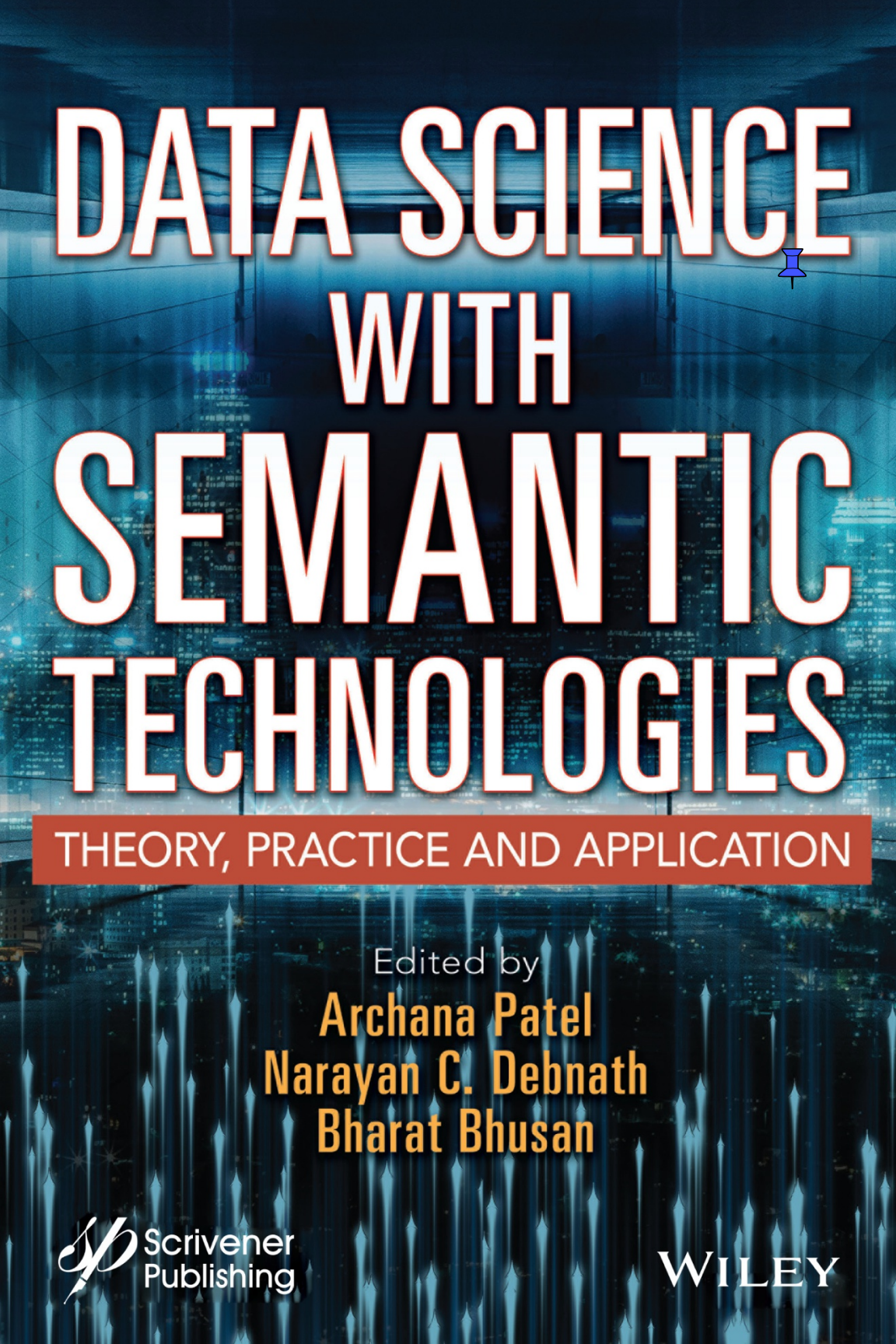
DATA SCIENCE
WITH
SEMANTIC
TECHNOLOGIES
THEORY, PRACTICE AND APPLICATION
Edited by
Archana Patel
Narayan C. Debnath
Bharat Bhusan
Scrivener Publishing
WILEY

Data Science with Semantic Technologies

Scrivener Publishing
100 Cummings Center, Suite 541J
Beverly, MA 01915-6106

Publishers at Scrivener
Martin Scrivener (martin@scrivenerpublishing.com)
Phillip Carmical (pcarmical@scrivenerpublishing.com)

