BIG DATA APPLICATIONS IN INDUSTRY 4.0

Edited by
P. Kaliraj and T. Devi

CRC Press
Taylor & Francis Group
AN AUERBACH BOOK

Big Data Applications in Industry 4.0

Big Data Applications in Industry 4.0

Edited by
P. Kaliraj
T. Devi

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
AN AUERBACH BOOK

First Edition published 2022
by CRC Press
6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742

and by CRC Press
4 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

CRC Press is an imprint of Taylor & Francis Group, LLC

Library of Congress Cataloguing-in-Publication Data
A catalog record has been requested for this book.

ISBN: 978-1-032-00811-0 (hbk)
ISBN: 978-1-032-19179-9 (pbk)
ISBN: 978-1-003-17588-9 (ebk)

DOI: 10.1201/9781003175889

Typeset in Garamond
by MPS Limited, Dehradun

Prof. P. Kaliraj dedicates this book to
esteemed *Bharathiar University*,
his father *Mr. M. Perumal*,
mother *Mrs. P. Rathinammal*,
grand children *Prithika Karthikeyan*,
Kayan Karthikeyan,
Ayaan Prashanth and
Akira Prashanth.

Prof. T. Devi dedicates this book to
Department of Computer Applications
her mother *Mrs. A. Suseela*,
mother-in-law *Mrs. D. Singari*,
son *R. Surya* and,
grandson *V. Deera*

Contents

Preface xiii
Acknowledgments xix
Editors xxi
Contributors xxiii

1 Data Science and Its Applications 1
PAUL ABRAHAM AND LAKSHMINARAYANAN S
1.1 Introduction to Data Science 2
1.2 Data Science and Its Application in the Healthcare Industry . . . 10
1.3 Data Science and Its Application in the Retail and Retail E-Commerce 16
1.4 Data Science and Its Application in the Banking, Financial Services and Insurance (BFSI) Sector 21
1.5 Statistical Methods and Analytics Techniques Used across Businesses 24
1.6 Statistical Methods and Analytics Techniques Used in Sales and Marketing 25
1.7 Statistical Methods and Analytics Techniques Used in Supply Chain Management 31
1.8 Statistical Methods and Analytics Techniques Used in Human Resource Management 34
References 37

2 Industry 4.0: Data and Data Integration 39
PAVAN GUNDARAPU
2.1 Introduction 40
2.2 Data Integration 41
2.3 Data Integration Solutions 41
2.4 Data Integration Methodologies 46
2.5 Service Providers 51
2.6 Brief on Each Software 52
2.7 Conclusion 53

References 53

3 Forecasting Principles and Models: An Overview 55
R. VIJAYARAGHAVAN
3.1 Introduction 56
3.2 Meaning of Forecasting 57
3.3 Applications of Forecasting 57
3.4 Limitations of Forecasting 59
3.5 Types of Forecasting Procedures 60
3.6 Process of Forecasting 63
3.7 Basic Forecasting Models 64
3.8 Software Tools for Forecasting 68
3.9 Conclusions 68
References 70

4 Breaking Technology Barriers in Diabetes and Industry 4.0 71
KRISHNAN SWAMINATHAN AND THAVAMANI D. PALANISWAMI
4.1 Brief Introduction to Diabetes 72
4.2 “Big Data” Concept 74
4.3 Recent Technological Advances in Diabetes Management 76
4.4 Barriers in Diabetes Technology 80
4.5 Technical Solutions to Break the Barriers 80
4.6 Summary 81
References 81

5 Role of Big Data Analytics in Industrial Revolution 4.0 85
V. BHUVANESWARI
5.1 Big Data Analytics 87
5.2 Big Data Components 92
5.3 Big Data & Industry 4.0 97
5.4 Big Data Use Cases 99
5.5 Big Data Roles 102
References 105

6 Big Data Infrastructure and Analytics for Education 4.0 107
CHANDRA ESWARAN AND RATHINARAJA JEYARAJ
6.1 Introduction 108
6.2 Industrial Revolutions 108
6.3 Advantages of Industry 4.0 in Education 109
6.4 System for Smart Education 111
6.5 Big Data Infrastructure for Smart Education 115
6.6 Big Data Analysis for Smart Education 118

6.7 Conclusion ... 123
References ... 123

7 **Text Analytics in Big Data Environments** ... 125
R. JANANI AND S. VIJAYARANI
7.1 Introduction ... 126
7.2 Text Analytics – Big Data Environment ... 127
7.3 Applications of Text Analytics ... 137
7.4 Issues and Research Challenges in Text Analytics ... 139
7.5 Tools for Text Analytics ... 140
7.6 Conclusion ... 141
References ... 141

8 **Business Data Analytics: Applications and Research Trends** ... 145
S. SHARMILA AND S. VIJAYARANI
8.1 Big Data Analytics and Business Analytics: An Introduction ... 146
8.2 Digital Revolution of Education 4.0 ... 148
8.3 Conceptual Framework of Big Data for Industry 4.0 ... 149
8.4 Business Analytics ... 153
8.5 Applications of Big Data and Business Analytics ... 160
8.6 Challenges of Big Data and Business Analytics ... 160
8.7 Open Research Directions ... 163
8.8 Conclusion ... 164
References ... 164

9 **Role of Big Data Analytics in the Financial Service Sector** ... 169
V. RAMANUJAM AND D. NAPOLEON
9.1 Introduction ... 171
9.2 The Effect of Finance 4.0 in a Nutshell ... 172
9.3 In The Banking Industry, Big Data ... 175
9.4 Big Data Analytics in Finance Industry ... 184
9.5 Sector of Finance Data Science ... 187
9.6 Conclusion ... 191
Acknowledgments ... 192
References ... 192

10 **Role of Big Data Analytics in the Education Domain** ... 195
C. SIVAMATHI AND S. VIJAYARANI
10.1 Introduction ... 196
10.2 Need for Big Data Analytics in Education ... 202
10.3 Applications of Big Data Analytics in Education ... 204
10.4 Advantages of Big Data in Education ... 206

10.5 Challenges in Implementing Big Data in Education 206
10.6 Education 4.0 in India 207
10.7 Case Study: Big Data Analytics in E-Learning 207
10.8 Conclusion 210
References 212

11 Social Media Analytics 217
E. SUGANYA AND S. VIJAYARANI
11.1 Introduction 218
11.2 Process of Social Media Analytics 219
11.3 Social Media Analytics 221
11.4 Techniques and Algorithms 224
11.5 Tools 226
11.6 Research Challenges 226
11.7 Case Studies in Social Media Analytics 229
11.8 Conclusion 230
References 230

12 Robust Statistics: Methods and Applications 233
R. MUTHUKRISHNAN
12.1 Introduction 234
12.2 History of Robust Statistics 235
12.3 Classical Statistics vs. Robust Statistics 235
12.4 Robust Statistical Measures 237
12.5 Robust Regression Procedures 243
12.6 Data Depth Procedures 245
12.7 Statistical Learning 248
12.8 Robust Statistics in R 252
12.9 Summary 253
References 254

13 Big Data in Tribal Healthcare and Biomedical Research 259
DHIVYA VENKATESAN, ABILASH VALSALA GOPALAKRISHNAN, NARAYANASAMY ARUL, CHHAKCHHUAK LALCHHANDAMA, BALACHANDAR VELLINGIRI AND N. SENTHIL KUMAR
13.1 Introduction 260
13.2 Data Lifecycle 264
13.3 Big Data in Genomic Research 267
13.4 Big Data in Biomedical Research 274
13.5 Healthcare as a Big Data Repository 286
13.6 Management of Big Data 286
13.7 Challenges in Healthcare Data 288
13.8 Tribal Research in India 290
13.9 Conclusion 291

Acknowledgments 291
References 291

14 PySpark towards Data Analytics 297
J. RAMSINGH
14.1 Introduction 299
14.2 PySpark – SparkContext 302
14.3 PySpark Shared Variables 305
14.4 PySpark – RDD (Resilient Distributed Dataset) 305
14.5 PySpark DataFrames 315
14.6 PySpark MLlib (Machine Learning Libraries) 323

15 How to Implement Data Lake for Large Enterprises 331
RAGAVENDRAN CHANDRASEKARAN
15.1 What Is a Data Warehouse? 332
15.2 What Is a Data Lake? 334
15.3 Why Do We Need Data Lake? 334
15.4 Overview of Data Lake in Cloud 334
15.5 Key Considerations for Data Lake Architecture 336
15.6 Phases of Data Lake Implementation 337
15.7 What to Load into Your Data Lake? 341
15.8 A Cloud Data Lake Journey 343
15.9 Conclusion 348
References 348

16 A Novel Application of Data Mining Techniques for Satellite Performance Analysis 349
S.A. KANNAN AND T. DEVI
16.1 Introduction 350
16.2 Data Generation and Analysis 351
16.3 Data Mining 352
16.4 Artificial Satellites and Data Mining 353
16.5 Statistical Techniques for Satellite Data Analysis 354
16.6 Novel Application 355
16.7 Selection of an Appropriate Data Mining Technique 356
16.8 Satellite Telemetry Data – Association Mining 359
16.9 Satellite Telemetry Data – Decision Tree Technique 360
16.10 Satellite Telemetry Data – A Modified Brute-Force Rule-Induction Algorithm 361
16.11 Methodology 362
16.12 Conclusion 363
References 364

17 Big Data Analytics: A Text Mining Perspective and Applications in Biomedicine and Healthcare 367
JEYAKUMAR NATARAJAN, BALU BHASURAN, AND GURUSAMY MURUGESAN
17.1 Introduction 369
17.2 Text Mining Overview and Related Fields 372
17.3 Phases and Tasks of Text Mining 377
17.4 Applications in Biomedicine 388
17.5 Applications in Healthcare 394
17.6 Conclusion 401
References 401

Index 409

Preface

Industry 4.0 is the latest technological innovation in manufacturing with the goal to increase productivity in a flexible and efficient manner. This revolutionary transformation which is changing the way in which manufacturers operate is powered by various technology advances including artificial intelligence, Big Data analytics, internet-of-things, and cloud computing. Big Data analytics has been identified as one of the significant components of industry 4.0 as it provides valuable insights for the purpose of smart factory management. This scenario requires the data to be processed with advanced tools and technologies in order to provide relevant information. Big Data and Industry 4.0 have the potential to shape up the industrial process in terms of resource consumption, process optimization, automation, and much more. It can be inferred that it also plays a key role in achieving sustainable development. However, keeping pace with these technologies require an individual to be highly skilled and well knowledgeable in identifying and solving any real-time problem. Such problems can be as small as a minute shift in the data generated, which may affect their surroundings, even their lives later. The exponentially rising generation rate of data has made the Big Data analytics a challenging area of research.

The Big Data Analytics Market growth is forecasted to grow at a compound annual growth rate of 29.7% to $40.6 billion by 2023 as per Frost & Sullivan. The growth in the Big Data analytics market will accelerate the need for specialists in Big Data analytics. And with demand for talented professionals more than doubling in the last few years, there are limitless opportunities for professionals who want to work on the cutting edge of Big Data research and development.

The awareness and practice on Big Data and its applications, skill development to face Industry 4.0 and technological advanced infrastructure become the keys for successful development of future pillars of our Globe. Linking Big Data analytics, which is one of the tools of Industry 4.0 with arts and science education is the need of the hour. Today, the rate at which the transformation happening is very disruptive in nature and also exponential changes are being witnessed. Educational institutions have to be way ahead of the requirement and prepare their students to meet the new challenges to be created by Industry 4.0. Currently, educational institutions are at crossroads and do not know how to interweave the Industry 4.0 tools into the arts, science, social science and teacher education programmes in Universities. The book can aid in imparting the concepts and knowledge of Big Data among graduates studying in Higher Education Institutions as it highlights the fundamentals and research trends of big data. It also describes applications of big data in various sectors such as finance, education, social media, remote sensing, and healthcare. Currently, there are no books on big data and its applications that can be used in the curriculum of higher education. The proposed book has a huge scope to be included in the higher education curriculum. Hence, the demand for the book among graduates and higher education institutions will be present as long as the curriculum of higher education institutions focuses on the development of Industry 4.0 skills. The students, scholars, and teachers can be from Arts & Science Universities, Engineering Institutions, and Teacher Education Universities. Practitioners – Scientists, Engineers, and Statisticians who are interested in building Big Data applications or analytical models to solve real world problems can also use this book for reference.

This book covers the recent advancements that have emerged in the field of Big Data and its applications. The exponentially rising generation rate of data has made the Big Data analytics, a challenging area of research. The book introduces the concepts, advanced tools and technologies for representing, and processing Big Data. It also covers applications of Big Data in domains such as financial services sector, education, tribal health care, biomedical research, healthcare, logistics, and warehouse management. Students of every discipline must be familiar with this fast growing technology since their future job prospects will be influenced by this technology. This book can be used in courses offered by Higher Education Institutions which strive to equip their graduates with Industry 4.0 skills. It can be used by scientists, engineers, and statisticians who are interested in building Big Data applications to solve real world problems.

Chapter 1 entitled "*Data Science and Its Applications*" introduces Data Science and discusses its applications in the business today. This chapter explores the possible types of data available in the business today, the many types of data analytics methods accessible today and covers uses cases through its applications.

Chapter 2 entitled "*Industry 4.0: Data and Data Integration*" provides an overview of what Data Integration is, the different Data Integration solutions available, and the different methodologies of Data Integration. The chapter also discusses about various Data Integration service providers available in the market.

Chapter 3 entitled "*Forecasting Principles and Models: – An Overview*" gives the readers a clear understanding of the general framework of forecasting principles, applications, limitations, and procedures for the data pertaining to such fields along with three basic forecast models, namely, naïve, moving average, and exponential smoothing models highlighting their significance.

Chapter 4 entitled "*Breaking Technology Barriers in Diabetes and Industry 4.0*" explores the application of Big Data in diagnosing diabetes. Diabetes is a fertile area to implement the concepts of Industry 4.0 that would directly impact lives of millions of people in India and world-wide. Barriers in Diabetes technology and technical solutions to break the barriers are detailed in this chapter. Healthcare is a fertile area with huge potential for Big Data, precision medicine, artificial intelligence, data mining, development of prediction models, health apps, machine automation, closed-loop technologies, and noninvasive monitoring systems.

Chapter 5 entitled "*Role of Big Data Analytics in Industrial Revolution 4.0*" provides readers a complete understanding emphasizing the need for Big Data for Industry 4.0 transformation. The chapter provides a detailed roadmap of Data evolution and its related technological transformation in computing with a brief description of data related terminologies as an introduction.

Chapter 6 entitled "*Big Data Infrastructure and Analytics for Education 4.0*" examines the application of Industry 4.0 and Big Data in the field of education. This chapter outlines how Industry 4.0 is being applied in education and discusses various Big Data infrastructure and analytics to build effective online teaching and learning.

Chapter 7 entitled "*Text Analytics in Big Data Environments*" explains the background of text analytics and text analytics in Big Data domain. It also discusses how machine learning techniques are applied over the huge volume of data in Big Data environment, addresses the research challenges and issues of text analytics over the Big Data environment, and discusses the tools for text analytics.

Chapter 8 entitled "*Business Data Analytics: Applications and Research Trends*" discusses the overview of Education 4.0, Big Data Analytics and Business Analytics, and the impact of Big Data Analytics in Education 4.0 as well as Business Analytics. Research perspectives and directions in these domains are also projected in this chapter.

Chapter 9 entitled "*Role of Big Data Analytics in the Financial Service Sector*" summarizes the features, prospects, and significant role of Big Data in banking industry and also its advantages in the financial sector. The chapter tries to identify the various use cases of Big Data in banking, finance services and insurance (BFSI) areas, where this analytics is turning out to be paramount.

Chapter 10 entitled "*Role of Big Data Analytics in the Education Domain*" describes the use of Big Data Analytics in Education domain. This chapter discusses how to analyze the educational data to improve the quality of education. It further discusses how Big Data technology will be used to assess the student performance, evaluation strategies, preparation of question papers, online examinations, comparison of curriculum, open-source educational tools, and web-based learning.

Chapter 11 entitled "*Social Media Analytics*" discusses the social media platforms and step-by-step processes of analysing the data available through social media. It describes domains of social media analytics (SMA), various types of analysis, techniques and algorithms for analysis such as natural language processing (NLP), news analytics, opinion mining, scraping, and text analytics. It introduces the machine learning and deep learning algorithms, software tools that are available for social media analytics, and research challenges.

Chapter 12 entitled "*Robust Statistics: Methods and Applications*" is a study on the assumptions and limitations of classical statistical procedures. It explores various robust statistical procedures developed in recent past, by considering the measure of location and scale, in the area of data depth, regression and multivariate analysis. This chapter analyzes data using robust statistical methods along with conventional statistical procedures using robust statistical packages in R programming.

Chapter 13 entitled "*Big Data in Tribal Healthcare and Biomedical Research*" confers the process of Big Data approaches in socio-economic status and in genomic research (NGS and Metagenomics). The chapter aims at deliberating healthcare as a Big Data repository, its analytics, and challenges in data retrieval and reiterates the necessity of Big Data in tribal community healthcare.

Chapter 14 entitled "*PySpark toward Data Analytics*" explores Pyspark in detail. The chapter explores how PySpark overcomes the drawbacks of Apache Hadoop MapReduce and how it extends the MapReduce model for its interactive queries and stream processing.

Chapter 15 entitled "*How to Implement Data Lake for Large Enterprises*" focuses on implementation of the Data Lake (DL) in cloud and the significance of DL where the pre-existence of a Data Warehouse (DW) helps businesses to take decisions.

Chapter 16 entitled "*A Novel Application of Data Mining Techniques for Satellite Performance Analysis*" provides a brief knowledge on how data mining techniques can be used to analyze satellite performance.

Chapter 17 entitled "*Big Data Analytics: A Text Mining Perspective and Applications in Biomedicine and Healthcare*" provides an overview of the text mining perspective of Big Data analytics with an emphasis on applications in biomedicine and healthcare. The chapter illustrates phases and tasks of text mining in Big Data scope and provides a description of two application areas of biomedicine and healthcare where text mining using Big Data analytics is applied.

How to Use the Book?

The method and purpose of using this book depend on the role that you play in an educational institution or in an industry or depend on the focus of your interest. We propose five types of roles: student, software developer, teacher, member of Board of Studies, and researcher.

If you are a student: Students can use the book to get a basic understanding of Big Data, its tools, and applications. Students belonging to any of the arts, science and social science disciplines will find useful information from chapters on complete insight on Big Data, fundamentals and applications. This book will serve as a starting point for beginners. Students will benefit from the chapters on applications of Big Data and data analytics in *biomedicine, healthcare, education, social media, finance, and satellite performance analysis*.

If you are a software developer: Software developers can use the book to get a basic understanding of Big Data, its tools, and applications. Readers with software development background will find useful information from chapters on fundamentals and applications. They will benefit from the chapters on *data integration, data lakes based on cloud, robust statististical methods given in R programming* and *PySpark.* Software developers will find the data analytics tool *PySpark* very useful from configuring runtime options, running in standalone, interactive jobs, writing simple programs, streaming analysis, and machine learning packages for data analysis. Cloud-based data lakes can be built by software developers using the concepts and architecture given in the chapter on implementation of *Data Lake for large enterprises.*

If you are a teacher, the book is useful as a text for several different university-level, college-level undergraduate and postgraduate courses. Chapters on *forecasting principles and models,* and *robust statistical methods* will help in gaining the knowledge on the Statistical models and methods that form the base for data analytics. A graduate course on Big Data can use this book as a primary textbook. It is important to equip the learners with a basic understanding on Big Data, a tool of Industry 4.0. Chapter on *Big Data – A Complete Insight* provides the fundamentals of Big Data. To teach the applications of Big Data in various sectors, say Healthcare, teachers will find useful information from chapters on diabetes, biomedicine and healthcare. A course on Big Data for Science too could use the chapters on diabetes, biomedicine and healthcare, and satellite performance analysis. A course on Big Data and Education could use the chapters that deal with application of Big Data, data analytics in Education 4.0.

If you are a member of the Board of Studies: Innovating the education to align with Industry 4.0 requires that the curriculum be revisited. Universities are looking for methods of incorporating Industry 4.0 tools across various disciplines of Arts, Science, and Social Science Education. This book helps in incorporating Big Data across Science and Education. The book is useful while framing the syllabus for new course that cut across Big Data and disciplines of Arts or Science or Social Science Education. For example, syllabi for courses entitled Big Data in science, Big Data in healthcare, Big Data in medical biotechnology, Big Data in education may be framed using the chapters in the book. Industry infusion into curriculum is given much importance by involving more industry experts – R&D managers, product development managers, technical managers as special invitees in the Board of Studies. Chapters given by industrial experts in this book will be very helpful to infuse the application part of Big Data into the curriculum.

