-customer-journey-analytics/) weaves hundreds of customer internet interactions together from across multiple channels. It combines thousands of events to create a journey for a business’ customers. It is a data-driven approach used to discover, analyze, and influence the customers’ journey. (However, when the information is “wrong,” it is both irritating and insulting, and may cause the loss of a customer.)
* The Customer Experience analysis (or [Voice of the Customer Analytics](https://www.gartner.com/smarterwithgartner/use-voice-of-customer-data-to-improve-customer-experience-analytics/)) uses tools and techniques to gather the customer’s attitudes, opinions, and emotions. Voice of the Customer Analytics emphasizes the mental state of customers. Other forms of Analytics normally focus on a customer’s actions and behavior, rather than their thoughts. Marketing organizations will often use this kind of analysis to manage reputations, manage products, and provide competitive Business Intelligence. Techniques for collecting this kind of information include short surveys and comprehensive software platforms.

***Smart Data and the Five Vs***

Big Data is commonly described as using the five Vs: value, variety, volume, velocity, veracity. A reduction in “volume” takes place with Smart Data. Only useful information for solving the problem is presented. Variety may, or may not, be reduced, depending on the screening process used in filtering the data. Value, velocity, and veracity (accuracy) should all increase with the decrease in volume.

***Machine Learning and Smart Data***

[Machine Learning](http://www.dataversity.net/machine-learning-vs-deep-learning/) is often a training process for Artificial Intelligence platforms, but can also be used as a recognition and decision-making program. As the use and popularity of Smart Data has increased, it has also been [used with Machine Learning](https://siliconangle.com/blog/2017/11/14/panoply-accelerates-data-integration-smart-data-warehouse/) algorithms designed to seek out Business Intelligence and insights. Machine Learning allows organizations to filter Data Lakes and [Data Warehouses](http://www.dataversity.net/data-warehouse-vs-data-lake-technology-different-approaches-managing-data/), creating Smart Data in the process.

Traditionally, organizations seeking Business Intelligence from Big Data have used Data Scientists, who spend time searching for insights and patterns within an enterprise’s datasets. [Machine Learning algorithms](http://www.dataversity.net/data-scientists-machine-learning-algorithms-data-driven-world/) using “Unsupervised Learning,” and combined with Big Data, have made it possible to perform Data Analytics more quickly, and without the Data Scientist. Machine Learning algorithms dramatically increase the accuracy, speed, and intelligence of screening Big Data, and with feedback, can continue to learn and refine the “filtering process.”

***Artificial Intelligence and Smart Data***

During the screening and filtering process of creating Smart Data, decisions are made as to which data should be blocked, and which should be presented. Machine Learning and [Artificial Intelligence](https://www.information-management.com/opinion/smart-data-gives-artificial-intelligence-meaning-in-2018) (AI)  use specific criteria during this process. AI is an ongoing attempt to create intelligence within machines, allowing them to work and respond like humans. Artificial Intelligence has provided flexibility and can address unique goals. For example, financial services firms can use AI-driven Smart Data for customer analysis, fraud detection, market analysis, and compliance.

***Smart Data and Big Data Use Cases***

* Smart Data and Beer
* Smart Data and Healthcare
* **Collecting Smart Data**

# ***Data Modeling Trends***

The Database Management and Data Modeling landscapes have evolved significantly in the past few years, from the traditional relational model to now include non-relational models as well. The growth of Big Data and its unstructured and semi-structured data formats, along with trends in Cloud Computing, Artificial Intelligence, Data Lakes, [Machine Learning](http://www.dataversity.net/what-is-machine-learning/), Blockchain and others pushing the need for more advanced concepts and practices.

Such a monumental shift has caused [Data Modeling](http://www.dataversity.net/what-is-data-modeling/) to also advance. The fundamental changes in data infrastructure and newer technology evolutions have together contributed to this Data Management development.