Data Science with Semantic Technologies

Theory, Practice, and Application

Edited by

Archana Patel

Department of Software Engineering, School of Computing and Information Technology, Eastern International University, Vietnam

Narayan C. Debnath

School of Computing and Department of Computer Science and Engineering, School of Engineering Vietnam

and

Bharat Bhusan

Technology, Sharda University Information Technology, India

WILEY

This edition first published 2022 by John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA and Scrivener Publishing LLC, 100 Cummings Center, Suite 541J, Beverly, MA 01915, USA

For more information about Scrivener publications please visit www.scrivenerpublishing.com.

Wiley Global Headquarters
111 River Street, Hoboken, NJ 07030, USA

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Library of Congress Cataloging-in-Publication Data

ISBN 9781119864981

Cover image: PixaBay.Com
Cover design by Russell Richardson

Set in size of 11pt and Minion Pro by Manila Typesetting Company, Makati, Philippines

Printed in the USA

10 9 8 7 6 5 4 3 2 1

Contents

Preface

Data Science is an invaluable resource that deals with vast volumes of data using modern tools and techniques to find unseen patterns, derive meaningful information, and make business decisions. To create intelligence in data science, it becomes necessary to utilize the semantic technologies which allow machine-readable representation of data. This intelligence uniquely identifies and connects data with common business terms, and also enables users to communicate with data. Instead of structuring the data, semantic technologies help users to understand the meaning of the data by using the concepts of semantics, ontology, OWL, linked data, and knowledge graphs. These technologies assist organizations in understanding all of the stored data, adding value to it, and enabling insights that were not available before. Organizations are also using semantic technologies to unearth precious nuggets of information from vast volumes of data and to enable more flexible use of data. These technologies can deal with the existing problems of data scientists and help them in making better decisions for any organization. All of these needs are part of a focused shift towards utilization of semantic technologies in data science that provide knowledge along with the ability to understand, reason, plan, and learn with existing and new data sets. These technologies also generate expected, reproducible, user-desired results.

This book aims to provide a roadmap for the deployment of semantic technologies in the field of data science. Moreover, it highlights how data science enables the user to create intelligence through these technologies by exploring the opportunities and eradicating the challenges in current and future time frames. It can serve as an important guide to applications of data science with semantic technologies for the upcoming generation and thus is a unique resource for scholars, researchers, professionals and practitioners in this field. Following is a brief description of the subjects covered in the 15 chapters of the book.

– Chapter 1 provides a brief introduction to data science. It addresses various aspects of data science such as what a data scientist does and why data science is in demand; the history of data science and how it differs from business intelligence; the life cycle of data science and data science components; why data science is important; the challenges of data science; the tools used for data science; and the benefits and applications of data science.

– Chapter 2 provides an overview of the top 10 tools and applications that should be of interest to any data scientist. Its objective includes, but is not limited to, realizing the use of Python in developing solutions to data science tasks; recognizing the use of R language as an open-source data science provider; traveling around the SQL to provide structured models for data science projects; navigating through data analytics and statistics using Excel; and using D3.js scripting tools for data visualization. Also, practical examples/case studies are provided on data visualization, data analytics, regression, forecasting, and outlier detection.

– Chapter 3 presents the use of data modeling for data science, revealing the possibility of a new side of the data. The chapter covers different types of data (unstructured data, semi-structured data, structured data, hybrid (un/semi)-structured data and big data) and data model design.

– Chapter 4 shows data management by considering language based on the novelty view of data. The chapter focuses on data life cycle, data distribution and CAP theorem.

– Chapter 5 presents the role of data science in healthcare. There are several fields in the healthcare sector, such as predictive modeling, genetics, etc., which make use of data science for diagnosis and drug discovery, thereby increasing usability of precision medicine.

– Chapter 6 provides a new balanced binary search tree that generates two kinds of nodes: simple and class nodes. Two advantages make the new structure attractive. First, it subsumes the most popular data structures of AVL and Red-Black trees. Second, it proposes other unknown balanced binary search trees in which we can adjust the maximal height of paths between 1.44 lg(n) and 2 lg(n), where n is the number of nodes in the tree and lg the base-two logarithm.