If you are a researcher: A crucial area where innovation is required is the research work carried out by universities and institutions so that innovative, creative, and useful products and services are made available to society through translational research. This book can serve as a comprehensive reference guide for researchers in the development of experimental Big Data applications. The chapters on *diabetes and Industry 4.0, Healthcare, biomedical research, Education 4.0, business data analytics, finance,* and *satellite performance analysis* provide researchers, scholars, and students with a list of important research questions to be addressed using Big Data.

*****

Acknowledgments

From Prof. P. Kaliraj

First and Foremost, I express my sincere gratitude to **Hon'ble Shri. Banwarilal Purohit,** Governor of Tamil Nadu, India who was instrumental in organising the conference on Innovating Education in the era of Industry 4.0 during Dec 14–15, 2019, in Ooty, which paved the way for further work in Industry 4.0 knowledge world.

My heartfelt thanks go to Hon'ble Chief Minister of Tamil Nadu, India and Hon'ble Minister for Higher Education, Government of Tamil Nadu. I thank Principal Secretary to Government, Higher Education Department, Government of Tamil Nadu.

I would like to express my thanks to Secretary to Governor, and Deputy Secretary to Governor, Universities Governor's Secretariat, Raj Bhavan, Chennai.

I thank my wife, Dr. Vanaja Kaliraj, and my son-in-law Mr. M. Karthikeyan, daughter Mrs. Ratna Priya Karthikeyan, son Dr. K. Siva Prashanth, daughter-in-law Dr. Ratna Priya Prashanth, grand children Prithika Karthikeyan, Kayan Karthikeyan, Ayaan Prashanth, and Akira Prashanth, and my family members for their support and being patient.

From Prof. T. Devi

I record my sincere thanks to **Prof. P. Kaliraj**, Hon'ble Vice-Chancellor of Bharathiar University, who identified the knowledge world gap when Professor searched for a book on Industry 4.0 and triggered the process of writing and editing books in the Industry 4.0 series. His continuous motivation during the lockdown period due to COVID-19, sensitisation and encouragement are unmatchable.

I express my profound thanks to the **Vice-Chancellor** and **Registrar** for the administrative support. Heartfelt thanks are due to the **authors of the chapters** for their contribution of chapters, continuous co-operation in improvising the chapters as and when requested and for timely communication. I thank all the expert members who served as reviewers for providing a quality and swift review.

We wish to thank **Mr. John Wyzalek, Senior Acquisitions Editor, Taylor & Francis/CRC Press**, who gave the reply mail in the first week of October 2020 mentioning that our proposal on Big Data book is **topical,** believed in this book's idea and helped us in realizing our dream.

Special thanks are due to Mr. Todd Perry, production editor, Taylor & Francis/CRC Press, Florida, for his excellent co-ordination and Mr. Manmohan Negi, Project Manager, MPS Ltd., for his untiring and swift support.

Thanks to Ms. Stephanie Kiefer, Editorial Assistant, Taylor & Francis/ CRC Press, for her efforts and communications during the project.

Thanks are due to **Dr. R. Rajeswari**, Associate Professor, Department of Computer Applications for her continuous support; **Sister Italia Joseph Maria** and **Ms. M. Lissa**, Project Assistants for providing earnest support.

Thanks to the faculty members **Prof. M. Punithavalli**, **Dr. T. Amudha**, **Dr. J. Satheeshkumar**, **Dr. V. Bhuvaneswari**, **Mr. S. Palanisamy**, **Dr. S. Gavaskar**, **Dr. R. Balu**, and **Dr. J. Ramsingh**.

Thanks to the Assistant technical officers **Mr. A. Elanchezian**, **Mr. A. Sivaraj**, **Mrs. B. Priyadarshini**, and office staff **Mr. A. Kalidas** of Department of Computer Applications of Bharathiar University, India.

Thanks are due to **Mrs. K. Kowsalya**, Assistant Registrar, **Mr. R. Karthick**, Assistant Section Officer, and **Mr. A. Prasanth** of office of the Vice-Chancellor and staff of the office of the Registrar of Bharathiar University, India.

Finally, I thank my husband Mr. D. Ravi, daughter Mrs. R. Deepiga, son Mr. R. Surya, son-in-law Mr. D. Vishnu Prakhash and grandson V. Deera and family members for their encouragement and support.

*****

Editors

Prof. P. Kaliraj, Hon'ble Vice-Chancellor, Bharathiar University, a Visionary and an Eminent Leader leading big academic teams is having more than three decades of teaching and research experience. He has held various renowned positions such as officiating Vice-Chancellor of Anna University, Head of Centre for Biotechnology of Anna University, Dean of faculty at A C College of Technology, and Member of the Syndicate for two decades at Anna University. Professor had research collaborations with the National Institute of Health in Maryland, United States, Glasgow University in Scotland, United Kingdom, and University of Illinois in Rockford, United States. University Grants Commission BSR Faculty Award and the Lifetime Achievement Award from the Biotechnology Research Society of India adorned the Professor. **42 scholars were gifted to receive the highest academic degree under his distinguished Guidance**. His remarkable **patent in the area of Filariasis is a boon in healthcare** and saving the lives of mankind. He is a Great Motivator and very good at sensitising the Faculty, Scholars and Students toward achieving Academic Excellence and Institutional Global Ranking. Professor is a recipient of **Life Time Achievement Award and Sir J.C. Bose Memorial Award** for his outstanding contribution in higher education research (e-mail: vc@buc.edu.in, pkaliraj@gmail.com).

Prof. T. Devi Ph.D. (UK), Professor, Centre for Research and Evaluation, Professor and Head, Department of Computer Applications, Bharathiar University, focuses on state-of-art technology that industries adopt in order to make the students ready for the future world. She is a **Gold Medalist** (1981–1984) from University of Madras and a **Commonwealth Scholar** (1994–1998) for her **Ph.D. from University of Warwick, United Kingdom**. She has three decades of teaching and research experience from Bharathiar University, Indian Institute of Foreign Trade, New Delhi and University of Warwick, United Kingdom. Professor is good in team building and setting goals and achieving. Her research interests include integrated data modeling and framework, meta-modeling, computer-assisted concurrent engineering, and speech processing. Professor had visited United Kingdom, Tanzania, and Singapore for academic collaborations. She has received various awards including **Commonwealth Scholarship**, **Best Alumni Award** from PSGR Krishnammal College for Women (PSGRKCW), Proficiency award from PSG College of Technology and awards from Bharathiar University for serving for BU-NIRF, Curriculum 4.0, and Roadmap 2030, and **guided 23 Ph.D. scholars**. Prof. T. Devi may be contacted at (e-mail: tdevi@buc.edu.in, tdevi5@gmail.com).

Contributors

V.G. Abilash
Department of Biomedical Sciences
Vellore Institute of Technology
Vellore, India

Paul Abraham
Data Science Evangelist
IT/ ITES Industry
Bangalore, India

N. Arul
Department of Zoology
Bharathiar University
Coimbatore, India

V. Balachandar
Department of Human Genetics and Molecular Biology
Bharathiar University
Coimbatore, India

Balu Bhasuran
DRDO-BU Center for Life Sciences
Bharathiar University
Coimbatore, India

V. Bhuvaneswari
Department of Computer Applications
Bharathiar University
Coimbatore, India

E. Chandra
Department of Computer Science
Bharathiar University
Coimbatore, India

Ragavendran Chandrasekaran
Senthil Nagar, Vandalur
Chennai, India

V. Dhivya
Department of Human Genetics and Molecular Biology
Bharathiar University
Coimbatore, India

Pavan Gundarapu
Data Integration solutions manager
Safran
Paris, France

R. Janani
Department of Computer Science
Bharathiar University
Coimbatore, India

Rathinaraja Jeyaraj
Kyungpook National University
Daegu, South Korea

N. Senthil Kumar
Department of Biotechnology
Mizoram University
Mizoram, India

C. Lalchhandama
Health and Family Welfare Department
Government of Mizoram
Mizoram, India

Gurusamy Murugesan
Department of Bioinformatics
Bharathiar University
Coimbatore, India

R. Muthukrishnan
Department of Statistics
Bharathiar University
Coimbatore, India

D. Napoleon
Department of Computer Science
Bharathiar University
Coimbatore, India

Jeyakumar Natarajan
Department of Bioinformatics
Bharathiar University
Coimbatore, India

Thavamani D. Palaniswami
Kovai Medical Center and Hospital
Coimbatore, India

V. Ramanujam
Bharathiar School of Management and Entrepreneur Development
Bharathiar University
Coimbatore, India

J. Ramsingh
Department of Computer Applications
Bharathiar University
Coimbatore, India

Lakshminarayanan S.
Department of Computer Applications
Bharathiar University
Coimbatore, India

Kannan S. A.
U R Rao Satellite Centre
Indian Space Research Organisation
Bangalore, India

S. Sharmila
Department of Computer Science
Bharathiar University
Coimbatore, India

C. Sivamathi
Department of Computer Science
Bharathiar University
Coimbatore, India

E. Suganya
Department of Computer Science
Bharathiar University
Coimbatore, India

Krishnan Swaminathan
Department of Endocrinology
Kovai Medical Center and Hospital
Coimbatore, India

R. Vijayaraghavan
Department of Statistics
Bharathiar University
Coimbatore, India

S. Vijayarani
Department of Computer Science
Bharathiar University
Coimbatore, India

Chapter 1

Data Science and Its Applications

Paul Abraham[1] and Lakshminarayanan S.[2]
[1]*Doddabanaswadi, Bangalore, India*
[2]*Department of Computer Applications Bharathiar University, Coimbatore, India*

Contents

1.1 Introduction to Data Science........ 2
1.1.1 Data Science: A Definition........ 3
1.1.2 Data in the Business........ 5
1.1.3 Types of Data Analytics........ 6
1.1.4 Use Cases in the Business........ 7
1.1.5 Data Analytics Process, Implementation and Measurement........ 9
1.2 Data Science and Its Application in the Healthcare Industry........ 10
1.2.1 Data Types Generated in the Healthcare Sector........ 10
1.2.2 Analytics Use Cases in Healthcare........ 11
1.2.3 Future and Challenges........ 15
1.3 Data Science and Its Application in the Retail and Retail E-Commerce........ 16
1.3.1 Data Types Generated in the Retail and Retail E-Commerce Sector........ 16
1.3.2 Analytics Use Cases in Retail and Retail E-Commerce........ 17
1.3.3 Future and Challenges........ 21
1.4 Data Science and Its Application in the Banking, Financial Services and Insurance (BFSI) Sector........ 21
1.4.1 Data Types Generated in the BFSI Sector........ 22
1.4.2 Analytics Use Cases in BFSI........ 22

DOI: 10.1201/9781003175889-1

1.4.3 Future and Challenges 24
1.5 Statistical Methods and Analytics Techniques Used across Businesses 24
1.6 Statistical Methods and Analytics Techniques Used in Sales and Marketing 25
1.6.1 Data Types Generated in Sales and Marketing Function 25
1.6.2 Statistical Methods and Analytical Techniques 26
1.6.3 Future and Challenges 31
1.7 Statistical Methods and Analytics Techniques Used in Supply Chain Management 31
1.7.1 Data Types Used in the SCM 31
1.7.2 Analytics Use Cases in SCM 32
1.7.3 Future and Challenges 34
1.8 Statistical Methods and Analytics Techniques Used in Human Resource Management 34
1.8.1 Data Types Generated in Human Resource Management 34
1.8.2 Analytics Use Cases in Human Resource Management 35
1.8.3 Future and Challenges 36
References 37

This chapter starts with a brief introduction to data science and aims to cover *three industry segments and three business functions*, where and how data science is applied.

Objectives

The objective of this chapter is to introduce data science and discuss its applications in the business today. Data science is about solving business problems, and businesses must recognize this fact. It examines which questions need answers and where to find the related data to support business decisions. This chapter defines and introduces the field of data science, possible types of data available in the business today, the many types of data analytics methods available today and covers use cases through its application. Though data science is used in all walks of life, this chapter restricts only its text to the scope of business or commercial activity. Going a little deeper, this chapter aims to cover three industry segments and three business functions where data science is applied.

1.1 Introduction to Data Science

Businesses see an uprising in transactions, leading to creating a huge repository of data comprising these transactions. This creates a need for information, insight, and intelligence about the business. Managers in the businesses moved from making decisions out of experience or institution to fact-based, data-driven decisions. This was effectively done by understanding the business objectives and their operative nuances and building intelligence around them.

The last decade has seen a huge transformation in the businesses moving toward a digital era by automating their process flows. In this trend, most businesses have also been collecting and storing their data in digital formats, and now the time has come to analyze and bring some value from the collected data. The collected data now demands to be cleaned by removing noises or unwanted information before being processed (Foster Provost & Tm Fawcett, 2018) to bring out meaningful insights for the business. Significant advancements related to storage spaces, thereby reducing the hardware costs, faster processing, and software products capable of performing complex calculations have become a boon to the business wanting to have a data-driven culture for decision making.

1.1.1 Data Science: A Definition

The loose definition of data science is to analyze data of a business to be able to produce actionable insights and recommendations for the business (Affine Analytics, 2018). The simplicity or the complexity of the analysis also impacts the quality and accuracy of results. As businesses and the data they collect became sophisticated, the need for technological skills, math/stats skills, and the necessary business acumen to define and deliver a relevant business solution became more relevant.

Data science is the process of examining data sets to conclude the information they contain, increasingly with the aid of specialized systems and software, using techniques, scientific models, theories, and hypotheses. These three pillars have very much been the mainstay of data science ever since it started getting embraced by businesses over the past two decades and should continue to be even in the future (Figure 1.1): Computer Science & IT, Business Acumen and Methods, Models, & Process.

Data Science expressed like this in the above picture is an idea accepted in academia and industry. It's an intersection of programming, analytical, and business skills that allows extracting meaningful insights from data to benefit business growth. However, this is used in social research, scientific & space programs, government planning, and so on, but this chapter will focus on its application in the Business Industry.

	DATA SCIENCE MODEL DEFINITION
❶	Business Acumen in its purest form means running a Business Enterprise. Any business existing to sell its product or services for a profit incurring some cost and generally having the functions like HR, Supply Chain, Finance, Sales & marketing to support it

(*Continued*)

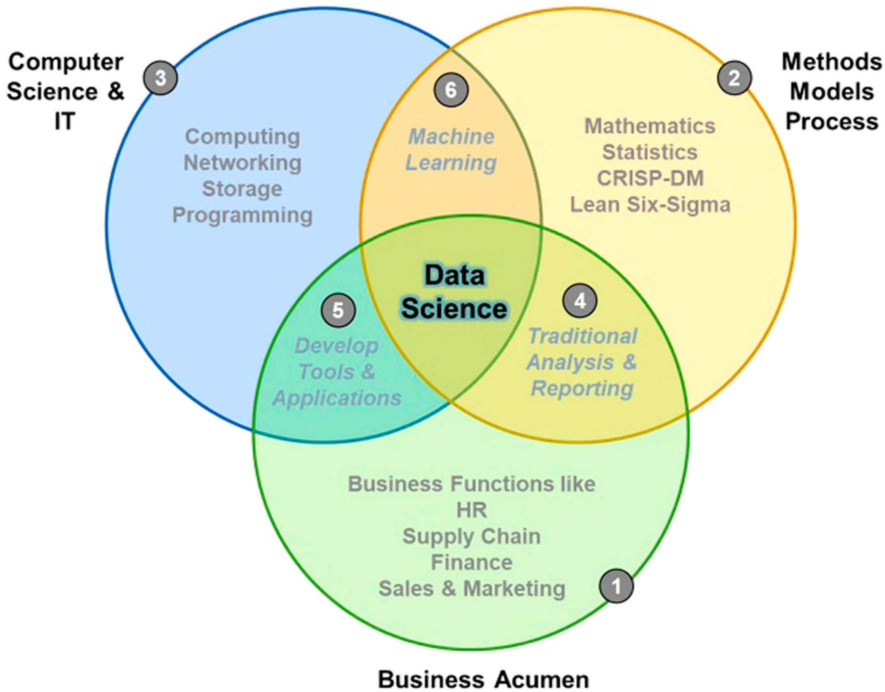

Figure 1.1 The Data Science model.

❷	Methods, Models, Process are defined as industry and academia proved practices that are the backbone to Data Science, including Mathematical models, theorems, Statistical methods, techniques, and process methodologies likes CRISP-DM, Six-Sigma, Lean, and so on
❸	Computer Science & IT practice is the full range of hardware, the software involved in providing computing for processing data, storage for storing and sharing data and networking for collecting and movement.
❹	When Business Acumen or Knowledge and Models methods process come together, it's classically called "traditional research." It involves using data collected in the business to make dashboards and reports to understand the business, plan for its future and make corrections if needed.
❺	Businesses take help from the Computer Science IT practice to help run business by building applications, web services, websites or plan IT-related strategies like going "digital" or adopting "cloud" based delivery of its products & services to serve their customers

(*Continued*)

❻	Machine learning is an idea to analyze data and automate the building of data models or algorithms. For example, medical diagnosis, image processing, prediction, classification, learning association, regression etc. Intelligent systems built on machine learning algorithms can learn from past experience or historical data.

1.1.2 Data in the Business

"In God we Trust, all others bring data"; this famous quote has been attributed to W. Edwards Deming. Deming was heavily involved in the economic reconstruction of post–World War 2 Japan. He proposed the philosophy to measure and analyze with data. This eventually helped in gaining increased performance in all areas of business. This philosophy is perfectly viable even today for any business or business situation. Leaving experience aside, many businesses seldom know about the performance of their business or, even more importantly, how they can improve it further.

In Industry 4.0, there is one key input or raw material playing the most critical role. This raw material is invisible and intangible in contrast to what we can see like oil, iron ore, or any physically visible components. It is nothing but "data," a special connection across a connected industry. With apt tools, techniques, and technology, companies can use data as their trump card. "Data are becoming the new raw material for business," says Craig Mundie, a senior advisor to the CEO at Microsoft and to the former American President Obama. Gone are the data when just rows and columns were data. Today everything is data.

An Industry pioneer was asked, "how do you deal with so much unstructured data that is generated through the social media, audio, chats, videos, emails, pictures, blogs posts?"; the pioneer believes that this data is rich in information and can be used to mine insights from it to be used to make a business decision. It's a challenge to work on such unstructured data because it calls for skillful hands to operate on it. People talk about or vent their feelings on social media. This means businesses can quickly gauge the sentiments among people for your business or brand. It just gives you an idea of what people out there are talking about you? Is there something valuable that you can use to make a course correction to your offer, brand, or business? In-store videos are used to generate heat maps to identify where people spend more time; this can be correlated with merchandise in those spaces, and now merchandise planning can be a lot more planned. A large tech giant is using product photographs to identify counterfeit products from real ones. Fraud detection is now possible. All of the above examples are big data analysis at work. So, there are a lot of possibilities to use unstructured data positively. How can unstructured data be paired with structured data to make it richer? How can this become a standard in the business are points to ponder over. This presents the business with immense opportunity to know what is happening in the business and react quickly to change and go faster to market.

Using data to prove some of the business decisions being taken gives confidence at different management levels to execute their plans, rather than depending on purely experience or trial measures. A data perspective provides structure and principles that give a framework to analyze any problems and implement a solution with confidence. Once industries develop data-analytical thinking, it clarifies the misconceptions and enriches the knowledge where they could apply these techniques in various domain topics.

1.1.3 Types of Data Analytics

For the given data collected and the business problems identified, numerous analysis methods can be identified. The below-mentioned four methods (Figure 1.2) can be considered to be generally accepted both by industry and academia alike.

The Analytics Advancement Model helps define, identify and illustrate what these types of analysis mean. In the above model, we can visualize four types of analysis possible and show them in terms of complexity of analysis and volume of analysis. Volume here means done often. There is no apparent relationship between volume and complexity.