### here are some other significant Data Modeling trends that will become apparent

### **Automated Data Modeling (Algorithms)**

### **Predictive Modeling**

### **Semantic Data Models**

Some of those advantages for the average business will gain through enhanced Data Modeling are:

1. The Data Model provides a [clear framework](http://www.dataversity.net/enterprise-data-modeling/) for development projects through the blueprint, thus business clients and developers get more clarity on agreed terms.
2. When the Data Model contains crisp guidelines, then it enables high performance levels. The guidelines help in resolving confusion during the development phase.
3. Corrupt datasets get detected and cleansed during the Data Modeling stage, which is another big advantage for developers.
4. Data Models offer a tested “blueprint” for building software, which great reduces the possibility of later development errors, and greatly reduces development cost and time resulting in shorter time to market (almost 70 percent of initial coding budget and allocated time are reduced).
5. The Data Model [clearly outlines both](http://www.dataversity.net/recent-innovation-nosql-data-modeling-forward-reverse-engineering-combined/) the “scope” and the associated “risks” of a managed development effort. Thus, the risks can be mitigated early on by appropriately scaling the model up or down.
6. As Data Models include detailed documentation, long-range maintenance becomes easy and transparent even when staff changes. The [documentation also](http://www.dataversity.net/facilitating-collaborative-data-governance-enterprise-data-architecture-data-modeling/) serves as the starting point of advanced data analysis.

# ***Data Science vs Business Intelligence***

**i**t is important to begin with some basic definitions of the two terms, taking a deeper look at the two distinct (though closely allied) fields within Data Analytics. Data Science, as used in business, is intrinsically data-driven, where many interdisciplinary sciences are applied together to extract meaning and insights from available business data, which is typically large and complex. On the other hand, Business Intelligence or BI helps monitor the current state of business data to understand the historical performance of a business.

So, in nutshell, while BI helps interpret past data, Data Science can analyze the past data (trends or patterns) to make future predictions. BI is mainly used for reporting or [Descriptive Analytics](http://www.dataversity.net/fundamentals-descriptive-analytics/); whereas Data Science is more used for [Predictive Analytics](http://www.dataversity.net/fundamentals-predictive-analytics/) or [Prescriptive Analytics](http://www.dataversity.net/fundamentals-prescriptive-analytics/).

### *****CONCLUSION*****

### **The following is a summary of the findings and recommendations discussed in this report**

For any company that wishes to enhance their business by being more data-driven, data science is the secret sauce. Data science projects can have multiplicative returns on investment, both from guidance through data insight, and development of data product. There is simply not enough supply of data scientists in the market to meet the demand ([data scientist salary](https://datajobs.com/big-data-salary) is sky high)

Today, enterprises are frequently at a loss about how to keep pace with the speed of technological change, and how to integrate newer and better capabilities with the existing ones. For example, advanced technologies such as Big Data, IoT, Machine Learning, and [Serverless Computing](http://www.dataversity.net/serverless-computing-serverless-architecture-overview/) can together transform the business landscape.

### **Data science education is well into its formative stages of development; it is evolving into a self-supporting discipline and producing professionals with distinct and complementary skills relative to professionals in the computer, information, and statistical sciences**

### **Today, the nation is in the formative phase of data science education, where educational organizations are pioneering their own programs, each with different approaches to depth, breadth, and curricular emphasis (e.g., business, computer science, engineering, information science, mathematics, social science or statistics). It is too early to expect consensus to emerge on certain best practices of Data Science education. However, it is not too early to envision the possible forms that such practices might take. Nor it is too early to make recommendations that can help the Data Science education community develop strategic vision and practices**

### *****REFRENCES*****

1. Press, Gil. [*"A Very Short History Of Data Science"*](https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/).
2. Wu, C. F. J. (1997). [*"Statistics = Data Science?"*](http://www2.isye.gatech.edu/~jeffwu/presentations/datascience.pdf) *(PDF). Retrieved 9 October 2014*
3. [*"IEEE Task Force on Data Science and Advanced Analytics"*](http://www.dsaa.co/).
4. [*"2014 IEEE International Conference on Data Science and Advanced Analytics"*](https://web.archive.org/web/20170329172857/http:/datamining.it.uts.edu.au/conferences/dsaa14/). Archived from [*the original*](http://datamining.it.uts.edu.au/conferences/dsaa14/) on 29 March 2017
5. [*"Journal on Data Science and Analytics"*](https://www.springer.com/41060).
6. [**^**](https://en.wikipedia.org/wiki/Data_science#cite_ref-NateSilver_26-0) [*"Nate Silver: What I need from statisticians"*](http://www.statisticsviews.com/details/feature/5133141/Nate-Silver-What-I-need-from-statisticians.html). 23 August 2013.