– Chapter 7 shows the study of machine learning and deep learning algorithms with detailed and analytical comparisons, which help new and inexperienced medical professionals or researchers in the medical field. The proposed machine learning model has an accurate algorithm that works with rich healthcare data, a high-dimensional data handling system, and an intelligent framework that uses different data sources to predict

heart disease. This chapter uses an ensemble-based deep learning model with optimal feature selection to improve accuracy.

– Chapter 8 presents an IoT-based automated fire control system in a mining area which will help to protect many valuable lives whenever an accident occurs due to fire. In the experimental application, different types of sensors for temperature, moisture, and gas are used to sense the different environmental data.

– Chapter 9 offers an aspect identification method for sentiment sentences in review documents. The chapter describes two key tasks—one for extracting significant features from the reviews and another for identification of degrees of product reviews.

– Chapter 10 shows the research that paved the way for semantic technology. It then describes each of the semantic pillars with examples and explanations of the business value of each technology.

– Chapter 11 describes the ontology evaluation tools and then focuses on the evaluation of the security ontologies. The existing ontology evaluation tools are classified under two categories; namely, domain-dependent ontology evaluation tools and domain-independent ontology evaluation tools. The evaluation of security ontology assesses the quality of ontology among the available ontologies.

– Chapter 12 discusses the main concepts of health data, data science, health data science, examples of the application of health data science and related challenges. In addition, it also highlights the application of semantic technologies in health data science and the challenges that lie ahead of using these technologies.

– Chapter 13 proposes an original hybrid optimization approach based on two different mixed integer programming statements. The first statement is based on minimizing the sum of pairwise distances between all objects (PDC clustering), while the second statement is based on minimizing the total distance from objects to cluster centers (CC clustering). Computational experiments showed that the hybrid method developed for solving the clustering problem combines the advantages of both approaches—the speed of the k-means method and the accuracy of PDC clustering—which makes it possible to get rid of the main drawback of the k-means, namely, the lack of guaranteed determining of the global optimum.

– Chapter 14 uses a model for the analysis of time series data which highly depend on the novel coronavirus 2019. This model predicts the future trend of confirmed, recovered, active, and death cases based on the available data from January 22, 2020 to May 29, 2021. The present model

predicted the spread of COVID-19 for a future period of 30 days. The RMSE, MSE, MAE, and MdAPE metrics are used for the model evaluation.

– Chapter 15 focuses on systems that incorporated real-world data utilized by actual users. It first describes a new methodology for the survey and then covers the various domains where semantic technology can be applied and some of the most impressive systems developed in each domain.

Finally, the editors would like to sincerely thank all the authors and reviewers who contributed their time and expertise for the successful completion of this book. The editors also appreciate the support given by Scrivener Publishing, which allowed us to compile the first edition of this book.

The Editors
Archana Patel
Narayan C. Debnath
Bharat Bhusan
June 2022

1

A Brief Introduction and Importance of Data Science

Karthika N.[1]*, Sheela J.[1] and Janet B.[2]

[1]Department of SCOPE, VIT-AP University, Amaravati, Andhra Pradesh, India
[2]Department of Computer Applications, National Institute of Technology, Tiruchirappalli, India

Abstract

Data is very important component of any organization. According to International Data Corporation, by 2025, global data will reach to 175 zettabytes. They need data to help them make careful decisions in business. Data is worthless until it is transformed into valuable data. Data science plays a vital role in processing and interpreting data. It focuses on the analysis and management of data too. It is concerned with obtaining useful information from large datasets. It is frequently applied in a wide range of industries, including healthcare, marketing, banking, finance, policy work, and more. This enables companies to make informed decisions around growth, optimization, and performance. In this brief monograph, we address following questions.

What is data science and what does a data scientist do? Why data science is in demand? History of data science, how data science differs from business intelligence? The lifecycle of data science, data science components, why data science is important? Challenges of data science, tools used for data science, benefits and applications of data science.