Descriptive analysis is termed as the first step in any analytical problem-solving project. It is the simplest to perform in the analysis ladder of knowledge. As a foundational analysis, it aims to answer the question "what happened?" For example, a company selling breakfast cereals through descriptive analysis will find

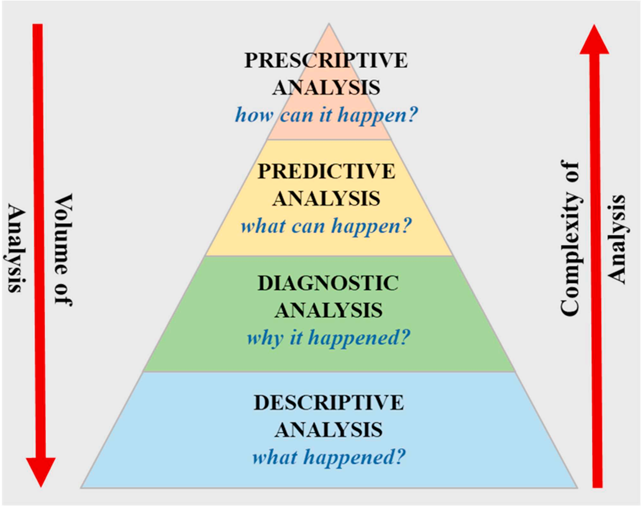

Figure 1.2 Analytics advancement model.

insight into its sales volume, units, value for a given geography and time period. With simple statistics like average sales, maximum sales, and minimum sales, the business can identify any trend, patterns, and seasonality in its sales. This will help understand what happened in its sales numbers.

The diagnostic analysis delves a little deeper to answer the question “why it happened?” and helps discover historical context through data. Continuing with the previous context, the question of “how effective was a promotional campaign based on the response in different geographies?” This type of analysis can help to identify causal relationships and anomalies in the data.

Predictive analysis is a little more complicated than the previous two discussed and answers “what can happen?” meaning looking into the future. The results from a predictive analysis should be treated as an estimate of the chance or probability of occurrence of that event. Widely used, a few examples are what the sales volume will be for the next time period? What is the propensity to buy for a new product release? Should I offer a loan to a particular applicant or no? This form of analysis uses knowledge and patterns from historical data to predict the future. In a world of uncertainty that businesses operate in, this is a very powerful tool to plan for the future.

The prescriptive analysis is almost the other end of the ladder, answering the question “how can it happen?” For example, businesses need the advice to understand the future course of action to take from all the available alternatives based on potential return and prescriptive analysis. For example, to achieve the outcome of a specific sale, it can suggest an alternative mix of investing in various types of promotions or media for advertising. This will be discussed more in-depth later with applications in supply chain, sales and marketing, and HR functions.

1.1.4 Use Cases in the Business

Businesses have come a long way in investing in groups that specialize only in data analytics. This group’s only objective is to fuel the business with insights, decisions, and knowledge using data. Businesses have invested in a strong data science talent, data infrastructure, tools, frameworks & methodologies, and industry-proven techniques. Change is constant, and the data analytics group is no different. They also innovate for the future by learning from the data and business problems thereby contributing to the bigger picture.

The banking and retail industries were pioneers in the Data Analytics Era, as they were in the digital era. Still, other sectors like manufacturing, telecom & communications, hospitality, and more recently public sector and government are significantly catching up and starting to using data science techniques. Finance, sales & marketing, IT functions that have adopted data science are faster than other functions. Figure 1.3 illustrates the various use cases in the business for analytics for decision making.

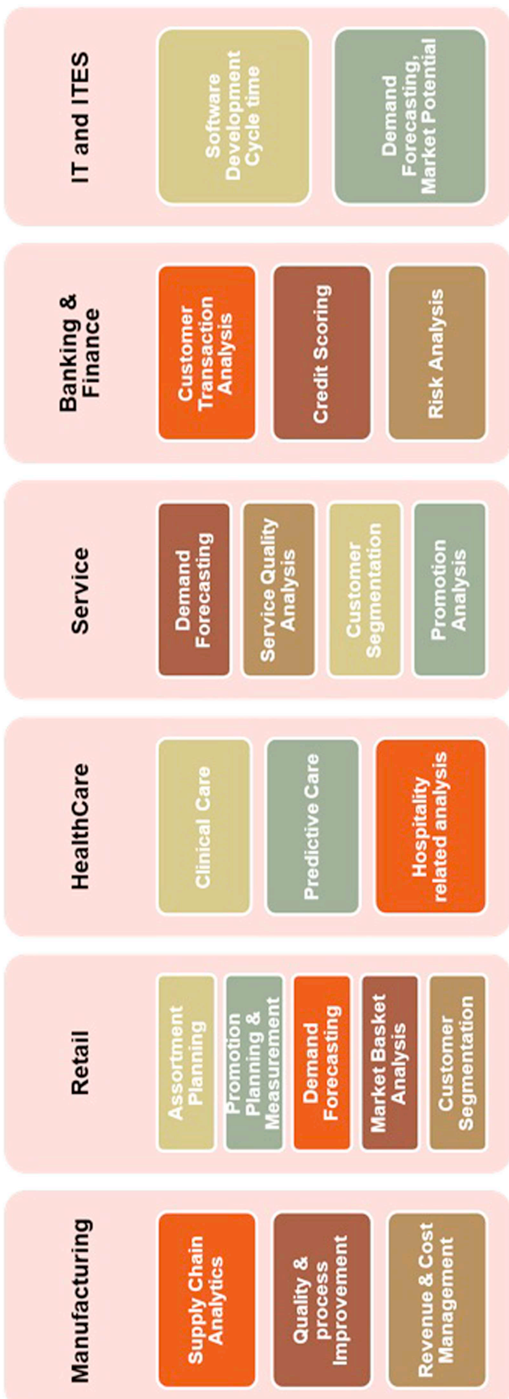

Figure 1.3 Industrywide application and its use cases (an illustration).

1.1.5 Data Analytics Process, Implementation and Measurement

The most important question to ask is how "data analytics" gets implemented in Industries? It all starts with a business problem. What are they? Industries collect a lot of data. What is the data telling? Are there commonalities that make eyebrows raise? Industries operate in an uncertain environment so decision-making is a challenge. Can industries forecast or predict the future? Be prepared! Be informed! is the key here. With limited resources like time, manpower, and material, how to get the best of them? Optimize! is the mantra. Industries cannot water all the roses in the garden, group points of data, help identify segments, make plans easier.

Solving a business problem using data science is a cycle of various tasks (Figure 1.4), namely:

- Always start with the question, "what is the business issue, problem to solve, goal to achieve, plan to support?"
- There is tons of data out there; gather the relevant data and prepare it for analysis.
- Explore the data, know what is data is saying as-is.
- Build a model to predict, associate, segment, and optimize as the case may be.
- Develop dashboard and visualize the results.
- Validate the findings with business and correct the findings if any.
- With the business teams, deploy the findings and measure results over time.

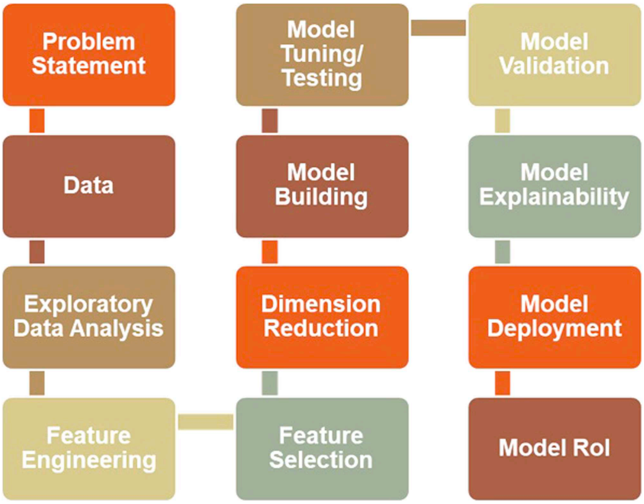

Figure 1.4 A typical data analytics process, implementation and measurement.

Data analytics is reshaped the way mankind historically was thinking of disaster response, business operations, media & entertainment, security and intelligence at all levels (Affine Analytics, 2021). The multitude of business exchanges, records, images, videos, sounds and signals are not simply being thought of as bits of data collected, marked, kept and retrieved, but as a possible wellspring of knowledge, which requires advanced analysis techniques that go from simple counts and aggregates to focus on finding relation and connected interpretations of the circumstance or situation present in the data.

The enormous collection of data, easily available hardware infrastructure, information management software, and advanced analytic capabilities have generated a celebrated moment in data analysis history. These connected trends mean that today mankind has the tremendous capacity and capability needed to analyze startling volume, variety, and velocity of data sets fast and cheaper than ever before. This body of knowledge is neither theoretical nor trivial. It represents a genuine attempt to leap forward and a fantastic opportunity to achieve big gains in efficiency, productivity, revenue, and profitability in any sphere. The business uses this to gain information, insight, and infer into its operations and thereby be ready to face the future with a bang!! Let us now look at how data analytics is used in various industries.

1.2 Data Science and Its Application in the Healthcare Industry

"Algorithm is the new doctor and data is the new drug," The "Healthcare Global Market Opportunities and Strategies to 2022" report (Business Wire, 2021) shows the Global Healthcare market at around \$8.4 trillion in 2018, with a 7.3% compound annual growth rate (CAGR) since 2014, and estimates to grow at 8.9% CAGR to around \$12 trillion by 2022.

One school of thought segments the healthcare market into largely healthcare service providers, pharmaceutical drugs manufacturing and distribution, medical equipment and supplies and veterinary care (Figure 1.5).

1.2.1 Data Types Generated in the Healthcare Sector

The move toward the adoption of technology in the healthcare sector has had a tremendously positive impact on the digitization of healthcare of both human health conditions and activities. This has created access to a large repository of knowledge and information. These milestones have presented various healthcare-related data through multiple resources (ETHealthWorld, 2019) like electronic health records (EHR), pharmaceutical research, healthcare digital platforms, medical imaging analysis, genomic sequencing, payer records, wearables and medical devices. Table 1.1 is an illustration of data sources in a general healthcare set-up.

Figure 1.5 Analytics in the healthcare ecosystem.

Source: https://healthtechmagazine.net/article/2017/04/healthcare-analytics-point-providers-patients-need-most-care

1.2.2 Analytics Use Cases in Healthcare

Data analytics is widely used in the healthcare industry today. Predicting the outcomes for a patient, fund allocation effectiveness and diagnostic technique improvement are only examples of how data analytics is transforming healthcare. The pharmaceutical industry is also experiencing this transformation through advanced analytics like machine learning and artificial intelligence (AI & ML). Drug discovery, a time-consuming and complex task with many parameters, is significantly improved through AI & ML. Pharma companies have been using data analytics to gain insights into their market, sales, consumers, and future predictions.

Healthcare analytics is used differently by each of its stakeholders. They include healthcare practitioners, government, healthcare providers, pharmaceutical companies and patients (Figure 1.6). Here are some use cases where analytics is used or in potential use in the industry, discussed by the stakeholder roles.

The (ETHealthWorld, 2019) healthcare practitioners are interested in clinical analytics, which aids in personalizing treatment, monitoring health, consulting remotely, and utilizing predictive health analysis to make decisions. Healthcare practitioners include doctors, therapists, caregivers, radiologists, biologists, and so on.

- Caregivers can monitor medicine refills for discharged patients through comprehensive dashboards and alerts. Analyze daily parameters during admission to classify levels of abnormality and predicting the reoccurrence of a health problem. It can prioritize critical care.

- Using artificial intelligence in medical imaging can classify medical images based on their criticality, this can help Radiologists (Dr. Sunil Kumar Vuppala, 2020) spend better time with patients rather than on medical reports. This will improve workflow where radiology is a key service and reduce misdiagnosis due to fatigue or other reasons.

(*Continued*)

Table 1.1 Data Sources in a General Healthcare Set-Up

Data Source	*Data Generated*	*Data Type*
Electronic records of patient's health	Clinical results, patient medical history, medical test results and patient prescription and diagnosis	numerical, text
Clinical records	Laboratory results like blood reports, tests	numerical, text
Diagnostic or monitoring instruments	Wide gamut of images (like CT Scan, MRI, X-Ray) to numbers (like patient vital signs) to text report (diagnosis)	image, text, voice, video, numerical
Insurance claims/ billing	Information on treatment, the cost of those services, expected payment and level of service	text, numerical
Pharmacy	Information on the fulfillment of medication orders	text, numerical, image
Human resources and supply chain	List of people employed and the role they play in the institution; resource allocationStock, storage and utilization of medical supplies	text, numerical, image
Digital wearables	Data generated about human vitals and activities coming from digital wearables like smartwatch, healthcare bands	text, numerical, image
Clinical trials	Results of drug testing, trials performed on drugs	image, text, voice, video, numerical
Healthcare surveys/projects	Samples, clinical records, analysis, results and findings from focused healthcare surveys/ projects	text, numerical, image
Sales	Sales data of medical insurance, pharmaceuticals, hospital beds, consultations and so on	numerical

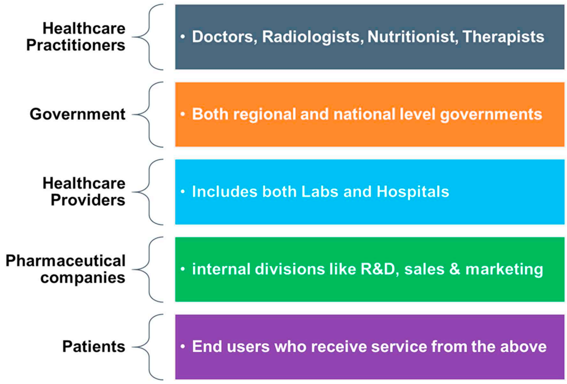

Figure 1.6 Stakeholders involved in healthcare.

- Data analytics takes claims data, providers' electronic health records (EHRs) and any other piece of information available to help physicians become more aware of the patients they're treating. They need not wait for patients to tell everything. This can help doctors learning about high-risk patients.

Using (ETHealthWorld, 2019) the data of patients, the government can identify health patterns and trends and analyze needs in healthcare at various geographical levels in a population. It also helps the government to draft health policies, identify interventions, plan programs for specific demographics, and prepare and respond to healthcare emergencies.

- Many health systems rely on government subsidies and support. Analytics help governments to have a clear picture of where the money is allocated and its reasons. Therefore, reducing the risk of resource wastage or unfair allocation of government subsidies.

- Health research institutes under the government study to prevent the spread of infectious diseases. Studying drug data and clinical trial results, and correlating data from pharmaceutical manufacturers, physicians and patients to build a model. For example, if a pandemic disease appears in a given population, data analytics can help find answers to questions like: how the population could get affected? how quickly could it spread? Actions that the government should take regarding quarantining an affected area, and what steps would be needed to control the pandemic before it spreads across the geography?

- Several governments also use data analytics to plan for population nutrition by promoting crops that help nutrition by season, region and prevalent health conditions of people.

For healthcare providers, like labs and hospitals, including insurance and claims processing companies' healthcare analytics entails mapping data into a form to better understand patients' health journey and know what contributes to improved healthcare outcomes.

■ Analytics helps understand the historical admission and discharge rates of patients helping to analyze the staff efficiency and productivity while able to predict and handle the different volumes of patients at a time.
■ Use data analytics to create consumer profiles, which will now allow the healthcare provider to send personalized messaging, improve retention and identify strategies meaningful for each individual. They use consumer behavioral patterns to draft impactful plans for care and keep their patients responsibly engaged through financial and clinical responsibilities.
■ More than ever, it is now when predictive tools are in high demand with hospitals, which are looking to reduce variation in their order patterns and supply utilization.
■ Hospitals are seeking to improve transition and deployment strategies of care coordination. Predictive analytics can warn the hospitals and other care providers when a patient's risk factors show a high probability of readmission, reducing financial burdens for the patient, hospitals and insurance companies alike.
■ Assess hospital claims and prescription fulfillment data to identify the potential for fraud by using predictive analytics to determine and notify at-risk claims.

The pharmaceutical and life sciences companies' various internal divisions such as finance, supply chain, R&D, sales, and marketing are benefitting from advanced analytics, AI & ML and using it in the areas of drug discovery, market assessment, brand knowledge, customer outreach and engagement.

■ Companies are currently using modeling to predict clinical outcomes, plan clinical trial designs, support the evidence of drug treatment effectiveness, optimize drug dosage, predict drug safety, and evaluate if any potential adverse event occurrence
■ It is not easy to release a drug out without an in-depth and rigorous process of creating the drug. It has to go through elaborate clinical trials before it is finally approved. Every pharmaceutical company must strictly follow this process before releasing and administering the drug to the patients. The use of data analytics tools, techniques, methodologies and algorithms

(*Continued*)

companies can shorten the time to go to market for the drug. Data analytics has played a significant role in developing a super effective, highly productive and impactful R&D pipeline.

- Healthcare analysts in these companies scrub both structured and unstructured data, including data coming from social media, text messages and pair that with classical tabular data to generate useable insights and work toward bringing better health outcomes for all stakeholders involved in the process.

Health data can encourage patients to be very proactive and involved in their care process. This is in a different point of view from the classical approach where doctors have the control and make the decisions.

- Digitalized periodical health, clinical and personal nutrition reports give individuals access to their health at their fingertips. Many healthcare apps have made this possible and empowered individuals to be even more focused on their health.
- The patients who are suffering from high blood pressure, asthma, migraine or other severe health problems, doctors can observe their lifestyle and bring changes if necessary through the data collected via wearable health-tracking devices.

1.2.3 Future and Challenges

Though data analytics has evolved and major health advantages are reached, there remain several challenges. First, large amounts of data produced remain in various decentralized systems that are accessible easily. Another challenge to conquer is the opposition of healthcare professionals against technological changes fearing risk or replacement. Information systems in the healthcare industry as a whole were not designed with analytics in mind to "get the data out" from it is not easy. The IT community in the healthcare system has not standardized these systems or their performance indicators. Within a given clinical information system, they are free to define their own data structures and standards for treatments and often do. Sharing and exchanging data through standard data formats requires a strong regulation in place and increases interoperability, privacy protection, and healthcare data exchange. Healthcare for analytical and research purposes is not created equal, needing standardization and quality improvement. Finally, data may be in a clinical narrative, images, and diagnosis as text that is more difficult to mine, requiring specialized talent and algorithms to bring them to a usable format.

The above challenges are wonderfully captured in this future-looking quote by Dr. Devi Shetty – nicknamed the Henry Ford of heart surgery, a renowned cardiac surgeon and entrepreneur who believes developments such as computerized diagnoses and technicians doing the work of highly trained medics are just around the corner. He said, "*Five to 10 years down the line, it will become mandatory for doctors to take a second opinion from the software before reaching the final diagnosis. This software will make doctors more efficient.*"

In the midst of the above-mentioned challenges, the future is brighter. Personal care, self-monitoring is becoming more and more popular. Today, individuals have access to enormous valuable health information, and, as a result, they have personally become involved in seeking information and improving their health. Market statistics say there are around 400,000 health apps that monitor a variety of personal health data like blood pressure, heart rate, sleep patterns, calory consumption, physical activity, cholesterol levels, and blood glucose among other parameters. This self-monitoring behavior is only set to increase, become more accurate, and alter the way how healthcare will be delivered.

1.3 Data Science and Its Application in the Retail and Retail E-Commerce

Global retail sales are projected to reach around $30 trillion by 2023, with a flat growth rate of about 4.5%, while retail e-commerce is projected to grow to $6.54 trillion by 2023. By 2023, the share of retail e-commerce will account for 22% of total retail sales. In this section, we will look at both the retail and retail e-commerce industries together.

In the market today, being customer-centric is everything. It demands that businesses stay a step ahead of their customers. Retail data gives information and insights to the retailers needed to stay valuable, ahead, and competitive. Thus making the retailers more informed about their customers and their behavior, habits, needs, wants, and spending patterns (Figure 1.7). This will enable the retailers to create a strong innovative retail experience for their customers. Retail data can analyze its customer data and segment them based on spending, demographics and behavior thus knowing which products sell are popular and in demand. This will help to make decisions and plans for products to promote.

1.3.1 Data Types Generated in the Retail and Retail E-Commerce Sector

Retail data is collected in raw form from several sources. Sales data comes mainly from point of sale (POS) or transaction systems, and this is a key source of data. However, additionally rich and valuable data is also generated from inventory, operational, campaign management, customer relationship management (CRM), supply chain,

Figure 1.7 Analytics in the retail and e-commerce ecosystem.

Source: https://www.comtecinfo.com/rpa/predictive-retail-analytics-use/

and partner relationship management (PRM) systems (Table 1.2). When analyzing, for decision making in retail, generally all of them or multiple parts are considered together. Retail e-commerce data is available through click stream, order, shipment management systems, logistics, supply chain and vendor systems. E-commerce collects much richer demographic data as compared to a traditional retail-like phone, email, physical address, IP addresses, and so on (pwc publications, 2016).

1.3.2 Analytics Use Cases in Retail and Retail E-Commerce

In this hyper-connected, information-driven era, data and data analytics are occupying a pivotal role in measuring and tracking growth and steering strategies for sustainable, profitable growth in the sector. Advances in digitization are swift, and the resultant changes in the behaviors of the consumer have the retail business redefine its operating model and its value proposition. While brick and mortar or physical retail is still a large share of total retail; its online counterpart continues to exhibit accelerated growth. Leading retailers have merged both their online and physical divisions such that the same teams oversee merchandising, planning and marketing for their physical stores and online businesses. Customers who like to shop in physical stores can now browse products and place orders on mobile devices, which then they can pick up their ordered products from a designated collection point. Analytics is assisting retailers to improve their profitability by enabling data-driven decision making for both their in-store and digital operations alike. Retail analytics can be studied with the help of the framework given in Table 1.3.

Table 1.2 Data Details of Retail E-Commerce Sector

Data Source	*Data Generated*	*Data Type*
Customer data	General data – Demographic details like name, age, address, IP address, phone, email, family members; Loyalty program data – status, program number	numerical, text, image
Sales	Detailed attributes of a: Product – name, brand, level, category, bundle, manufacturer; Price – list price, discount, sale price, promotion; Geography – city, store; Measure – units, value, volume	numerical
Inventory	Incoming stock, stock in-store, stock out rate	image, text, numerical
Logistics, supply chain	Delivery schedule, shipment, transport carrier, packaging details	text, numerical
Clickstream (retail e-commerce)	Onsite traffic metrics, IP address, record of every single click on the website, login details	text, numerical, image
Human resources	List of staff who are assigned to various tasks/units of the firm and their role. Attendance, timesheets	text, numerical, image
Promotions	Type of promotion, promotion material, duration, location, level, sponsor, budget; Email engagement attributes; Social media engagement attributes	numerical, image, text
Pricing & discount	Unit price, vendor price, profit, sale price, trade price, discount rate, discount level	numerical
Store surveillance	Store videos collected for surveillance	text, numerical, image, video
Surveys, house panels	Customer satisfaction surveys, house panel to measure consumption	numerical, text

Table 1.3 Framework to Study Retail Analytics

Area of Application	*Analytics Use Case*
Sales and marketing	Sales and demand forecasting using time series modeling are always necessary to understand the future and plan for it in the present. This will help businesses to optimize stock purchase, plan staff and promotions using predictive modeling. Retail companies both offline and online businesses want the Customer Lifetime Value (CLV) to plan personalized communication for those customers. In the same manner, supplier value is also equally important and predicting that will allow promoting high valued suppliers Using text mining and natural language processing (NLP) firms conduct e-commerce review analytics to understand the sentiments of customers. Even customer satisfaction surveys or now social media posts/tweets can be a rich source to use for sentiment analysis Using advanced clustering techniques retail companies now can develop and measure micro segmentations from price, store, customer and product data and (Ramesh Ilangovan, 2017) create multiple what-if scenarios for various clusters. This process helps identify optimal clusters to help improve planning, decision-making, and execution. Attribution modeling helps retailers understand how to optimize their marketing spend based on how customers reach and navigate through their sites. Dynamic pricing has been a go-to methodology to push retail sales, especially in intensely competitive segments like electronics. Using internal factors like supply, sales goals, margins, etc. and external factors like traffic, conversion rate, popularity of the products, etc. to build optimized pricing models such as price elasticity and ensemble models, product prices increase or decrease based on market situations.
Merchandising	Using predictive and prescriptive analytics to improve merchandising, which product where and when within the store with respect to demand patterns can be identified. This means assortment varies from one store to another. Association rule mining and Recommender algorithms tell a retailer what customers are buying together. This will help retailers place such products, categories next to one another or in websites that offer those products as a bundle or recommend the next product as a pop-up. This helps in promotion planning and pricing. Deep learning techniques are used by online retailers to identify and stop

(*Continued*)

Table 1.3 (Continued) Framework to Study Retail Analytics

Area of Application	*Analytics Use Case*
	fraudulent suppliers/sellers who sell defective, counterfeit products online which is an illegal activity.
Supply chain and logistics	In e-commerce business models, learning about returns is key because returns is a cost to the company. Using predictive analytics companies are now able to predict returns and also financially and logistically plan for it. Using optimization techniques, vehicle routing is planned for logistics and product delivery such that cost of transport is low and efficient reach to the location. This is sometimes absolutely required to meet delivery SLAs promised to the customer. Pricing optimization models are used to understand where and when to buy products from vendors. Today's retailers have a global model in sourcing and these pricing model allow them to get a good bargain on sourcing by combining it with demand forecasting data. Warehouse planning is another key decision and a backbone to the entire supply chain planning. Availability of space, distances to vendors, stores, closeness to highways, size of the warehouse are some key inputs into a warehousing optimization model.
Store operations	Using optimization techniques and location data companies can plan to optimize the mix of physical and online locations or to identify new store locations and plan franchise territories. Strong descriptive analytics resulting in a highly efficient dashboard for retail company management to understand store wise performance can help bring the right intervention for growth. There are thousands of stock-keeping units (sku) in a store both online or offline, optimization techniques are used to identify how much and what inventory to buy and stock or sell. Simple descriptive analysis dashboard to understand stock out scenarios and using predictive modeling be able to predict the stock-outs help in inventory planning for such items in store. Through IoT devices, cameras, and website navigation data is collected of customer movement with a store applicable both offline or online. This data is paired with advanced deep learning and computer vision analysis to optimize store layout, enhance merchandising, assess product performance, and improve customer experience.

1.3.3 Future and Challenges

New technologies (Datarade, 2020) like big data, artificial intelligence, machine learning, cloud infrastructure, new-age retail practices like 100% outsourced supply chain, e-commerce delivery, food-tech companies, mobile phone retail, QSRs and so on are on the move today. The most important question to be asked now is "why traditional companies have failed to keep pace with these modern developments?" Managers in these traditional companies continue to be doubtful about the claims made by these revolutionary technologies and often claims that they are greatly exaggerated. The knowledge of data analytics is often confined at most times to reporting and business intelligence. The few vendors of analytics who could have bridged this gap, in turn, lack business knowledge and understanding of challenges faced by retailers of today and are unconvinced about analytics application in their business beyond just tactics.

The party to retail business, "the customer" cannot (forbes, 2018) be seen as single community, but several communities across geographies are disparate in their habits and culture and expanding every day. The competition in the retail space is now not restricted to the neighborhood store, but many channels as mobile and web expand. Many me-too retailers imitate the more successful retailers, who were the early adopters of analytics, but only half-realizing its full benefits.

Finally, increasing conflict on pricing, discounts and range between traditional retail, e-commerce and modern retail is increasing pricing and margin pressure on companies as they juggle their volume growth ambitions with prices and margins, while trying to build their "Omni" presence across channels seldom realizing importance of each (Ramesh Ilangovan, 2017). While the end consumer may be benefitting in this conflict through lower prices, the pressure on margins across the value chain continues to grow. This makes us ask "Will analytics be the answer?"

1.4 Data Science and Its Application in the Banking, Financial Services and Insurance (BFSI) Sector

According to most of the studies conducted, out of the huge amounts (~2.5 quintillion [1018] bytes of data) of financial data collected nearly 85% of them were created in the last two years only. Further, with the continuous increase in the adoption of mobile technologies and IoT, **the scale of data was expected to grow exponentially** as stated above.

Due to the **increasing and changing customer expectations** and the **increased competition of Fintech players**, the financial services sector can simply not permit itself to leave those huge amounts of data unexploited (Joris Lochy, 2019). It is thus better for banks and insurers to leverage data science to maximize customer understanding and gain a competitive advantage (Figure 1.8).

Figure 1.8 Analytics in the banking, financial services and insurance (BFSI) ecosystem.

Source: http://fusionanalyticsworld.com/social-media-analytics-bfsi-part/

1.4.1 Data Types Generated in the BFSI Sector

Banks have huge amounts of data from customers in the form of their payments done online, customer profile data collected for KYC, deposits/withdrawals at ATMs, purchases at point-of-sales and others, but these all are not linked and hence at times not able to (Joris Lochy, 2019) utilize these rich data sets (Table 1.4). This while the financial services industry has been investing heavily for more than a decade in data collection and processing technologies (such as *data warehouses* and *business intelligence*) and is one of the forerunners in investments in data science areas.

1.4.2 Analytics Use Cases in BFSI

1. Identifying the change in customer behavior for personalized service:
   i. With digital usage, increasing more customer data is captured easily, which was not the case during the in-person nondigital era. These captured data points could be leveraged for building a personalized service that was present during the in-person connects.
   ii. Customers are comfortable using digital mediums for their bank transactions and purchases of stocks. Customers search using their mobile devices before buying any stock or products, and these footprints are also digitally captured. Now, the financial sectors are also reaching out to the customers via social media channels and selling their products/insurance premiums.
   iii. Now the stage has reached wherein the customers have expected more personalized and right information for their specific interest needs, than a generic recommendation. By integrating multiple data footprints of the specific customer along with like-minded customer data, this could be achieved.
2. Generate cross- and up-selling opportunities (Joris Lochy, 2019)

Table 1.4 Data Details of BFSI Sector

Data Source	*Data Generated*	*Data Type*
Customer data	General data – Demographic details like name, age, address, IP address, phone, email, family members Loyalty program data – status, program number	numerical, text, image
Product portfolio information	Credits, accounts, payments, securities, insurances…	Numerical, text
Clickstream (Banking site)	Onsite traffic metrics, IP address, record of every single click on the website, login details	text, numerical, image
Human resources	List of staff who are assigned to various tasks/ units of the firm and their role. Attendance, timesheets	text, numerical, image
Promotion details	Type of promotion, promotion material, duration, location, level, sponsor, budgetEmail engagement attributes Social media engagement attributes	numerical, image, text
Pricing & discount	Unit price, vendor price, profit, sale price, trade price, discount rate, discount level	numerical
ATM surveillance	ATM videos collected for surveillance	text, numerical, image, video

i. Through notifications, customer call agents or web ads, cross- and up-selling could be generated which is based on individual behavior of the customer.
ii. Customer is self-buying a bond on the stock market, this shows that it's a knowledgeable customer and would be open to the product: an up-selling opportunity for similar structured notes' basic info need not be explained.
iii. It is easier to understand about the customer that he does not have a home yet and is currently located at a house for sale, which helps to a selling opportunity for mortgage. These information could be obtained through the geo-location information and the public information or advertisements on the houses for sale.

   iv. Customer (Joris Lochy, 2019) modifies certain customer information (e.g. change of address due to move/relocation, change of civil status, e.g. following a wedding): selling opportunities for loans (e.g. mortgage, car loan) or insurances (home insurance, car insurance).
3. Helping with managing customer risks:
   i. *Cyber fraud prevention* can be addressed by continuously assessing the outliers or fraudulent transactions with restrictions and additional measures of security as required. These techniques would be useful for both physical money at branches as for the overall liquidity management of the bank/insurer.
   ii. Credit risk management improve the credit models at regular intervals based on customer patterns separately for private and corporate customers, thus having to improve credit scoring too. The models could help in deriving new rules, once the machine understands the data pattern, and this data can also be used to better manage the collateral of credits, thus also reducing credit risk for the bank.
   iii. Fraud detection for insurance: Many frauds happen during the claims of insurances; a good mechanism to identify these common fraudulent practices would help in managing the risks. Common past fraudulent data, sensor data, image of accident impact, etc. could help in guiding the adjudicator with the right estimate and reduce in insurance fraudulent claims.

1.4.3 Future and Challenges

New-age digital source data like the IoT data (e.g. sensors in home, equipments) in combination with the legacy old data sources (like transaction history, reports of companies) has a completely difficult task. Special care must be given to new data formats and data types because the underlying data structure changes may not be easily or readily updated on the trained models. The data privacy and intrusion, along with personalized services, are provided based on customer-specific data and their transactions, which is a fine line between being intrusive and helpful.

1.5 Statistical Methods and Analytics Techniques Used across Businesses

Statistical methods and analytics techniques help us systematically apply them to (Simran Kaur Arora, 2020) describe the data scope; modularize the data structure; condense the data representation; illustrate via images, tables and graphs and evaluate statistical inclinations and probability and to derive meaningful conclusions. These analytical procedures enable us to induce the underlying inference from data by eliminating the unnecessary chaos created by the rest of it.

These methods and techniques can be applied to analyze different styles of data like qualitative, quantitative, image, voice or speech, videos and text. Qualitative data mainly answers questions such as "why," "what" or "how." Each of these questions is addressed via quantitative techniques using scaling. Quantitative data is just numbers either point or with decimals. Now, data collection has evolved and so its analysis. Social media presents to us rich text-based data that is converted into numbers before analyzing. Images used for classification or recognition are converted into numbers based on color and pixels and then used for analysis. Video is nothing but multiple frames of pictures that are treated similar to images. Sounds and speech are converted into waves and frequency and that can, in turn, be converted into numbers before analysis.

There are numerous techniques to analyze data depending upon the business problem or question at hand, the type of data and the amount of data collected (Michael, J. A. Berry et al., 2011). Each of these techniques focuses on mining data, identifying meaningful information, deriving insights and transforming them into decision-making parameters.

In further sections, discussion will revolve around focused statistical methods and analytics techniques used in different functions of the business namely sales and marketing, HR and supply chain.

1.6 Statistical Methods and Analytics Techniques Used in Sales and Marketing

Sales and marketing are two business functions within an organization – they both lead generations and revenue along with creating an impact. The term sales refers to all activities that lead to the selling of goods and services. And marketing is the process of getting people interested in the goods and services being sold. Marketing informs and attracts leads and prospects to the business or product or service. Sales, on the other hand, works directly with prospects to reinforce the value of the company's solution to convert prospects into customers (Figure 1.9). The fundamental distinction between the two departments is that the marketing department's efforts cost the organization expenses, whereas the sales department generates revenue to the company.

1.6.1 Data Types Generated in Sales and Marketing Function

Sales data is usually information used to manage sales and key trends around the pipeline. The data may concern from market to opportunities and deals to the third party to actual sales performance. Marketing data as shown in Table 1.5 encompasses data collected from leads, spends for campaigns, advertising, branding and can be used to improve product development, promotion, sales, pricing, distribution and related strategies (Simran Kaur Arora, 2020).

Figure 1.9 Analytics in the sales and marketing ecosystem.

Source: https://talkinginfluence.com/2019/12/12/improve-influencer-analytics/

1.6.2 Statistical Methods and Analytical Techniques

As defined by SAS,

> Marketing analytics comprises the processes and technologies that enable marketers to evaluate the success of their marketing initiatives by measuring performance using important business metrics, such as ROI, marketing attribution and overall marketing effectiveness. In other words, it tells you how your marketing programs are performing.

Unlike marketing, sales have always been number-driven and now with the explosion of data and computational power; sales analytics has become central to any large sales organization. So, what is sales analytics? Sales analytics is the process used to identify, model, understand and predict sales trends and sales results while helping in the understanding of these trends and finding improvement points. The best practice is to closely tie all activities to determine revenue outcomes and set objectives for your sales team.

Sales and marketing analytics are essential to unlocking commercially relevant insights, increasing revenue and profitability and improving brand perception. With the help of the right analytics, you can uncover new markets, new audience

Table 1.5 Data Details of Sales and Marketing

Data Source	*Data Generated*	*Data Type*
Market share/size	Mostly data from syndicated research studies or secondary sources of data collected by internal teams	numerical, text, image
Quote & config	All quotes given to customers/ channel partners during opportunity stages (before sale)	numerical, text
Campaign management	All sales and marketing related campaign data like budgets, programs, expenses, program details	image, text, numerical
Compensation	Both Channel Partner and Sales persons' compensation details like target, achievement variable pay	numerical
Partner relationship management (PRM)	All information about a contracted past, prospective and current partners. Also called Master data of channel partners. Should contain demographic and contact data. Often stored in PRM systems	text, numerical, image
Customer relationship management (CRM)	All information about a company's past, current and prospect customers. Also called Master data of customers. Should contain demographic and contact data. Often stored in CRM systems	text, numerical, image
Contracts	Lists of all contracts signed by the company with its partners and customers to be used for sales and marketing purpose	text, numerical, image

(*Continued*)

Table 1.5 (Continued) Data Details of Sales and Marketing

Data Source	*Data Generated*	*Data Type*
Digital click stream	Data generated from the company's website, social media, software, knowledge management and campaign related web pages	text, numerical, image
Pipeline/opportunity/ deals	Mostly part of a CRM system will contain sales leads or opportunities	text, numerical
Third party	Data used to enrich data from internal transaction systems, mostly syndicated studies, surveys	text, numerical, image
Sales	Actual sales performance by product, customer, geography for a time-period and a measure like value, volume or unit. Should be available in the company's Order, Shipment and Revenue management ERP systems	numerical

niches, areas for future development and much more. Figure 1.10 shows the possible statistical methods and analytical techniques used in sales and marketing.

Sales and marketing analytics comprises analytics for each silo and at various levels including at a strategy level, sales function, marketing function, consumers and partners and not to leave out the sales representatives themselves.

1. Sales and marketing strategy – Analytics supporting sales and marketing strategy are today part of every analytics CoE. They are generally done at a corporate level and analysis provided at a geography, business unit, customer segment and sometimes product level too. It all starts with knowing the market size and of the most popular analytical techniques is TAM (total addressable market), which is a funnel-like analysis to identify and quantify the overall opportunity in the market that the business can address. In a more mature business generally White space analysis is done to bridge gaps with a new product, service release. Knowing your competitor and their strategies is like half the sale done. Win–Loss analysis using text mining from

Figure 1.10 Statistical methods and analytical techniques used in sales and marketing.

salespersons' comments from CRM systems is a wealth of information in the hands of the business to plan a competitive attack. Price wars are everywhere, especially with the rise in e-commerce business. At what price will the customer stop buying? How elastic is my price for a product? Are questions answered from price elasticity analysis? Price-sensitive industries have a full pricing analytics team to feed business teams with decision-making insights.

2. Marketing analytics – Marketing analytics can be divided into analytics done on above the line (ATL) and below the line (BTL) activities. "ATL" meaning that the strategy is going to be deployed around a wider target audience, e.g. television, radio or billboards. ATL is most applicable when a product is directed at a broader spectrum of consumers. With so many options and limited resources where to focus most to maximize RoI question is answered using marketing mix modeling, sometimes media mix modeling thanks to the tech burst and the rich availability of media platforms today. While "BTL" strategies are going to target a specific group of potential consumers using tools like direct emailing or direct product demonstrations. Test control analysis, promotion effectiveness techniques drive the use of the right technique for the right product for the right audience.
3. Sales analytics – sales analytics can be understood using the sales towers starting with qualified leads called opportunities where win/loss analysis through ensemble techniques are used to predict which opportunity will win and optimization techniques used to identify where to spend time and

available funds. During quoting its always measuring and identifying opportunities for attaching meaning for every value of the main product sold what can be attached with that thereby increasing the overall bill value. A salesperson's time, funds and other resources are limited using which maximizing revenue is the key, so forecasting sales, possible trend and seasonality are very critical for businesses to better plan go to market strategies. Businesses use time series forecasting techniques or regression techniques for this depending on the criticality and data availability. After-sales predicting annuity sales like renewals in insurance, financial products or cross-sell/upsell possibilities is an important fuel for growth. The propensity to buy during a sales campaign offer will tell businesses to give offers for that work.

4. Consumer analytics – Many book articles combine customer analytics with sales or marketing analytics, but there is significant merit to call it out separately thanks to its drive and importance. Understanding consumer segments help in driving focused sales or marketing strategies. Unsupervised techniques are used to group customers and then profile them to understand them deeper followed by building targeted campaigns for them. Any product launch or upgrade that happens in the business is oftentimes followed with a propensity to buy a predictive model to target the highest-scoring consumers. This will help skim the market and realize quick sales. Acquire, build and retain consumers is the strategy of a growing business, so they focus on predicting customer lifetime value (CLV). CLV means loyalty to the business defined by purchase, repurchase and referral to other consumers. Another business question on why has a particular customer not purchasing is answered via churn analytics. Businesses need to know if a customer will churn and when are they likely to churn? Such that an intervention strategy or program could be designed to this target.
5. Partner analytics – With a significant share of businesses depending on the channel for their sales, partner-focused analytics are gaining importance. Businesses mine data from their PRM systems to segment their partners based on their value to the business using Recency, Frequency, Monetary technique or other unsupervised techniques like K Means. Onboarding a partner into the business requires a contract and legal clearances. With go-to-market pressures analytics teams often find analytics to their rescue to identify opportunities to reduce onboarding turnaround time (TAT) using simple descriptive statistic techniques. It is better to stop sales fraud before it occurs; predictive analytics techniques like logistic regression and artificial neural networks are used to predict if a sales deal is likely to turn into a fraud one.
6. Sales representative analytics – Businesses want their sales representatives to spend most of their time meeting customers and spend time in selling. Analytics teams in the sales operations often conduct time spent analysis to benchmark their sales representatives and their time spent on sales-related activities with that of the industry.