***Keywords*:** Data science, history, lifecycle, components, tools

**Corresponding author*: bharathikarthika@gmail.com

Archana Patel, Narayan C. Debnath and Bharat Bhusan (eds.) Data Science with Semantic Technologies: Theory, Practice, and Application, (1–30) © 2022 Scrivener Publishing LLC

1.1 What is Data Science? What Does a Data Scientist Do?

Data is very important component of any organization. According to International Data Corporation, by 2025, global data will reach to 175 zettabytes. They need data to help them make careful decisions in business. Data is worthless until it is transformed into valuable data. Data science plays a vital role in processing and interpreting data. It focuses on the analysis and management of data too. It is concerned with obtaining useful information from large datasets. It is frequently applied in a wide range of industries, including healthcare, marketing, banking, finance, policy work, and more. This enables companies to make informed decisions around growth, optimization, and performance. In nutshell, Data science is an integrative strategy for deriving actionable insights from today's organizations' massive and ever-increasing data sets. Preparing data for analysis and processing, performing advanced data analysis, and presenting the findings to expose trends and allow stakeholders to make educated decisions are all part of data science [1, 2]. Data science experts are both well-known, data-driven individuals with advanced technical capabilities who can construct complicated quantitative algorithms to organize and interpret huge amounts of data in order to address questions and drive strategy in their company. This is combined with the communication and leadership skills required to provide tangible results to numerous stakeholders throughout a company or organization.Data scientists must be inquisitive and results-driven, with great industry-specific expertise and communication abilities that enable them to convey highly technical outcomes to non-technical colleagues. To create and analyze algorithms, they have a solid quantitative background in statistics and linear algebra, as well as programming experience with a focus on data warehousing, mining, and modeling [3].

1.2 Why Data Science is in Demand?

Data science is the branch of science concerned with the discovery, analysis, modeling, and extraction of useful information which has become a buzz in a lot of companies. Firms are increasingly aware that they have been sitting on data treasure mines the priority with which this data must be analyzed, and ROI generated is obvious. We look at the most important reasons that data science professions are in high demand [4].

- **Data Organization**
  During the mid-2000s IT boom, the emphasis was on "lifting and shifting" offline business operations into automated computer systems. Digital content generation, transactional data processing, and data log streams have all been consistent throughout the last two decades. This indicates that every company now has a plethora of information that it believes can really be valuable but does not know how to use. This is apparent in Glassdoor's recent analysis, which identifies the 50 greatest jobs in modern era.
- **Scarcity of Trained Manpower**
  According to a McKinsey Global Institute study, by 2018, the United States will be short 190,000 data scientists, 1.5 million managers, including analysts who would properly comprehend and make judgments based on Big Data. The need is particularly great in India, where the tools and techniques are available but there are not enough qualified people. Data scientists, who can perform analytics, and analytics consultants, who can analyze and apply data, are two sorts of talent shortages, according to Srikanth Velamakanni, co-founder and CEO of Fractal Analytics. The supply of talent in these fields, particularly data scientists, is extremely limited, and the demand is enormous."
- **The Pay Is Outstanding**
  A data science position is currently one of the highest paying in the market. The national average income for a data scientist/analyst in the United States, according to Glass Door, is more than $62,000. In India, pay is heavily influenced by experience. Those with the appropriate skillset can earn up to 19 LPA. (source: PayScale.)
- **The "X" Factor**
  A data scientist's major responsibility are exceptional and specific to the position. Because of nature of the profession, they may flourish in their careers by integrating several analytical expertise across diverse areas such as big data, machine learning, and so on. This vast knowledge base gives them an unsurpassed reputation or X-factor.
- **Data Scientists' Democratization**
  Tech behemoths are not the only ones who need data scientists. According to a Harvard Business Report issued many years ago, "Organizations in the top list of their area in the

use of data-driven decision making were, on average, 5% more productive and 6% more profitable than their peers". Even mid-sized and small organizations have been driven to adopt data science because of this. In truth, many small businesses are trying to hire entry-level data scientists for a fair wage. This works well for both. The scientist will be able to further develop his or her skills, and the company will be able to pay less than it would otherwise.

- **Fewer Barriers for Professionals**
  Data science is open to a wide range of experts from varied backgrounds because it is a relatively new discipline. Math/statistics, computer science and engineering, and natural science are all areas of knowledge for today's data scientists. Some perhaps have social science, economics, or business degrees. They have all devised a problem-solving technique and improved their skills through formal or online education.
- **Abundance of Jobs**
  Data science is employed in a wide range of business sectors, from production to healthcare, Information Technology to finance, therefore there are plenty of data science jobs available for individuals who are interested and willing to put in the effort. It is true not only in terms of industries, but also in terms of geography. So, regardless of one's geographical location or current domain, data science and analytics are available to everybody.
- **A Wide Range of Roles**
  Even if data science job is indeed a broad term, there are numerous subroles that fall under its scope. Data scientists, data architects, business intelligence engineers, business analysts, data engineers, database administrators, and data analytics managers are all in considerable demand.