1.6.3 Future and Challenges

Over the years, as businesses expand into digitalization, the need for advanced targeting and tracking is becoming the main focus of sales and marketing initiatives. With the higher demand for efficient analytics solutions, the challenges started to rise.

The new technologies were typically deployed in isolations, and the result was a huge set of tools and platforms of a disconnected data environments. There were always be instability and mismatching results coming from the different platforms causing data discrepancies. At the end of the day, you will be facing the issue of which data source is the most reliable for analysis leading to decision making. Each business has its own technology stack and infrastructure therefore connecting internal company sales and marketing data with online data is sometimes one of the biggest challenges for marketers. Businesses should establish a strong privacy policy to address legal and ethical concerns for sales and marketing data, analytics and its implementation. Privacy laws like GDPR and other issues may affect some industries more strongly than others.

1.7 Statistical Methods and Analytics Techniques Used in Supply Chain Management

By the end of 2010, most of the companies had integrated all of their own and external resources available in the market. This integration has enabled their working pattern of a system for any quick response to the needs in the market. Creating a visualization dashboard to help in taking some quick decisions for ad hoc solutions has been made possible and such a system is referred to as supply chain management (SCM) system (Figure 1.11).

In this decade, the advancement in the supply chain field has been driven through implementing advanced analytics (data science) methods like time-series forecasting, route optimization techniques and hierarchical structuring. Business decision makers are now able to understand the fact on how data science is helping their companies to make the right decisions at the right time saving millions of dollars.

1.7.1 Data Types Used in the SCM

The supply chain is a great place to apply analytics for gaining a competitive advantage because of the uncertainty, complexity, varied data sources and the significant role it plays in the overall cost structure and profitability for almost any firm (Table 1.6).

Figure 1.11 Analytics in the supply chain management ecosystem.

Source: https://medium.com/o4s-io/how-supply-chain-analytics-can-transform-business-2fc9e16bc9ac

1.7.2 Analytics Use Cases in SCM

A supply chain management is a network of multiple businesses and relationships. Supply chain users need to be aware of the benefits given by the data analytics for their operations. Some of the key areas of SCM where data science plays a vital role are:

1. Demand prediction: Demand forecasting is essential in planning for sourcing, manufacturing, logistics, distribution and sales, which are is done in various forms in the past decade; with the advent of data science, it is becoming more effective and easier to handle these. In the current stages, firms are even starting to predict the demands for new products that are yet to be launched too this is helping in decisions like manufacturing, procurement planning and strategizing the OEMs. There is huge volatility in demand, which causes problems in the entire supply chain from supply planning, production and inventory control to shipping, hence it's challenging and equally important to forecast helping in planning at every level in the organizations, regions, stores, etc.
2. Optimal route identification: Route optimization is a very important factor, and it is more than just identifying the shortest route from point A (source) to point B (destination). For a perfect route optimization to handle the flow of supply chain and control it efficiently, the following are to be adhered to:
   1. planning to be done to manage the entire fleet;
   2. set processes and adherence to them;

Table 1.6 Data Details of SCM

Data Source	*Data Generated*	*Data Type*
Supplier provided data	General data – Demographic details like name, age, address, IP address, phone, email, family members Loyalty program data – status, program number,	numerical, text, image
Transactional sales data	Detailed attributes: Product – name, brand, level, category, bundle, vendor/ manufacturer Price – list price, discount, sale price, promotion Geography – city, store Measure – units, value, volume	numerical
Public open data	GPS tracking data - Vehicle info, route details Government data – Policy info, guidelines details	text, numerical
Third-party/OEM proprietary data	Incoming stock, stock in-store, stock out rate	numerical, text, image
Logistic data	Delivery schedule, shipment, transport carrier, packaging details	text, numerical

iii. Real-time traffic information updates, helping to change the route directions;

iv. Foresee and flexibility to handle any ad hoc situations.

Already many works have been proceeding in the above-mentioned area, and some have also reached advanced stages during this decade.

3. Space/inventory optimization: It's a trade-off between how many items to be stocked to handle the supply-demand effectively. Challenges faced in the decision could be out of stock, over dumped stocks, planning of space utilizations. A better understanding of the moving/non-moving items, cost benefits. With the right data points in hand, we could easily maximize the space utilization thereby improve productivity with an increase in profits.
4. Consignments track and trace: Each of the consignments is now attached with unique bar codes and using a proper RFID reader, complete info is

retrieved which could track and identify where the consignments are currently in transit. But just providing this info is just one stage, but currently, companies are using this and understanding better various stages in the complete SCM logistics cycle and optimizing the cycle time that is taking longer than expected.

1.7.3 Future and Challenges

To maintain the quality of customer service, the challenge is to adapt to a fast-changing environment and the delays during transit, probably due to unforeseen challenges. Finally, in the complete SCM cycle, data science is helping to understand and create transparency in the complete flow. This is also helping to set correct SLA and adhere to them to have increased inefficiency in the processes involved, thereby increasing customer satisfaction.

1.8 Statistical Methods and Analytics Techniques Used in Human Resource Management

In the current digital era, it is now evident that the HR team should use the available tools to aid in their core activities – whether it is talent acquisition, resource optimization, training & development or employee payments (Figure 1.12). Data science in HR is used for effective improvement in overall employee performance who have several open questions:

- How can we acquire the right talent? How can we decide which profiles are right for the job description?
- How can we identify the highly skilled people and retain them?
- How can we retain and engage our top talent?
- How can we leverage social network data for human resources operations?

1.8.1 Data Types Generated in Human Resource Management

HR analytics is the process of addressing a strategic HR concern using HR data (and business and external data if necessary), thereby identifying the HR issues and further preparing a subsequent action plan. Table 1.7 shows the data, data type and data source relevant to human resource management.

Figure 1.12 Analytics in the human resource management ecosystem.

Source: https://www.peoplematters.in/article/hr-analytics/workforce-analytics-how-mature-are-organizations-13135

Table 1.7 Data Details of Human Resource Management

Data Source	*Data Generated*	*Data Type*
Employee data	General data – Demographic details like name, age, address, IP address, phone, email, project details	numerical, text
Behavioral data	Detailed attributes, Performance details, previous ratings, payroll info	Numerical, text
Social media data	Social Engagement data, sentiment index data, campaigns data	text, numerical
Third-party/OEM proprietary data	In-premises camera video data	image, video

1.8.2 Analytics Use Cases in Human Resource Management

1. Employee profiling and segmentation
   i. All employees are not similar; their career planning/benefits should be on a case by case basis, hence there is a need to understand the exiting workforce better.

   ii. Understand the demographics, skills, educational background, experience and designation, and all these can be combined with information on each roles and responsibilities.
   iii. This would help in coming up with planned targeted programs for each segment profile and help in achieving better relationship and higher satisfaction from employees.
2. Employee attrition model
   i. Employee attrition is a major issue, as this has various other impacts like high financial costs, productivity losses, negative impact on customer service, loss of expertise, loss of business opportunities, job dissatisfaction of remaining employees and a bad image of the organization.
   ii. Devise a retention strategy for potential churners. To identify potential churners, we need a predictive model that can assist us with this.
   iii. The model can help in determining future possibilities and reducing employee turnover if desired. KPIs such as employee satisfaction, staff advocacy, etc. are helpful in this analysis.
3. No shows – post-offer roll out
   i. To estimate the employees joining probability, post-offer role out. A plan should be devised accordingly to reduce the no-show percentage.
   ii. Use the existing no-show data and accepted offers data across various skill set, job roles and experience to understand if there is any similar pattern or trend.
   iii. Once there are some identical patterns, we could validate and devise a proper mitigation plan to help reduce the no-shows.
4. Employee sentiment analysis
   i. Healthy presence on social media platforms via running campaigns, posting ideas, shouting achievements and initiatives increases the social HR brand for employees to follow and employers to measure.
   ii. Social identification for potential candidates and understanding resource profile. Empowering with an additional information.
   iii. Helps to define and manage social engagement with employees, accurately measuring sentiment and understanding each employee's social sentiment index.

1.8.3 Future and Challenges

Being transparent is better, and one way to achieve this is to have the correct data in front during discussions. Please make sure everyone in the organization knows and understands it. This would help during the challenging times and bringing in a positive organizational culture too. It is always advisable to start small and grow with the challenges to handle along the way – have conversations with employees, record their responses, add managers in the loop, involve various functions, make a plan, share it with everybody and commit to it. HR analytics will help you monitor and improve employee engagement, employee retention, employee wellness, employee productivity, employee experience and work culture.

References

Affine Analytics. (2021). The Evolution of Data Analytics – Then, Now and Later. Retrieved Oct 15, 2021 from https://www.affineanalytics.com/blog/the-evolution-of-data-analytics-then-now-and-later/

Bridgei2i. (2021). Interview with BRIDGEi2i CEO and Co-Founder, Prithvijit Roy in conversation with Neeraj Krishnamoorthy. Retrieved Oct 15, 2021 from https://bridgei2i.com/interview-with-bridgei2i-ceo-prithvijit-roy/

Business Wire. (2021). The "Healthcare Global Market Opportunities And Strategies To 2022". Retrieved Oct 15, 2021 from https://www.businesswire.com/news/home/20190625005862/en/11.9-Trillion-Global-Healthcare-Market-Key-Opportunities

Driving Retail Growth by Leveraging Analytics, PWC Publications 2016. Retrieved from https://www.pwc.in/assets/pdfs/publications/2016/profitable-growth-for-retail-businesses-online.pdf

Enabling various types of Healthcare Data to build Top 10 DL applications Dr. Sunil Kumar Vuppala Apr 5. Retrieved from https://medium.com/@sunil.vuppala/enabling-various-types-of-healthcare-data-to-build-top-10-dl-applications-f5c6f45eddba

ETHealthWorld. (2019). Healthcare Analytics: How Data Is Transforming the Healthcare Landscape in India, ETHealthWorld July 19, 2019. Retrieved from https://health.economictimes.indiatimes.com/news/health-it/healthcare-analytics-how-data-is-transforming-the-healthcare-landscape-in-india/70288906

Foster Provost & Tm Fawcett. (2018). *Data science for business*. Newton, MA: O'REILLY.

Gartner. (2018). *Analytics*. Retrieved from Gartner: https://www.gartner.com/it-glossary/analytics/

Healthcare Analytics Point Providers to Patients that Need the Most Care by Juliet Van Wagenen. Retrieved from https://healthtechmagazine.net/article/2017/04/healthcare-analytics-point-providers-patients-need-most-care

How Retailers Can Make The Most of Their Data, Hugo Moreno Contributor, Thought Leaders Contributor Group, Leadership Strategy, Jun 28, 2018. Retrieved from https://www.forbes.com/sites/forbesinsights/2018/06/28/how-retailers-can-make-the-most-of-their-data/#4bead88d453c

How Supply Chain Analytics can Transform Business? By O4S Team. Retrieved from https://medium.com/o4s-io/how-supply-chain-analytics-can-transform-business-2fc9e16bc9ac

Ilangovan, R. (2017). Retail Analytics Trends — 2017 and beyond Ramesh Ilangovan May 15, 2017. Retrieved from https://towardsdatascience.com/retail-analytics-trends-2017-and-beyond-374bc6627cb0

Linoff, G.S., and Berry, M.J.A. (2011). *Data mining techniques*. New York: Wiley Publishing, Inc.

Lochy, J. (2019). Big Data in the Financial Services Industry - From Data to Insights, Sep 2019. Retrieved from https://www.finextra.com/blogposting/17847/big-data-in-the-financial-services-industry---from-data-to-insights

Predictive Retail Analytics – Why Should You Use It? by Chandra Gogineni. Retrieved from https://www.comtecinfo.com/rpa/predictive-retail-analytics-use/

Search Data Management Tech Target Data Analytics (DA) by Margaret Rouse. Retrieved from https://searchdatamanagement.techtarget.com/definition/data-analytics

Social Media Analytics – BFSI by Kalyan Banga. Retrieved from http://fusionanalyticsworld.com/social-media-analytics-bfsi-part/
Six Ways to Improve Your Influencer Marketing Analytics by Kayla Matthews. Retrieved from https://talkinginfluence.com/2019/12/12/improve-influencer-analytics/
The Ultimate Guide to Retail & Commerce Data 2020, Datarade. Retrieved from https://datarade.ai/data-categories/retail-commerce-data/guide
Types of Analytics in Human Resource Management. Retrieved from https://talentedge.com/articles/analytics-hr-management/
What is Data Analysis? Methods, Techniques & Tools; Posted in Data Analytics by Simran Kaur Arora Apr, 2020. Retrieved from https://hackr.io/blog/what-is-data-analysis-methods-techniques-tools
Workforce Analytics: How Mature are Organizations? Retrieved from https://www.peoplematters.in/article/hr-analytics/workforce-analytics-how-mature-are-organizations-13135

Data Science and Its Applications

Affine Analytics . (2021). The Evolution of Data Analytics – Then, Now and Later. Retrieved Oct 15, 2021 from https://www.affineanalytics.com/blog/the-evolution-of-data-analytics-then-now-and-later/

Bridgei2i . (2021). Interview with BRIDGEi2i CEO and Co-Founder, Prithvijit Roy in conversation with Neeraj Krishnamoorthy. Retrieved Oct 15, 2021 from https://bridgei2i.com/interview-with-bridgei2i-ceo-prithvijit-roy/

Business Wire . (2021). The “Healthcare Global Market Opportunities And Strategies To 2022”. Retrieved Oct 15, 2021 from https://www.businesswire.com/news/home/20190625005862/en/11.9-Trillion-Global-Healthcare-Market-Key-Opportunities

Driving Retail Growth by Leveraging Analytics, PWC Publications 2016. Retrieved from https://www.pwc.in/assets/pdfs/publications/2016/profitable-growth-for-retail-businesses-online.pdf

Enabling various types of Healthcare Data to build Top 10 DL applications Dr. Sunil Kumar Vuppala Apr 5. Retrieved from https://medium.com/@sunil.vuppala/enabling-various-types-of-healthcare-data-to-build-top-10-dl-applications-f5c6f45eddba

ETHealthWorld . (2019). Healthcare Analytics: How Data Is Transforming the Healthcare Landscape in India, ETHealthWorld July 19, 2019. Retrieved from https://health.economictimes.indiatimes.com/news/health-it/healthcare-analytics-how-data-is-transforming-the-healthcare-landscape-in-india/70288906

Foster Provost & Tm Fawcett . (2018). Data science for business. Newton, MA: O’REILLY.

Gartner . (2018). Analytics. Retrieved from Gartner: https://www.gartner.com/it-glossary/analytics/

Healthcare Analytics Point Providers to Patients that Need the Most Care by Juliet Van Wagenen. Retrieved from https://healthtechmagazine.net/article/2017/04/healthcare-analytics-point-providers-patients-need-most-care

How Retailers Can Make The Most of Their Data, Hugo Moreno Contributor, Thought Leaders Contributor Group, Leadership Strategy, Jun 28, 2018. Retrieved from https://www.forbes.com/sites/forbesinsights/2018/06/28/how-retailers-can-make-the-most-of-their-data/#4bead88d453c

How Supply Chain Analytics can Transform Business? By O4S Team. Retrieved from https://medium.com/o4s-io/how-supply-chain-analytics-can-transform-business-2fc9e16bc9ac

Ilangovan, R. (2017). Retail Analytics Trends — 2017 and beyond Ramesh Ilangovan May 15, 2017. Retrieved from https://towardsdatascience.com/retail-analytics-trends-2017-and-beyond-374bc6627cb0

Linoff, G.S. , and Berry, M.J.A. (2011). Data mining techniques. New York: Wiley Publishing, Inc.

Lochy, J. (2019). Big Data in the Financial Services Industry - From Data to Insights, Sep 2019. Retrieved from https://www.finextra.com/blogposting/17847/big-data-in-the-financial-services-industry---from-data-to-insights

Predictive Retail Analytics – Why Should You Use It? by Chandra Gogineni. Retrieved from https://www.comtecinfo.com/rpa/predictive-retail-analytics-use/

Search Data Management Tech Target Data Analytics (DA) by Margaret Rouse. Retrieved from https://searchdatamanagement.techtarget.com/definition/data-analytics

Social Media Analytics – BFSI by Kalyan Banga. Retrieved from http://fusionanalyticsworld.com/social-media-analytics-bfsi-part/

Six Ways to Improve Your Influencer Marketing Analytics by Kayla Matthews. Retrieved from https://talkinginfluence.com/2019/12/12/improve-influencer-analytics/

The Ultimate Guide to Retail & Commerce Data 2020, Datarade. Retrieved from https://datarade.ai/data-categories/retail-commerce-data/guide

Types of Analytics in Human Resource Management. Retrieved from https://talentedge.com/articles/analytics-hr-management/

What is Data Analysis? Methods, Techniques & Tools; Posted in Data Analytics by Simran Kaur Arora Apr, 2020. Retrieved from https://hackr.io/blog/what-is-data-analysis-methods-techniques-tools

Workforce Analytics: How Mature are Organizations? Retrieved from https://www.peoplematters.in/article/hr-analytics/workforce-analytics-how-mature-are-

organizations-13135

Industry 4.0: Data and Data Integration

Apachespark . (2020). Retrieved from spark: https://spark.apache.org/
ETL . (2020, April 16). Retrieved from Extract, transform, load: https://en.wikipedia.org/wiki/Extract,_transform,_load
Flink . (2020). Retrieved from apache Flink: https://flink.apache.org/
kafka . (2020). Retrieved from kafka: https://kafka.apache.org/
Keys, D. (2020, April 25). Daniel Keys Moran quotes. Retrieved from BrainyQuote.com: https://www.brainyquote.com/quotes/daniel_keys_moran_230911
Mark Cotteleer, B.S. (2020). Forces of change: Industry 4.0. Retrieved from https://www2.deloitte.com/us/en/insights/focus/industry-4-0/overview.html
matillion . (2020). Retrieved from matillion: https://www.matillion.com/
McDaniel, S. (2019, October 2). Data source. Retrieved from talend.com: https://www.talend.com/resources/data-source/
McDaniel, S. (2019, September 16). ETL architecture. Retrieved from talend.com: https://www.talend.com/resources/etl-architecture/
Microsoft . (2020). SSIS. Retrieved from Microsoft: https://docs.microsoft.com/en-us/sql/integration-services/sql-server-integration-services?view=sql-server-ver15
Pearlman, S. (2019, March 14). Batch processing. Retrieved from talend.com: https://www.talend.com/resources/batch-processing/
Pearlman, S. (2019, September 3). Retrieved from https://www.talend.com/resources/what-is-data-integration/
Real-time big data . (2020). Retrieved from Talend: https://www.talend.com/products/big-data/real-time-big-data/
Talend . (2020). Retrieved from talend.com: https://www.talend.com/products/data-integration/
Tuovila, A. (2019, May 24). ObsoleteInventory. Retrieved from investopedia.com: https://www.investopedia.com/terms/o/obsoleteinventory.asp

Forecasting Principles and Models: An Overview

Hanke, J. E. , & Wichern, D. (2014). Business forecasting (9th ed.). Essex, UK: Pearson Education Limited.
Hoshmand, A. R. (2010). Business forecasting – A practical approach (2nd ed.). New York: Routledge Publications.

Breaking Technology Barriers in Diabetes and Industry 4.0

Akturk, H. K. , Giordano, D. , Champakanath, A. , et al. (2020, Apr). Long-term real-life glycaemic outcomes with a hybrid closed-loop system compared with sensor-augmented pump therapy in patients with type 1 diabetes. Diabetes Obes Metab, 22(4), 583–589.
Anjana, R. M. , Deepa, M. , Pradeepa, R. , et al. (2017, Aug). ICMR–INDIAB Collaborative Study Group. Prevalence of diabetes and prediabetes in 15 states of India: Results from the ICMR-INDIAB population-based cross-sectional study. Lancet Diabetes Endocrinol, 5(8), 585–596.
Anjana, R. M. , Shanthi Rani, C. S. , Deepa, M. , et al. (2015, Aug). Incidence of diabetes and prediabetes and predictors of progression among Asian Indians: 10-year follow-up of the Chennai Urban Rural Epidemiology Study (CURES). Diabetes Care, 38(8), 1441–1448.

Das, A. K. (2015, Apr). Type 1 diabetes in India: Overall insights. Indian J Endocrinol Metab, 19(Suppl 1), S31–S33.
Das, J. , Holla, A. , Das, V. , Mohanan, M. , Tabak, D. , & Chan, B. (2012). In urban and rural India: A standardized patient study showed low levels of provider training and huge quality gaps. Health Aff Millwood, 31, 2774–2784.
Forouhi, N. G. , & Wareham, N. J. (2014). Epidemiology of diabetes. Medicine (Abingdon), 42, 698–702.
Haffner, S. M. , Lehto, S. , Rönnemaa, T. , Pyörälä, K. , & Laakso, M. (1998, Jul 23). Mortality from coronary heart disease in subjects with type 2 diabetes and in nondiabetic subjects with and without prior myocardial infarction. N Engl J Med, 339(4), 229–234.
Hamet, P. , & Tremblay, J. (2017, Apr). Artificial intelligence in medicine. Metabolism, 69S, S36–S40.
IDF. International Diabetes Federation . (2015). IDF Diabetes Atlas Seventh edition. https://www.diabetesatlas.org/upload/resources/previous/files/7/IDF%20Diabetes%20Atlas%207th.pdf, ISBN: 978-2-930229-81-2
Jha, V. (2004). End-stage renal care in developing countries: The India experience. Ren Fail, 26(3), 201–208. doi:10.1081/JDI-120039516
Kher, V. (2002). End-stage renal disease in developing countries. Kidney Int, 62(1), 350–362. doi:10.1046/j.1523-1755.2002
Kumar, K. M. (2015, Apr). Incidence trends for childhood type 1 diabetes in India. Indian J Endocrinol Metab, 19(Suppl 1), S34–S35.
Kumar, P. , Krishna, P. , Reddy, S. C. , et al. (2008, Nov). Incidence of type 1 diabetes mellitus and associated complications among children and young adults: Results from Karnataka Diabetes Registry 1995–2008. J Indian Med Assoc, 106(11), 708–711.
Leasher, J. L. , Bourne, R. R. , Flaxman, S. R. , et al. (2016, Sep). Global estimates on the number of people blind or visually impaired by diabetic retinopathy: A meta-analysis from 1990 to 2010. Diabetes Care, 39(9), 1643–1649.
Padhy, S. K. , Takkar, B. , Chawla, R. , & Kumar, A. (2019). Artificial intelligence in diabetic retinopathy: A natural step to the future. Indian J Ophthalmol, 67(7), 1004–1009. doi:10.4103/ijo.IJO_1989_18
Ramachandran, A. , Snehalatha, C. , & Krishnaswamy, C. V. (1996, Oct). Incidence of IDDM in children in urban population in southern India. Madras IDDM Registry Group Madras, South India. Diabetes Res Clin Pract, 34(2), 79–82.
Rumbold, J. M. M. , O’Kane, M. , Philip, N. , & Pierscionek, B. K. (2020, Feb). Big Data and diabetes: The applications of Big Data for diabetes care now and in the future. Diabet Med, 37(2), 187–193.
Sakhuja, V. , & Sud, K. (2003). End-stage renal disease in India and Pakistan: Burden of disease and management issues. Kidney Int Suppl, 83(83), S115–S118.
Singh, N. , Armstrong, D. G. , & Lipsky, B. A. (2005, Jan 12). Preventing foot ulcers in patients with diabetes. JAMA, 293(2):217–228.
Swaminathan, K. , & Thangavel, G. (2015). Pesticides and human diabetes: A pilot project to explore a possible link. Practical Diabetes, 32(3), 111–113.
Swaminathan, K. , Veerasekar, G. , Kuppusamy, S. , et al. (2017, Jan–Feb). Noncommunicable disease in rural India: Are we seriously underestimating the risk? The Nallampatti noncommunicable disease study. Indian J Endocrinol Metab, 21(1), 90–95.
Tabák, A. G. , Herder, C. , Rathmann, W. , Brunner, E. J. , & Kivimäki, M. (2012). Prediabetes: A high-risk state for diabetes development. Lancet, 379, 2279–2290.
Unnikrishnan, A. G. (2019, Sep–Oct). Artificial intelligence in health care: Focus on diabetes management. Indian J Endocrinol Metab, 23(5), 503–506.
Velmurugan, G. , Swaminathan, K. , Sundaresan, M. et al . (2020). Association of co-accumulation of arsenic and organophosphate insecticides with diabetes and atherosclerosis in a rural agricultural community: KMCH-NNCD-I study. Acta Diabetol, 57(10), 1159–1168.
WHO. World Health Organization . (2016). Global report on diabetes. https://apps.who.int/iris/bitstream/handle/10665/204871/9789241565257_eng.pdf ISBN 978 92 4 156525 7 France

Role of Big Data Analytics in Industrial Revolution 4.0

Analytics Magazine . (2012). How Big Data is Changing the oil & gas industry – Analytics Magazine.analytics-magazine.org. Retrieved April 11, 2020, from http://analytics-magazine.org/how-big-data-is-changing-the-oil-a-gas-industry/

Ayers, R. , Panova, E. , Deen, K. , Panova, E. , Deen, K. , Panova, E. , Deen, K. , & Panova, E. (2018). How Big Data Is Revolutionizing Sports - Dataconomy. Dataconomy. Retrieved April 12, 2020, from https://dataconomy.com/2018/01/big-data-revolutionizing-favorite-sports-teams/

Berger-De Leon, M. , Reinbacher, T. , & Wee, D. (2018). The IoT as a growth driver. McKinsey & Company. Retrieved April 04, 2020, from https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/the-iot-as-a-growth-driver

Brennan, E. (2019). A new global agriculture: Using Big Data to bring farmers together. foodtank.com. Retrieved April 11, 2020, from https://foodtank.com/news/2019/04/a-new-global-agriculture-using-big-data-to-bring-farmerstogether/?gclid=CjwKCAjwvtX0BRAFEiwAGWJyZGgIcE6i5Hpv MT-lI5uv7AoH7QqBaAXzEhE8WcYOhfstLqICiyfexoC3UwQAvD_BwE

Cattell, J. , Chilukuri, S. , & Levy, M. (2013). How Big Data can revolutionize pharmaceutical R&D. McKinsey & Company. Retrieved April 20, 2020, from https://www.mckinsey.com/industries/pharmaceuticals-and-medical-products/our-insights/how-big-data-can-revolutionize-pharmaceutical-r-and-d

Imanuel . (2014). Pat research: B2B reviews, buying guides & best practices. What is predictive analytics?. Retrieved April 05, 2020, from https://www.predictiveanalyticstoday.com/what-is-predictive-analytics/

SAS . (2020). Big Data in real life: The impact of analytics on car manufacturing. sas.com. Retrieved April 13, 2020, from https://www.sas.com/en_au/insights/articles/analytics/impact-of-analytics-on-car-manufacturing.html

Team Machine Learning. (2017). 8 different job roles in data science Big Data industry. HackerEarth Blog. Retrieved April 09, 2020, from https://www.hackerearth.com/blog/developers/8-different-job-roles-data-science-big-data-industry/

UBS . (2015). The evolution of artificial intelligence. AI's Coming Of Age. ubs.com. Retrieved April 10, 2020, from https://www.ubs.com/microsites/artificial-intelligence/en/ai-coming-age.html

Wire, B. (2014). Research and markets: Big Data in extraction and natural resource industries report: Mining, water, timber, oil and gas markets 2014–2019. *Business Wire* . Retrieved April 08, 2020, from https://www.businesswire.com/news/home/20140707005873/en/Research-Markets-Big-Data-Extraction-Natural-Resource

Woodie, A. (2015). 9 must-have skills to land top Big Data jobs in 2015. Datanami. Retrieved April 08, 2020, from https://www.datanami.com/2015/01/07/9-must-skills-land-top-big-data-jobs 2015/

Big Data Infrastructure and Analytics for Education 4.0

Buyya, R. , Yeo, C. S. , Venugopal, S. , Broberg, J. , & Brandic, I. (2009). Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. Future Generation Computer Systems, 25(6), 599–616.

Dean, J. , & Ghemawat, S. (2004). Mapreduce: Simplified data processing on large clusters. 6th ACM Conference on Symposium on Operating Systems Design Implementation.

Díaz, M. , Martín, C. , & Rubio, B. (2016). State-of-the-art, challenges, and open issues in the integration of Internet of things and cloud computing. Journal of Network and Computer Applications, 67, 99–117.

Paul, A. , & Jeyaraj, R. (2019). Internet of Things: A primer. Human Behavior and Emerging Technologies, 1(1), 37–47.

Yang, X. , et al. (2014). Cloud computing in e-Science: Research challenges and opportunities. The Journal of Supercomputing, 70(1), 408–464.

Text Analytics in Big Data Environments

Abraham, A. , Das, S. , & Konar, A. (2006, July). Document clustering using differential evolution. In 2006 IEEE International Conference on Evolutionary Computation (1784–1791). IEEE.
Agarwal, B. , & Mittal, N. (2014). Text classification using machine learning methods-a survey. In Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28–30, 2012 (701–709). Springer, New Delhi.
Anandarajan, M. , & Nolan, T. (2019). Practical text analytics. Maximizing the value of text data. In Advances in Analytics and Data Science, Vol. 2, p. 285. New York: Springer.
Bifet, A. (2013). Mining Big Data in real time. Informatica, 37, 15–20.
Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77–84.
Blei, D. M. , & Lafferty, J. D. (2006). Dynamic topic models. In Proceedings of the 23rd International Conference on Machine Learning (113–120). ACM.
Blei, D. M. , & Lafferty, J. D. (2007). A correlated topic model of science. The Annals of Applied Statistics, 1(1), 17–35.
Blei, D. M. , Ng, A. Y. , & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022.
Cloud Security Alliance . (2013). Big Data analytics for security intelligence. Retrieved from www.cloudsecurityalliance.org/research/big-data62
Debortoli, S. , Müller, O. , Junglas, I. , & vomBrocke, J. (2016). Text mining for information systems researchers: An annotated topic modeling tutorial. Communications of the Association for Information Systems, 39(1), 7.
Edwards, R. , & Fenwick, T. (2016). Digital analytics in professional work and learning. Studies in Continuing Education, 38(2), 213–227.
Feldman, R. , & Sanger, J. (2007). The text mining handbook: Advanced approaches in analyzingunstructured data. Cambridge: Cambridge University Press.
Gaikwad, D. K. , & Mahender, C. N. (2016). A review paper on text summarization. International Journal of Advanced Research in Computer and Communication Engineering, 5(3), 154–160.
Granello, D. H. , & Wheaton, J. E. (2004). Online data collection: Strategies for research. Journal of Counseling & Development, 82(4), 387–393.
Hadoop . (2015). Apache Software Foundation (ASF). Retrieved December 18, 2015, from http://hadoop.apache.org
Ikonomakis, M. , Kotsiantis, S. , & Tampakas, V. (2005). Text classification using machine learning techniques. WSEAS Transactions on Computers, 4(8), 966–974.
Isa, D. , Lee, L. H. , Kallimani, V. P. , & Rajkumar, R. (2008). Text document pre-processing with the Bayes formula for classification using the support vector machine. IEEE Transactions on Knowledge and Data Engineering, 20(9), 1264–1272.
Joulin, A. , Grave, E. , Bojanowski, P. , & Mikolov, T. (2016). Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759.
Kabir, S. M. S. (2016). Basic guidelines for research. Chittagong: Book Zone Publication. Chapter 9.
Keim, D. , Qu, H. , & Ma, K. L. (2013). Big-data visualization. IEEE Computer Graphics and Applications, 33(4), 20–21.
Kim, G. H. , Trimi, S. , & Chung, J. H. (2014). Big-data applications in the government sector. Communications of the ACM, 57(3), 78–85.
Krippendorff, K. (2012). Content analysis: An introduction to its methodology. Thousand Oaks: Sage.
Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 5(1), 1–167.
Lovins, J. B. (1968). Development of a stemming algorithm. Mechanical Translation and Computational Linguistics, 11(1–2), 22–31.

Manning, C. D. , Schütze, H. , & Raghavan, P. (2008). Introduction to information retrieval. Cambridge: Cambridge University Press.
Mirończuk, M. M. , & Protasiewicz, J. (2018). A recent overview of the state-of-the-art elements of text classification. Expert Systems with Applications, 106, 36–54.
Mohammed, A. J. , Yusof, Y. , & Husni, H. (2015). Document clustering based on firefly algorithm. Journal of Computer Science, 11(3), 453–465.
Nasukawa, T. , & Yi, J. (2003). Sentiment analysis: Capturing favorability using naturallanguage processing. In Proceedings of the 2nd International Conference on Knowledge Capture (70–77). ACM.
Nigam, K. , McCallum, A. K. , Thrun, S. , & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. Machine Learning, 39(2–3), 103–134.
PAT Research . Retrieved from https://www.predictiveanalyticstoday.com/top-software-for-text-analysis-text-mining-text-analytics/
Pipino, L. L. , Lee, Y. W. , & Wang, R. Y. (2002). Data quality assessment. Communications of the ACM, 45(4), 211–218.
Porter, M. F. (2001). Snowball: A language for stemming algorithms. Retrieved from http://snowball.tartarus.org/texts/introduction.html
Rahm, E. , & Do, H. H. (2000). Data cleaning: Problems and current approaches. IEEE Data Engineering Bulletin, 23(4), 3–13.
Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys (CSUR), 34(1), 1–47.
Silva, C. , & Ribeiro, B. (2003, July). The importance of stop word removal on recall values in text categorization. In Proceedings of the International Joint Conference on Neural Networks, 2003 (Vol. 3, 1661–1666). IEEE.
Talib, R. , Hanif, M. K. , Ayesha, S. , & Fatima, F. (2016). Text mining: Techniques, applications and issues. International Journal of Advanced Computer Science & Applications, 1(7), 414–418.
Tas, O. , & Kiyani, F. (2007). A survey automatic text summarization. Press Academia Procedia, 5(1), 205–213.
Wang, L. , Wang, G. , & Alexander, C. A. (2015). Big data and visualization: Methods, challenges and technology progress. Digital Technologies, 1(1), 33–38.
Xu, R. , & Wunsch, D. (2005). Survey of clustering algorithms. IEEE Transactions on Neural Networks, 16(3), 645–678.
Zhang, W. , Yoshida, T. , Tang, X. , & Wang, Q. (2010). Text clustering using frequent itemsets. Knowledge-Based Systems, 23(5), 379–388.

Business Data Analytics: Applications and Research Trends

Apache Avro . Retrieved April 12, 2020, from https://avro.apache.org/
Apache Cassandra . Retrieved April 12, 2020, from http://cassandra.apache.org/
Apache Chukwa . Retrieved April 12, 2020, from https://chukwa.apache.org/
Apache Flink . Retrieved April 12, 2020, from https://flink.apache.org/
Apache Flume . Retrieved April 12, 2020, from https://flume.apache.org/
Apache HBase . Retrieved April 12, 2020, from http://hbase.apache.org/
Apache Hive . Retrieved April 12, 2020, from http://hive.apache.org/
Apache Oozie Workflow Scheduler for Hadoop . Retrieved April 12, 2020, from http://oozie.apache.org/
Apache Pig . Retrieved April 12, 2020, from http://pig.apache.org/
Apache Spark . Retrieved April 12, 2020, from https://spark.apache.org/
Apache Sqoop . Retrieved April 12, 2020, from http://sqoop.apache.org/
Apache Storm . Retrieved April 12, 2020, from https://storm.apache.org/
Apache Tez . Retrieved April 12, 2020, from http://tez.apache.org/
Apache Zookeeper . Retrieved April 12, 2020, from https://zookeeper.apache.org/
Bhadani, A. K. , & Jothimani, D. (2016). Big data: Challenges, opportunities, and realities. Effective big data management and opportunities for implementation (pp. 1–24). Hershey, PA: IGI Global.

Bloomberg Businessweek Research Services, The current state of business analytics: Where do we go from here? Retrieved April 12, 2020, from https://www.sas.com sources/asset/busanalyticssstudy_wp_08232011.pdf

Chahal, H. , Jyoti, J. , & Wirtz, J. (2019). Business analytics: Concept and applications. In Understanding the role of business analytics (pp. 1–8). Singapore: Springer.

Chatzimilioudis, G. , et al. (2012). Crowdsourcing with smartphones. IEEE Internet Computing, 16(5), 36–44.

Davenport, T. (2014). Big data at work: Dispelling the myths, uncovering the opportunities. Boston, MA: Harvard Business Review Press.

Davenport, T. , & Harris, J. (2017). Competing on analytics: Updated, with a new introduction: The new science of winning. Boston, MA: Harvard Business Press.

Davenport, T. H. , Barth, P. , & Bean, R. (2012).How “big data” is different. MIT Sloan Management Review, 54(1), 43–46.

De Domenico, M. , Lima, A. , González, M. C. , & Arenas, A. (2015). Personalized routing for multitudes in smart cities. EPJ Data Science, 4(1), 1.

Dobre, C. , & Xhafa, F. (2014). Intelligent services for big data science. Future Generation Computer Systems, 37, 267–281.

Dong, H. , et al. (2015). Traffic zone division based on big data from mobile phone base stations. Transportation Research Part C: Emerging Technologies, 58, 278–291.

Douglass, R. W. , et al. (2015). High resolution population estimates from telecommunications data. EPJ Data Science, 4(1), 4.

Finger, F. , et al. (2016). Mobile phone data highlights the role of mass gatherings in the spreading of cholera outbreaks. Proceedings of the National Academy of Sciences, 113(23), 6421–6426.

Gokalp, M. , et al. (2016). Big data for industry 4.0: A conceptual framework. 2016 International Conference on Computational Science and Computational Intelligence (CSCI). IEEE.

Grover, V. , et al. (2018). Creating strategic business value from big data analytics: A research framework. Journal of Management Information Systems, 35(2), 388–423.

Hashem, Ibrahim AbakerTargio , et al. (2015). The rise of “big data” on cloud computing: Review and open research issues. Information Systems, 47, 98–115.

IDC . Big data big opportunities. Retrieved April 12, 2020, from http://www.emc.com/microsites/cio/articles/big-databig-opportunities/LCIA-Big Data Opportunities -Value.pdf

Jackson, T. W. , & Lockwood, S. (2018). Business analytics: A contemporary approach. London: Macmillan International Higher Education.

Khatib, E.J. , Barco, R. , Munoz, P. , La Bandera, I. D. , & Serrano, I. (2016). Self-healing in mobile networks with big data. IEEE Communications Magazine, 54(1), 114–120.

LaValle, S. , et al. (2010). Analytics: The new path to value. MIT Sloan Management Review, 52(1), 1–25.

LaValle, S. , et al. (2011). Big data, analytics and the path from insights to value. MIT Sloan Management review, 52(2), 21–32.

Lim, E.-P. , Chen, H. , & Chen, G. (2013). Business intelligence and analytics: Research directions. ACM Transactions on Management Information Systems (TMIS), 3(4), 1–10.

Lima, A. (2016). Digital traces of human mobility and interaction: Models and applications. Dissertation, University of Birmingham.

Lokanathan, S. , et al. (2016). The potential of mobile network big data as a tool in Colombo's transportation and urban planning. Information Technologies & International Development, 12(2), 63.

Mahout . Retrieved April 12, 2020, from http://mahout.apache.org/

MLLib . Retrieved April 12, 2020, from https://spark.apache.org/mllib/

Mohammadi, M. , et al. (2018). Deep learning for IoT big data and streaming analytics: A survey. IEEE Communications Surveys & Tutorials, 20(4), 2923–2960.

Nweke, H. F. , et al. (2018). Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges. Expert Systems with Applications, 105, 233–261.

Nweke, H. F. , et al. (2019). Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions. Information Fusion, 46, 147–170.

Ochoa, S. F. , Fortino, G. , & Di Fatta, G. (2017). Cyber-physical systems, internet of things and big data. Future Generation Computer Systems, 75, 82–84.
Python Programming . Retrieved April 12, 2020, from https://www.python.org/
Salehan, M. , & Kim, D. J. (2016). Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics. Decision Support Systems, 81, 30–40.
Stubbs, E. (2011). The value of business analytics. Hoboken, NJ: John Wiley & Sons.
Teaching tools (edudemic) . (2012). 10 incredible powerful tools of the future. Retrieved from https://educationprospector.wordpress.com/2012/08/18/teaching-toolsedudemic.
The R Project for Statistical Computing . Retrieved April 12, 2020, from http://www.r-project.org/
Tom, W. (2012). Mobile big data fault-tolerant processing for ehealth network. IEEE Network, 30(1), 36–42. Hadoop: The definitive guide (p. 36). Newton, MA: O'Reilly Media, Inc.
Wang, K. , Shao, Y. , Shu, L. , Zhu, C. , & Zhang, Y. (2016). Mobile big data fault-tolerant processing for ehealth networks. IEEE Network, 30(1), 36–42.
Watson, H.J. (2009). Tutorial: Business intelligence – Past, present, and future. Communications of the Association for Information Systems, 25(1), 39.
Wyber, R. , Vaillancourt, S. , Perry, W. , Mannava, P. , Folaranmi, T. , & Celi, L.A. (2015). Big data in global health: Improving health in low-and middle-income countries. Bulletin of the World Health Organization, 93, 203–208.
Xu, Z. , et al. (2016). Crowdsourcing based description of urban emergency events using social media big data. IEEE Transactions on Cloud Computing, 8(2), 387–397.
Yang, T.-Y. , et al. (2017). Behavior-based grade prediction for MOOCs via time series neural networks. IEEE Journal of Selected Topics in Signal Processing, 11(5), 716–728.
Zhan, X. , Ukkusuri, S. V. , & Zhu, F. (2014). Inferring urban land use using large-scale social media check-in data. Networks and Spatial Economics, 14(3–4), 647–667.