1.3 History of Data Science

The terminology "data science" was just recently coined a new profession interested in trying to make sense of large volumes of data. Making sense of data, on the other hand, has a significant background, and it has been addressed for years by many computer scientists, scientists, librarians, statisticians, and others. The history below shows how the terminology:

data science" evolved over time, as well as attempts to describe it and associated concepts [5].

In 1974, Peter Naur's book gives a broad overview of modern data processing techniques that are employed in a variety of applications. The IFIP Guide to Data Processing Concepts and Terms states that it is organized around the data principle: "Data is a codified representation of ideas or facts that may be communicated and even perhaps changed by certain process." According to the book's preface, a course plan titled "Datalogy, the science of data and data processes, and its position in education" was presented at the 1968 IFIP Congress, and the name "data science" has been widely used since then. Data science, according to Naur, is defined as "the science of working with data after it has been established, but the relationship of the data to what it represents is assigned to other disciplines and sciences."

In 1977, the International Association for Statistical Computing (IASC) was founded as an ISI chapter. "The goal of the IASC is to connect conventional statistical techniques, innovative computer technology, and domain specialists' skills to transform data into knowledge and information," says the organization." 1989 The first Knowledge Discovery in Databases (KDD) workshop is arranged and chaired by Gregory Piatetsky-Shapiro. Figure 1.1 shows the proceeding of IJCAI-Workshop. It was renamed the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) in 1995. September 1994, an article on "Database Marketing" appears in Business Week: "Enterprises are acquiring vast amounts of data on you, processing it to determine how likely someone really is to purchase a product, after then using that intelligence to design a marketing strategy perfectly tuned to find a way to convince you to do so... A prior spike of anticipation in the 1980s, sparked by the extensive use of checkout scanners, resulted in severe disappointment. Several organizations were overwhelmed with vast amount of data and were unable to do anything valuable with it... Despite this, many corporations recognize that they have no other option except to enter the database marketing arena."

In 1997, the journal Data Mining and Knowledge Discovery is set up, with the reversal of the two terms in the title emphasizing the rise of "data mining" as the standard term for "extracting information from vast datasets." December 1999, "Existing statistical procedures perform effectively with relatively small data sets, Jacob Zahavi says in Knowledge @ Wharton's "Mining Data for Nuggets of Knowledge." Today's databases, on the other hand, can have trillions of rows and columns of data. In data mining, scalability is a major concern. Another technical issue is creating models that would better analyze data, recognize nonlinear relations, and interaction

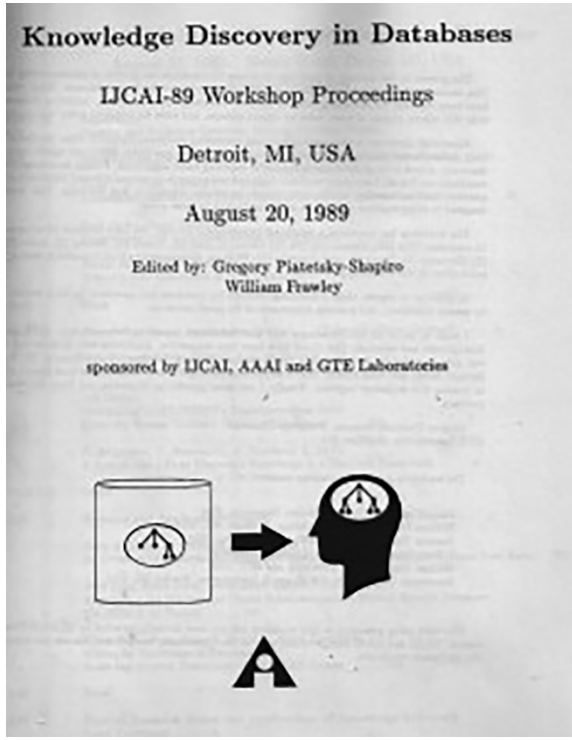

Figure 1.1 Proceeding of IJCAI-workshop.

among elements. To handle web-site issues, specialized data mining techniques may need to be developed."