Role of Big Data Analytics in the Financial Service Sector

Hariharasudan, A. , Kot, S. (2018). A scoping review on digital English and Education 4.0 for Industry 4.0. Social Sciences, 7, 227.
Hussin, A. A. (2018, July). Education 4.0 made simple: Ideas for teaching. International Journal of Education and Literacy Studies, 6(3), 92.
Lawrence R. , Ching L. F. , & Abdullah H. Strengths and Weaknesses of Education 4.0 in the Higher Education Institution. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 9(2S3). Retrieved from https://prezi.com/i/f_6nqbbm11gw/strengths-and-weaknesses-of-education-40-in-the-higher-education-institution/
Linh, P. K. (2019, June). Education in Industry 4.0. International Journal of Engineering Science Invention (IJESI), 8(06), Series. I, 09–13.
Lopez-Garcia, T. J. , et al. (2019). Review of trends in the educational model of distance education in Mexico, towards an Education 4.0. Computer Reviews Journal, 3. ISSN: 2581-6640.
Morabito, V. (2015). Big data and analytics: Strategic and organizational impacts. New York: Springer
Mourtzisa, D. , Vlachoua, E. , Dimitrakopoulosa, G. , & Zogopoulos, V. (2018). Cyber-physical systems and education 4.0 – The teaching factory 4.0 concept. Procedia Manufacturing, 23, 129–134.
O’Riáin, S. , Curry, E. , & Harth, A. (2012). XBRL and open data for global financial ecosystems: A linked data approach. International Journal of Accounting Information Systems, 13, 141–162.
Pence, H. E. (2014). What is Big Data and why is it important? Journal of Educational Technology Systems, 43(2), 159–171.
Shahroom, A. A. , & Hussin, N. (2018). Industrial revolution 4.0 and education. International Journal of Academic Research in Business and Social Sciences, 8(9), 314–319.
Sharma, P. (2019, December). Digital revolution of Education 4.0. International Journal of Engineering and Advanced Technology (IJEAT), 9(2), 314–319.
Tandon R. , & Tandon S. (2020, February). Education 4.0: A new paradigm in transforming the future of education in India. International Journal of Innovative Science, Engineering &

Technology, 7(2). ISSN (Online): 2348–7968.
Technavio . (2013). Global Big Data market in the financial services sector 2012–2016. Retrieved from https://inkwoodresearch.com/reports/big-data-market/
Tsai, B.-H. (2014). Examination of ex-dividend day trading using Big Data of American depositary receipts. *Proc. 2nd Int’l Conf. Advanced Cloud and Big Data (CBD)* , pp. 34–38.
Wang, Y. , Li, S. , & Lin, Z. (2013). Revealing key non-financial factors for online credit-scoring in e-financing. *Proc. 10th Int’l Conf. Service Systems and Service Management (ICSSSM)* , pp. 547–552.

Role of Big Data Analytics in the Education Domain

Admiraal, W. , Post, L. , Guo, P. , Saab, N. , Makinen, S. , Rainio, O. , … Danford, G. (2019). Students as future workers: Cross-border multidisciplinary learning labs in higher education. International Journal of Technology in Education and Science, 3(2), 85–94. Retrieved from https://www.learntechlib.org/p/207262/
Alexander, B. , Ashford-Rowe, K. , Barajas-Murph, N. , Dobbin, G. , Knott, J. , McCormack , … Weber, N. (2019). Horizon report 2019 higher education edition. EDU19. Retrieved from https://www.learntechlib.org/p/208644/
Altbach, P. G. , Reisberg, L. , & Rumbley, L. E. (2009). Trends in global higher education: Tracking an academic revolution in global perspectives on Higher Education. UNESCO Report, from the World Conference on Higher Education. Retrieved from http://www.cep.edu.rs/public/Altbach,_Reisberg,_Rumbley_Tracking_an_Academic_Revolution,_UNESCO_2009.pdf
Ameen, N. (2019). What robots and AI may mean for university lecturers and students? Retrieved from http://theconversation.com/what-robots-and-ai-may-mean-for-university-lecturers-and-students-114383
Anirban, S. (2014). Big data analytics in the education sector: needs, opportunities, and challenges. International Journal of Research in Computer and Communication Technology (IJRCCT), 3(11), 2278–5841.
Bakhshi, H. , Downing, J. , Osborne, M. , & Schneider, P. (2017). The future of skills: employment in 2030. London: Pearson & Nesta. Retrieved from https://media.nesta.org.uk/documents/the_future_of_skills_employment_in_2030_0.pdf
Barrow, J. , Forker, C. , Sands, A. , O'Hare, D. , & Hurst, W. (2019). Augmented reality for enhancing life science education. Paper presented at VISUAL 2019 – The Fourth International Conference on Applications and Systems ofVisual Paradigms, Rome, Italy.
Bates, G. , Rixon, A. , Carbone, A. , & Pilgrim, C. (2019). Beyond employability skills: Developing professional purpose. Journal of Teaching and Learning for Graduate Employability, 10(1). Retrieved from https://ojs.deakin.edu.au/index.php/jtlge/article/view/794
Berners-Lee, T. , Hendler, J. , & Lassila, O. (2001). The semantic web. Scientific American, 284(5), 34–43.
Boulton, G. (2017). The digital revolution and the future of science. Retrieved from https://www.timeshighereducation.com/hub/p/jisc-futures-digital-revolution-and-future-science
Cheng, L. , Ritzhaupt, A. D. , & Antonenko, P. (2018). Effects of the flipped classroom instructional strategy on students’ learning outcomes: A meta-analysis. Education Technology Research and Development, 67(4). Retrieved from 10.1007/s11423-018-9633-7
Ciechanowski, L. , Przegalinska, A. , Magnuski, M. , & Gloor, P. (2019). In the shades of the uncanny valley: An experimental study of human–chatbot interaction. Future Generation Computer Systems, 92, 539–548. Retrieved from 10.1016/j.future.2018.01.055
Connor, A. , Karmokar, S. , & Whittington, C. (2015). From STEM to STEAM: Strategies for enhancing engineering technology education. International Journal of Engineering Pedagogy, 5(2), 37–47. Retrieved from https://online-journals.org/index.php/i-jep/article/view/4458
Daniel, J. S. (2018). Open Universities: Old concepts and contemporary challenges. IRRODL Special Issue on theFuture of Open Universities. Retrieved from http://sirjohn.ca/wpontent/uploads/2018/11/20180718_IRRODL_RevNov.pdf
Davies, S. (2019). Moving towards Education 4.0. Blog. Retrieved from https://www.jisc.ac.uk/blog/member-stories-towards-higher-education-40-15-jan-2019

Dobozy, E. , & Cameron, L. (2018). Editorial: Special issue on learning design research: Mapping the terrain. Australasian Journal of Educational Technology, 34(2), i–v. Retrieved from https://ajet.org.au/index.php/AJET/article/view/4390
Ehlers, U. D. , & Kellermann, S. A. (2019). Future skills – The future of learning and higher education. Results of the International Future Skills Delphi Survey. Karlsruhe. Retrieved from https://nextskills.files.wordpress.com/2019/03/2019-02-23-delphi-report-final.pdf
Feldman, P. (2018). The potential of 4.0 is huge – UK must take the lead. Blog. Retrieved from https://www.jisc.ac.uk/blog/the-potential-of-education-4-is-huge-the-uk-must-take-the-lead-now-12-sep-2018
Fisk, P. (2017). Education 4.0 … the future of learning will be dramatically different, in school and throughout life. Retrieved from https://www.thegeniusworks.com/2017/01/future-education-young-everyone-taught-together (Also see Video 'The Future of Learning' on this site.)
Gallagher, M. (2019). Learning for the future of work. Retrieved from https://www.swinburne.edu.au/new-workforce (See Website and video.)
Golembiewski, L. (2019). How wearable AI will amplify human intelligence. Harvard Business Review. Retrieved from https://hbr.org/2019/04/how-wearable-ai-will-amplify-human-intelligence
Heaven, D. (Ed). (2017). Machines that think. Boston, MA: Nicholas Barley Publishing.
Hussin, A. (2018). Education 4.0 Made simple: Ideas for teaching. International Journal of Education and LiteracyStudies, 6(3), 92–98. Retrieved from http://www.journals.aiac.org.au/index.php/IJELS/article/view/4616
Johansson, F. (2017). The Medici Effect. Boston, MA: Harvard Business Review Press.
John, J. (2019). How to create: A broader, fairer and smarter education system. Retrieved from https://www.jisc.ac.uk/blog/how-to-create-a-broader-fairer-and-smarter-education-system-08-mar-2019
JISC. (2018). Digital skills crucial to the success of fourth industrial revolution. Retrieved from https://www.jisc.ac.uk/news/digital-skills-crucial-to-the-success-of-fourth-industrial-revolution-28-jun-2018
JISC. (2019). Preparing for Education 4.0. Retrieved from https://www.timeshighereducation.com/hub/jisc/p/preparing-education-40
Jongbloed, B. (2015). Universities as hybrid organizations: Trends, drivers, and challenges for the European university. International Studies of Management & Organization, 45(3), 207–225.
Luckin, R. (2019). AI and education: The reality and the potential. The Knowledge Illusion. Retrieved from https://knowledgeillusion.blog/2019/04/30/ai-and-education-the-reality-and-the-potential/
McGregor, A. , & Hamilton, M. (2019). Shaping education for a hyper connected world. Digifest Magazine.
McVitty, D. (2019). The more universities are thinking about value for money, the better the sector looks. WONKHEblog. Retrieved from https://tinyurl.com/y5bxhwmz.
Navitas Ventures . (2017). Digital transformation in higher education, report. Retrieved from http://www.navitasventures.com/wp-content/uploads/2017/08/HE-Digital-Transformation-_Navitas_Ventures_-EN.pdf
OECD . (2018). A brave new world: Technology & education. Trends Shaping Education. Retrieved from https://www.oecd.org/education/ceri/Spotlight-15-A-Brave-New-World-Technology-and-Education.pdf
Ruiz-Palmero, J. , Colomo-Magaña, E. , Ríos-Ariza, J. M. , & Gómez-García, M. (2020). Big Data in education: Perception of training advisors on its use in the educational system. Social Sciences, 9(4), 53.
Salmon, G. , & Asgari, S. (2019). Higher education – The last bastion? European Journal of Open, Distance and E-learning. Retrieved from http://www.eurodl.org/?p=current&sp=brief&article=792
Samans, R. (2019). Globalization 4.0 shaping a new global architecture in the age of the Fourth Industrial Revolution: A call for engagement. World Economic Forum Report. Retrieved from http://www3.weforum.org/docs/WEF_Globalization_4.0_Call_for_Engagement.pdf
Schwab, K. (2016). The Fourth Industrial Revolution: What it means, how to respond. Retrieved from https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/
Vlachopoulos, P. (2018). Curriculum digital transformation through learning design: The design, develop, implement methodology. In K. Ntalianis , A. Andreatos , & C. Sgouropoulou (Eds.),

Proceedings of the 17th European Conference on e-Learning (pp. 585–591). Reading, UK: Academic Conferences and Publishing International.
Xing, B. (2019). Towards a Magic Cube Framework in understanding Higher Education 4.0 for the Fourth Industrial Revolution. In D. B. A. Khosrow-Pou (Ed.), Handbook of research on challenges and opportunities in launching a technology driven international university (pp. 107–130). Hershey: IGI Global.

Social Media Analytics

Antonakaki, D. , Spiliotopoulos, D. , Samaras, C. V. , Pratikakis, P. , Ioannidis, S. , & Fragopoulou, P. (2017, October 31). Social media analysis during political turbulence. PloS One, 12(10), e0186836.
Bao, Y. , & Dawood, H. (2019). Towards deep learning prospects: Insights for social media analytics. Retrieved from http://www.ieee.org/publications_standards/publications/rights/index.html
Batrinca, B. , & Treleaven, P. C. (2014). Social media analytics: A survey of techniques. Tools and platforms, Published online: 26 July. This article is published with open access at Springerlink.com
Bengio, Y. (2009). Learning deep architectures for AI. Retrieved from http://www.iro.umontreal.ca/bengioy
Bright, J. , & Margetts, H. (2014). Scott Hale and TahaYasseri. The Use of Social Media for Research and Analysis: A Feasibility Study DWP ad hoc research report no. 13, Oxford Internet Institute on behalf of the Department for Work and Pensions. ISBN 978-1-78425-407-0.
Chae, J. , Thomy, D. , Boschy, H. , Sejong, Y. J. , Maciejewsk, R. , Ebert, D. S. , & Ertl, T. (2012). Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. Visual Analytics Science and Technology (VAST), 2012 IEEE Conference.
Cioffi-Revilla, C. (2010). Computational social science. WIREs Computational Statistics. doi:10.1002/wics.95
Feldman, R. (2013). Techniques and applications for sentiment analysis. Communications of the ACM, 56(4), 82–89.
Gastelum, Z. N. , & Whattam, K. M. (2013). State-of-the arts of social media analytics research Pacific Northwest National Laboratory Richland. Washington 99352. PNNL-22171.
Jindal, N. , & Liu, B. (2008). Opinion spam and analysis. In Proceedings of the 2008 International Conference on, Web Search and Data Mining, WSDM '08, 219–230. ACM, New York, NY.
Kalia, G. (2013). A research paper on social media: An innovative educational tool. Punjab: Chitkara University.
Kavitha, D. (2017). Survey of data mining techniques for social networking websites. International Journal of Computer Science and Mobile Computing, 6(4), 418–426.
Khan, A. , Baharudin, B. , Lee, L. H. , & Khan, K. (2010). A review of machine learning algorithms for text-documents classification. Journal of Advances in Information Technology, 1(1). Retrieved from http://www.jait.us/uploadfile/2014/1223/20141223050800532.pdf
Kularathne, S. D. , Dissanayake, R.B. , Samarasinghe, N.D. , Premalal, L.P.G. , & Premaratne, S. C. (2017). Customer behavior analysis for social media. International Journal of Advanced Engineering. Management and Science (IJAEMS), 3(1). ISSN: 2454–1311.
Kumar, A. , Khorwal, R. , & Chaudhary, S. (2016). A survey on sentiment analysis using swarm intelligence. Indian Journal of Science and Technology, 9(39), 1–7.
Mihanovic, A. , Gabelica, H. , & Krstic, Z. (2012). Big Data and sentiment analysis using Knime: Online reviews vs. social media. Croatia: MIPRO Opatija.
Mukherjee, A. , Liu, B. , & Glance, N. (2012). Spotting fake reviewer groups in consumer reviews. In Proceedings of the 21st International Conference on World Wide Web, WWW '12, 191–200. ACM, New York, NY.
Patnaik, S. , & Sucharita Barik, S. (2018). Social media analytics using visualization. International Journal of Scientific Research Engineering & Technology (IJSRET), 7(4). ISSN 2278–0882.

Praveena, T. L. , & Lakshmi, N. V. Muthu . (2017). An overview of social media analytics. International Journal of Advanced Scientific Technologies. Engineering and Management Sciences, 3(1). IJASTEMS-ISSN: 2454-356X.

Umar, R. (2014). Social media analytics as a business intelligence practice: Current landscape & future prospects. International Conference on Parallel Processing (ICPP 2014).

Weiguo Fan Michael D.Gordon . (2014). Unveiling the power of social media analytics. Retrieved from

https://www.researchgate.net/publication/259148570_The_Power_of_Social_Media_Analytics

https://aurus5.com/blog/cisco/cisco-logo-history-and-evolution/

https://barnraisersllc.com/2015/11/23/7-case-studies-show-social-media-analytics-pay-off/

https://biznology.com/2016/08/12-inspiring-social-media-monitoring-case-studies/

https://en.wikipedia.org/wiki/File:Samsung_Logo.svg

https://internetretailing.net/retail-directory-listing/apparel-and-footwear/lipsy

https://monkeylearn.com/blog/text-analysis-tools/

https://www.businessinsider.com/pizza-hut-brings-back-old-logo-2019-6

https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.freelogovectors.net%2Fkeen-logo-eps-file%2F&psig=AOvVaw1SkarYBiiiwHoki8-0j7v4&ust=1587628689027000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCKDY-PrH--gCFQAAAAAdAAAAABAM

https://www.google.com/url?sa=i&url=https%3A%2F%2Fen.wikipedia.org%2Fwiki%2FKmart&psig=AOvVaw3dy7zw97LS4EiRnTd6aKfs&ust=1587628790156000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCJCVkanI--gCFQAAAAAdAAAAABAD

https://www.google.com/url?sa=i&url=https%3A%2F%2Fen.wikipedia.org%2Fwiki%2FTV_Land&psig=AOvVaw0WDWd41Nk-1XS10E4AW0yG&ust=1587628808483000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCOiw57nI--gCFQAAAAAdAAAAABAD

https://www.google.com/url?sa=i&url=https%3A%2F%2Fonjalazbroz.blogspot.com%2F2018%2F11%2Fmulticulturalism-necessity.html&psig=AOvVaw3Ij4r-4LxXCAkMmwnjqrQK&ust=1587999174434000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCKCK_e-shukCFQAAAAAdAAAAABAD

https://www.google.com/url?sa=i&url=https%3A%2F%2Fseeklogo.com%2Fvector-logo%2F205129%2Fgatorade&psig=AOvVaw350C6Xi2l-6s0AS3yrKiiX&ust=1587628782314000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCLjguNPI--gCFQAAAAAdAAAAABAH

https://www.google.com/url?sa=i&url=https%3A%2F%2Fblogs.helsinki.fi%2Fquantitative-communication%2Fmethods%2Fcontent-analysis%2F&psig=AOvVaw0ClNC5ASe4o4wrn4GxPST-&ust=1587999275197000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCMDC28CshukCFQAAAAAdAAAAABAM

https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.fte.org%2Fsite-spotlight-yale-university%2Fyale-logo%2F&psig=AOvVaw19DjEpMeTZByocZtsYVGZV&ust=1587628774468000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCPjmuqHI--gCFQAAAAAdAAAAABAD

https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.pinterest.com%2Fpin%2F452963674998706292%2F&psig=AOvVaw2vezEFuc6dUujMj8YejGzI&ust=1587628798183000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCPCC0a7I--gCFQAAAAAdAAAAABAD

https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.premierleague.com%2Fpartners%2Fbarclays&psig=AOvVaw15ndhLVdUXJ1ncuZ7DmVWp&ust=1587628628488000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCNiV2-TH--gCFQAAAAAdAAAAABAD

https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.prnewswire.com%2Fnews-releases%2Fmoneygram-launches-new-online-money-transfer-service-platform-with-walmart-300322812.html&psig=AOvVaw06_Nkd8OzRy5X4oHz0RAzl&ust=1587628793952000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCNiWza7I--gCFQAAAAAdAAAAABAJ

https://www.predictiveanalyticstoday.com/top-free-software-for-text-analysis-text-mining-text-analytics/

https://www.predictiveanalyticstoday.com/top-software-for-text-analysis-text-mining-text-analytics/

https://www.socialmediatoday.com/social-business/12-best-social-media-monitoring-tools-consider
https://www.softwaretestinghelp.com/data-visualization-tools/
https://www.toptal.com/designers/data-visualization/data-visualization-tools