2001 "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics," by William S. Cleveland, is published. That is a plan to "expand the key areas of technological endeavor in statistics." The new area will be dubbed "data science" because the notion is ambitious and requires significant change. Cleveland compares and contrasts the proposed new discipline to computer science and current data mining research.2001 "Statistical Modeling: The Two Cultures," by Leo Breiman, is published. When it comes to utilizing statistical modeling to draw conclusions from data, there are two contrasting cultures. A stochastic data model is assumed to have generated the data.

The statistical community has been devoted to using data models nearly exclusively. This emphasis has resulted in useless theory, spurious findings, and the exclusion of statisticians from a wide spectrum of current situations. Algorithmic modeling has advanced significantly in theory and in

practice in domains other than statistics that may be used on huge challenging as well as on small datasets as the more useful and consistent alternative to data modeling. If we want to utilize data to resolve problems as a domain, we ought to move away from merely using data models and use a wider range of tools. The International Council for Science's Committee on Data for Science and Technology (CODATA) publishes the journal (ICSU). January in the year 2003 "By data science, people mean practically the whole thing that has to do with data: acquiring, examining, and modeling," according to the launch of the Journal of Data Science. However, the most important feature is its applicability—it may be used in a wide range of situations. The journal is primarily concerned with the application of statistical methods in general. All data scientists will be able to submit their viewpoints and ideas in the Journal of Data Science. The data science venn diagram is depicted by Figure 1.2.

In "A Taxonomy of Data Science," Chris Wiggins and Hilary Mason write in September 2010: One potential taxonomy... of what a data scientist does may be: obtain, scrub, explore, model, and perceive, in approximately chronological order. Data science is evidently a merger of the hacker arts of statistics and machine learning, as well as math and data topic knowledge for the analysis to be understandable. It necessitates creative thinking and a willingness to learn. "To become a properly trained data scientist, one needs comprehend a great deal," writes Drew Conway in "The Data Science

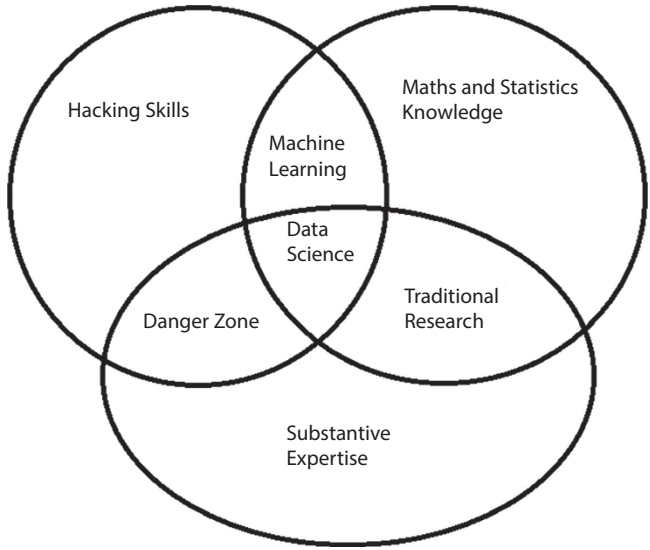

Figure 1.2 Data science Venn diagram.

Venn Diagram" from September 2010. Regrettably, simply citing literature and teachings is insufficient to untangle the tangles.

In the month of May 2011, in his article "Why the phrase 'data science' is wrong but useful," Pete Warden writes: According to P., there is no commonly agreed-upon boundary between what is inside and beyond the purview of data science. Rather than choosing a topic in the beginning and later gathering data to shed light on it, they tend to be more concerned with what the data can disclose and then picking fascinating strands to pursue. Matthew J. Graham presented "The Art of Data Science" at the Astro statistics and Data Mining in Large Astronomical Databases workshop in June 2011. "We need to learn new abilities to succeed in the modern data-intensive world of twenty-first-century science," he says. "We

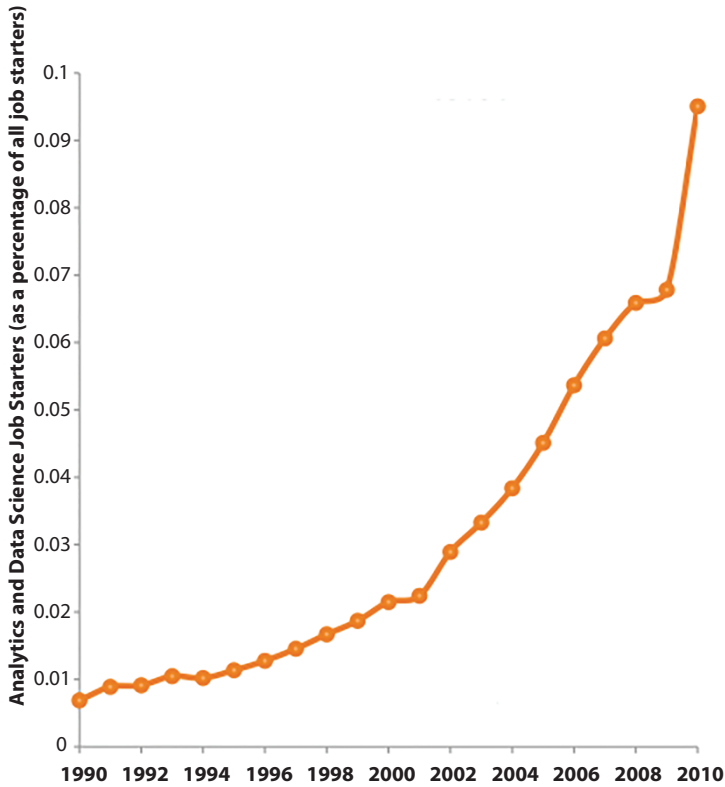

Figure 1.3 Job growth on analytics and data science.

should understand how [data] is perceived and expressed, as well as how it links to physical space and time."

The title of "business analyst" felt too constrictive. The title of "data analyst" was a candidate, but we believed it limited what people could do. After all, many of the members of our teams were highly skilled engineers. The phrase "research scientist" was used by firms, including Sun, Yahoo, HP, IBM, and Xerox as a fitting job title. We assumed, however, that the majority of research scientists worked on future and abstract issues in labs apart from the product development groups. It would take ages for lab research to have an impact on big items. As a replacement, our teams concentrated on developing data applications that can have an instant and significant influence on the firm. Data scientist seems to be the greatest fit: individuals who combine data and research to produce something new [6]. Figure 1.3 shows the job growth on analytics and data science.

1.4 How Does Data Science Differ from Business Intelligence?

Business intelligence (BI) is a method of analyzing descriptive data using technology and knowledge in order to make informed business decisions [7]. The BI toolkit is used to collect, govern, and transform data. It allows internal and external stakeholders to communicate data, making decision-making easier. Business intelligence is the process of extracting useful information from data. Some of the things that BI can help you with are:

- Developing a better grasp of the marketplace
- Identifying new revenue streams
- Enhancing business procedures
- Keeping one step ahead of the competition

Cloud computing has been the most important facilitator of BI in recent years. The cloud has enabled organizations to process more data, from more sources, and more efficiently than they could before cloud technologies were introduced.

Data science vs. business intelligence: Understanding the differences between data science and business intelligence is beneficial. Understanding how they work together is also beneficial. It is not a question of choose one over the other [8]. It all boils down to picking the proper solution to

receive the information you need. Most of the time, this means combining data science and business intelligence. The simplest approach to distinguish between the two is to think about data science in terms of the future and business intelligence in terms of the past and present. Data science is concerned with predictive and prescriptive analysis, whereas business

Table 1.1 Comparison of data science and business intelligence.

Factor	Data science	Business intelligence
Concept	It is a discipline that employs mathematics, statistics, and other methods to uncover hidden patterns in data.	It is a collection of technology, applications, and processes that businesses employ to analyze business data.
Focus	It is concentrated on the future.	It concentrated the past and present.
Data	It can handle both structured and unstructured data.	It is mostly concerned with structured data.
Flexibility	Data science is more adaptable since data sources can be added as needed.	It is less flexible because data sources for business intelligence must be planned ahead of time.
Method	It employs the scientific process.	It employs the analytic method.
Complexity	In comparison to business intelligence, it is more sophisticated.	When compared to data science, it is a lot easier.
Expertise	Data scientist is its area of competence.	Its area of specialization is for business users.
Questions	It addresses the questions of what will happen and what might happen.	It is concerned with the question of what occurred.
Tools	SAS, BigML, MATLAB, Excel, and other tools are used.	Insight Squared Sales Analytics, Klipfolio, ThoughtSpot, Cyfe, TIBCO Spotfire, and more solutions are among them.