Robust Statistics: Methods and Applications

Campbell, N. , Lopuhaa, H. P. , & Rousseeuw, P. J. (1998). On the calculation of a robust S-estimator of a covariance matrix. Statistics in Medicine, 17(23), 2685–2695.
Cook, R. D. , Hawkins, D. M. , & Weisberg, S. (1993). Exact iterative computation of the robust multivariate minimum volume ellipsoid estimator. Statistics and Probability Letters, 16, 213–218.
Chork, C. , & Rousseeuw, P. J. (1992). Integrating a high-breakdown option into discriminant analysis in exploration geochemistry. Journal of Geochemical Exploration, 43(3), 191–203.
Crawley, M. J. (2007): The R book. New York: John Wiley and Sons, Limited.
Croux, C. , & Dehon, C. (2001). Robust linear discriminant analysis using s-estimators. Canadian Journal of Statistics, 29, 473–493.
Croux, C. , Filzmoser, P. , & Oliveira, M. R. (2007). Algorithms for projection-pursuit robust principal component analysis. Chemometrics and Intelligent Laboratory Systems, 87, 218–225.
Cuesta-Albertos, J. , Gordaliza, A. , & Matran, C. (1997). Trimmed k-means: An attempt to robustify quantizers. The Annals of Statistics, 25(2), 553–576.
Davies, P. L. (1987). Asymptotic behavior of S-estimates of multivariate location parameters and dispersion matrices. The Annals of Statistics, 15(3), 1269–1292.
Donoho, D. L. (1982). Breakdown properties of multivariate location estimators. Ph.D. qualifying paper. Cambridge: Harvard University.
Filzmoser, P. , Hron, K. , & Templ, M. (2012). Discriminant analysis for compositional data and robust parameter estimation. Computational Statistics, 27(4), 585–604.
Filzmoser, P. , & Todorov, V. (2013). Robust tools for the imperfect world. Information Sciences, 245, 4–20.
Fischler, M. A. , & Bolles, R. C. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography, commun. ACM, 24(6), 381–395.
Gallegos, M. T. , and Ritter, G. (2005). A robust method for cluster analysis. The Annals of Statistics, 33(1), 347–380.
Garcia-Escudero, L. A. , & Gordaliza, A. (1999). Robustness properties of k-means and trimmed k-means. Journal of Americal Statistical Association, 94, 956–969.
Garcia-Escudero, L. A. , & Gordaliza, A. (2005). A proposal for robust curve clustering. Journal of Classification, 22, 185–201.
Garcia-Escudero, L. A. , Gordaliza, A. , Matran, C. , & Mayo-Iscar, A. (2008). A general trimming approach to robust cluster analysis. Annals of Statistics, 36, 1324–1345.
Garcia-Escudero, L. A. Gordaliza, A. , Matran, C. , & Mayo-Iscar, A. (2010). A review of robust clustering methods. Advances in Data Analysis and Classification, 4(2–3), 89–109.
Hakim, A. E. , & Saban, M. (2012). FRPCA: Fast robust principal component analysis for online observations. 21st International Conference on Pattern Recognition, 413–416.
Hampel, F. R. (1968). Contributions to the theory of robust estimation, Ph.D. Thesis, University of California, Berkeley.
Hampel, F. R. (1974). The influence curve and its role in robust estimation. Journal of the American Statistical Association, 69, 383–393.
Hampel, F. R. , Ronchetti, E. M. , Rousseeuw, P. J. , & Stahel, W. A. (1986). Robust statistics: The approach based on influence functions. New York: Wiley.
Hawkins D. M. (1993). The feasible set algorithm for least median of squares regression. Computational Statistics and Data Analysis, 16, 81–101.
Hawkins, D. M. (1994a). The feasible solution algorithm for least trimmed squares regression. Computational Statistics and Data Analysis, 17, 185–196.
Hawkins D. M. (1994b). The feasible solution algorithm for the minimum covariance determinant estimator in multivariate data. Computational Statistics and Data Analysis, 17, 197–210.

Hawkins, D. M. , & McLachen, G. (1997). High-breakdown linear discriminant analysis. Journal of the American Statistical Association, 92, 136–143.
Hawkins, D. M. , & Olive, D. J. (1999). Improved feasible solution algorithms for high breakdown estimation. Computational Statistics and Data Analysis, 30, 1–11.
He, X. , & Fung, W. K. (2000). High breakdown estimation for multiple populations with applications to discriminant analysis. Journal of Multivariate Analysis, 72, 151–162.
Huber, P. J. (1964). Robust estimation of location parameters. Annals of Mathematical Statistics, 35, 73–101.
Huber, P. J. (1981). Robust statistics. New York: Wiley.
Huber, P. J. , & Ronchetti, E. M. (2009). Robust statistics. New York: John Wiley & Sons.
Hubert, M. , & Driessen, V. K. (2004). Fast and robust discriminant analysis. Computational Statistics & Data Analysis, 45(2), 301–320.
Hubert, M. , Rousseeuw, P. J. , & Verdonck, T. (2012). A deterministic algorithm for robust location and scatter. Journal of Computational and Graphical Statistics, 21, 618–637.
James, G. , Witten, D. , Hastie, T. , & Tibshirani, R. (2013). An introduction to statistical learning with applications in R. New York: Springer.
Kent, J. T. , & Tyler, D. E. (1996). Constrained M-estimation for multivariate location and scatter. The Annals of Statistics, 24(3), 1346–1370.
Koshevoy, G. , & Mosler, K. (1997). Zonoid trimming for multivariate distributions. Annals of Statistics, 25, 49–69.
Lopuhaa, H. P. (1989). On the relation between S-estimators and m-estimators of multivariate location and covariance. The Annals of Statistics, 17(4), 1662–1683.
Lopuhaa, H. P. , & Rousseeuw, P. J. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. The Annals of Statistics, 19(1), 229–248.
Liu, R. Y. (1990). On a notion of data depth based on random simplices. Annals of Statistics, 18, 405–414.
Mahalanobis, P. C. (1936). On the generalized distance in statistics. Proceedings of National Institute of Sciences, 12, 49–55.
Mallows, C. L. (1975). On some topics in robustness, technical memorandum. Murray Hill, NJ: Bell Telephone Laboratories.
Maronna, R. A. (1976). Robust M-estimators of multivariate location and scatter. The Annals of Statistics, 4(1), 51–67.
Muthukrishnan, R. , Boobalan, E. D. , & Mathaiyan, R. (2014). MCD based principal component analysis in computer vision. International Journal of Computer Science and Information Technologies, 5(6), 8293–8296.
Muthukrishnan, R. , & Udaya Prakash, N. (2019). Performance of classification techniques along with support vector machines. International Journal of Innovative Technology and Exploring Engineering, 9(2), 4366–4369.
Oja, H. (1983). Descriptive statistics for multivariate distributions. Statistics and Probability Letter, 1, 327–332.
R Core Team . (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/
Rousseeuw, P. J. (1984). Least median of squares regression. Journal of the American Statistical Association, 79, 871–880.
Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. Mathematical Statistics and Applications, 8, 283–297.
Rousseeuw, P. J. , & Croux, C. (1993). Alternatives to the median absolute deviation. Journal of the American Statistical Association, 88, 1273–1283.
Rousseeuw, P. J. , & Driessen, V. K. (1999). A fast algorithm for the minimum covariance determinant estimator. Technometrics, 41, 212–223.
Rousseeuw, P. J. , & Leroy, A. M. (1987). Robust regression and outlier detection. Wiley series in probability and mathematical statistics. In Applied probability and statistics. New York: Wiley.
Rousseeuw, P. J. , & Yohai, V. J. (1984). Robust regression by means of S-estimators. In robust and nonlinear time series analysis. Lecture Notes in Statistics, 26, 256–276.
Rousseeuw, P. J. , & Zomeren, B. C. (1991). Robust distances: Simulations and cutoff values, directions in robust statistics and diagnostics (pp. 195–203). New York: Springer.
Sirkia,S. , Taskinen, S. , & Oja, H. (2007). Symmetrized M-estimators of multivariate scatter. Journal of Multivariate Analysis, 98, 1611–1629.

Skocaj, D. , Bischof, H. , and Leonardis, A. (2002). A robust PCA algorithm for building representations from panoramic images. In European Conference Computer Vision, 761–775.
Tatsuoka, K. S. , & Tyler, D. E. (2000). On the uniqueness of S-functionals and M-functionals under nonelliptical distributions. The Annals of Statistics, 28(4), 1219–1243.
Tiku, M. L. , & Akkaya, A. D. (2004). Robust estimation and hypothesis testing. New Delhi: New Age International Limited.
Todorov, V. , & Filzmoser, P. (2009). An object-oriented framework for robust multivariate analysis. Journal of Statistical Software, 32(3), 1–47.
Todorov, V. , & Pires, A. M. (2007). Comparative performance of several robust linear discriminant analysis. REVSTAT Statistical Journal, 5(1), 63–83.
Torre, F. D. , & Black, M. J. (2003). A framework for robust subspace learning. International Journal of Computer Vision, 54, 117–142.
Tukey, J. W. (1960). A survey of sampling from contaminated distributions. Contributions to probability and statistics (I. Olkin , Ed.). Stanford, CA: Stanford University Press.
Tukey, J. W. (1962). The future of data analysis. The Annals of Mathematical Statistics, 33, 1–67.
Tukey, J. W. (1975). Mathematics and picturing data, In Proceedings of the International Congress on Mathematics, R. D. James (ed.). Canadian Mathematics Congress, 2, 523–531.
Tyler, D. E. (1987). A distribution-free m-estimator of multivariate scatter. The Annals Statistics, 15, 234–251.
Vardi, Y. , & Zhang, C. H. (2000). The multivariate L1-median and associated data depth. Proceedings of the National Academy of Sciences, USA, 97, 1423–1426.
Wilcox, R. (2010). Fundamentals of modern statistical methods. New York: Springer.
Wilcox, R. (2017). Introduction to robust estimation and hypothesis testing. San Diego, CA: Elsevier.
Yohai, V. J. (1987). High breakdown-point and high-efficiency robust estimates for regression. The Annals of Statistics, 15(2), 642–656.
Zuo, Y. (2003). Projection-based depth functions and associated medians. Annals of Statistics, 31, 1460–1490.
Zuo, Y. , & Serfling, R. (2000). General notions of statistical depth function. Annals of Statistics, 28, 461–482.

Big Data in Tribal Healthcare and Biomedical Research

1000 Genomes Project Consortium . (2015). A global reference for human genetic variation. Nature, 526, 68–74.
Albanese, D. , & Donati, C. (2017). Strain profiling and epidemiology of bacterial species from metagenomic sequencing. Nature Communications, 8, 1–14.
Andreu-Perez, J. , Poon, C. C. , Merrifield, R. D. , Wong, S. T. , & Yang, G.-Z. (2015). Big data for health. IEEE Journal of Biomedical and Health Informatics, 19, 1193–1208.
Basu, S. (2000). Dimensions of tribal health in India. Health and Population Perspectives and Issues, 23, 61–70.
Belle, A. , Thiagarajan, R. , Soroushmehr, S. , Navidi, F. , Beard, D. A. , & Najarian, K. (2015). Big data analytics in healthcare. BioMed Research International, 2015: 370194.
Benke, K. , & Benke, G. (2018). Artificial intelligence and big data in public health. International Journal of Environmental Research and Public Health, 15, 2796.
Blazquez, D. , & Domenech, J. (2018). Big Data sources and methods for social and economic analyses. Technological Forecasting and Social Change, 130, 99–113.
Boyd, D. , & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, Communication & Society, 15, 662–679.
Breitwieser, F. P. , Lu, J. , & Salzberg, S. L. (2019). A review of methods and databases for metagenomic classification and assembly. Briefings in Bioinformatics, 20, 1125–1136.
Cancilla, M. , Powell, I. , Hillier, A. , & Davidson, B. (1992). Rapid genomic fingerprinting of Lactococcus lactis strains by arbitrarily primed polymerase chain reaction with 32P and fluorescent labels. Applied and Environmental Microbiology, 58, 1772–1775.

Chapman, P. , Clinton, J. , Kerber, R. , Khabaza, T. , Reinartz, T. , Shearer, C. , & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. SPSS Inc, 9, 13.

Choudhury, S. , Fishman, J. R. , McGowan, M. L. , & Juengst, E. T. (2014). Big data, open science and the brain: Lessons learned from genomics. Frontiers in Human Neuroscience, 8, 239.

Church, D. (2015). Schneider V a, Steinberg K, Schatz MC, Quinlan AR, Chin CS, et al. Extending reference assembly models. Genome Biology, 16, 13.

Clifford, L. (2008). Big Data: How do your data grow. Nature, 455, 28–29.

De Mandal, S. , Panda, A. , Bisht, S. , & Kumar, N. (2015). Microbial ecology in the era of next generation sequencing. Next Generation Sequencing and Applications, 1, 2.

Dean, J. , & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. Communications of the ACM, 51, 107–113.

Elend, C. , Schmeisser, C. , Leggewie, C. , Babiak, P. , Carballeira, J. D. , Steele, H. , Reymond, J.-L. , Jaeger, K.-E. , & Streit, W. (2006). Isolation and biochemical characterization of two novel metagenome-derived esterases. Applied and Environmental Microbiology, 72, 3637–3645.

Fayyad, U. , Piatetsky-Shapiro, G. , & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, 39, 27–34.

Forbes, S. A. , Bindal, N. , Bamford, S. , Cole, C. , Kok, C. Y. , Beare, D. , Jia, M. , Shepherd, R. , Leung, K. , & Menzies, A. (2010). COSMIC: Mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Research, 39, D945–D950.

Grasnick, B. , Perscheid, C. , & Uflacker, M. (2018). A framework for the automatic combination and evaluation of gene selection methods. Presented at the International Conference on Practical Applications of Computational Biology & Bioinformatics, Springer, pp. 166–174.

Griffith, M. , Spies, N. C. , Krysiak, K. , McMichael, J. F. , Coffman, A. C. , Danos, A. M. , Ainscough, B. J. , Ramirez, C. A. , Rieke, D. T. , & Kujan, L. (2017). CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nature Genetics, 49, 170.

Handelsman, J. (2005). Metagenomics: Application of genomics to uncultured microorganisms. Microbiology and Molecular Biology Reviews, 69, 195–195.

Jin, X. , Wah, B. W. , Cheng, X. , & Wang, Y. (2015). Significance and challenges of big data research. Big Data Research, 2, 59–64.

Khomtchouk, B. B. , Hennessy, J. R. , & Wahlestedt, C. (2017). shinyheatmap: Ultra fast low memory heatmap web interface for big data genomics. PloS One, 12(5), e0176334.

Landrum, M. J. , Lee, J. M. , Riley, G. R. , Jang, W. , Rubinstein, W. S. , Church, D. M. , & Maglott, D. R. (2014). ClinVar: Public archive of relationships among sequence variation and human phenotype. Nucleic Acids Research, 42, D980–D985.

Langmead, B. , Schatz, M. C. , Lin, J. , Pop, M. , & Salzberg, S. L. (2009). Searching for SNPs with cloud computing. Genome Biology, 10, R134.

Muyzer, G. , De Waal, E. C. , & Uitterlinden, A. G. (1993). Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. Applied and Environmental Microbiology, 59, 695–700.

Narain, J. P. , Jain, S. , Bora, D. , & Venkatesh, S. (2015). Eradicating successfully yaws from India: The strategy & global lessons. The Indian Journal of Medical Research, 141, 608.

Naseriparsa, M. , Bidgoli, A.-M. , & Varaee, T. (2014). A hybrid feature selection method to improve performance of a group of classification algorithms. arXiv preprint arXiv:1403.2372.

Nichols, D. S. , Sanderson, K. , Buia, A. , Van De Kamp, J. , Holloway, P. , Bowman, J. P. , Smith, M. , Mancuso Nichols, C. , Nichols, P. , & McMeekin, T. (2002). Bioprospecting and biotechnology in Antarctica. The Antarctic: past, present and future, Antarctic CRC research report 28.

Nickerson, M. (2017). Characteristics of a nation-to-nation relationship. Ottawa: Institute on Governance.

O'Driscoll, A. , Daugelaite, J. , & Sleator, R. D. (2013). “Big data,” Hadoop and cloud computing in genomics. Journal of Biomedical Informatics, 46, 774–781.

Oulas, A. , Pavloudi, C. , Polymenakou, P. , Pavlopoulos, G. A. , Papanikolaou, N. , Kotoulas, G. , Arvanitidis, C. , & Iliopoulos, I. (2015). Metagenomics: Tools and Insights for Analyzing Next-Generation Sequencing Data Derived from Biodiversity Studies. Bioinformatics and Biology Insights, 9, 75-88

Pace, N. R. (1997). A molecular view of microbial diversity and the biosphere. Science, 276, 734–740.
Pasolli, E. , Truong, D. T. , Malik, F. , Waldron, L. , & Segata, N. (2016). Machine learning meta-analysis of large metagenomic datasets: Tools and biological insights. PLoS Computational Biology, 12. doi:10.1371/journal.pcbi.1004977
Pireddu, L. , Leo, S. , & Zanetti, G. (2011). SEAL: A distributed short read mapping and duplicate removal tool. Bioinformatics, 27, 2159–2160.
Polyakova, A. G. , Loginov, M. P. , Serebrennikova, A. I. , & Thalassinos, E. (2019). Design of a socio-economic processes monitoring system based on network analysis and big data. International Journal of Economics and Business Administration, VII(1), 130–139.
Rainie, S. C. , Schultz, J. L. , Briggs, E. , Riggs, P. , & Palmanteer-Holder, N. L. (2017). Data as a strategic resource: Self-determination, governance, and the data challenge for Indigenous nations in the United States. International Indigenous Policy Journal, 8(2). doi:10.18584/iipj.2017.8.2.1
Reiman, D. , Metwally, A. , & Dai, Y. (2017). Using convolutional neural networks to explore the microbiome. Presented at the 2017 39th annual international conference of the IEEE engineering in medicine and biology society (EMBC), IEEE, pp. 4269–4272.
Reisman, M. (2017). EHRs: The challenge of making electronic data usable and interoperable. Pharmacy and Therapeutics, 42, 572.
Rinke, C. , Schwientek, P. , Sczyrba, A. , Ivanova, N. N. , Anderson, I. J. , Cheng, J.-F. , Darling, A. , Malfatti, S. , Swan, B. K. , & Gies, E. A. (2013). Insights into the phylogeny and coding potential of microbial dark matter. Nature, 499, 431–437.
Rodriguez-Lonebear, D. (2016). Building a data revolution in Indian country. In Tahu Kukutai & Taylor John (Eds.), Indigenous data sovereignty: Toward an agenda (pp. 253–272). Australia: Australian National University Press.
Ruegg, J. , Gries, C. , Bond-Lamberty, B. , Bowen, G. J. , Felzer, B. S. , McIntyre, N. E. , Soranno, P. A. , Vanderbilt, K. L. , & Weathers, K. C. (2014). Completing the data life cycle: Using information management in macrosystems ecology research. Frontiers in Ecology and the Environment, 12, 24–30.
Schatz, M. C. (2009). CloudBurst: Highly sensitive read mapping with MapReduce. Bioinformatics, 25, 1363–1369.
Shahrivari, S. (2014). Beyond batch processing: towards real-time and streaming big data. Computers, 3, 117–129.
Shameer, K. , Badgeley, M. A. , Miotto, R. , Glicksberg, B. S. , Morgan, J. W. , & Dudley, J. T. (2017). Translational bioinformatics in the era of real-time biomedical, health care and wellness data streams. Briefings in Bioinformatics, 18, 105–124.
Sheffield, V. , Beck, J. , Stone, E. , & Myers, R. (1992). A simple and efficient method for attachment of a 40-base pair, GC-rich sequence to PCR-amplified DNA. BioTechniques, 12, 386–388.
Shvachko, K. , Kuang, H. , Radia, S. , & Chansler, R. (2010). The hadoop distributed file system. Presented at the 2010 IEEE 26th symposium on mass storage systems and technologies (MSST), Ieee, pp. 1–10.
Simonet, A. , Fedak, G. , & Ripeanu, M. (2015). Active data: A programming model to manage data life cycle across heterogeneous systems and infrastructures. Future Generation Computer Systems, 53, 25–42.
Sims, D. , Sudbery, I. , Ilott, N. E. , Heger, A. , & Ponting, C. P. (2014). Sequencing depth and coverage: Key considerations in genomic analyses. Nature Reviews Genetics, 15, 121.
Singh, D. , & Reddy, C. K. (2015). A survey on platforms for big data analytics. Journal of Big Data, 2, 8.
Torsvik, V. L. , & Øvreås, L. (2011). DNA reassociation yields broadscale information on metagenome complexity and microbial diversity. In Handbook of Molecular Microbial Ecology I: Metagenomics and Complementary Approaches (pp. 3–16). New York: Wiley-Blackwell.
Truong, D. T. , Franzosa, E. A. , Tickle, T. L. , Scholz, M. , Weingart, G. , Pasolli, E. , Tett, A. , Huttenhower, C. , & Segata, N. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nature Methods, 12, 902–903.
Tsymbal, A. , Pechenizkiy, M. , & Cunningham, P. (2005). Diversity in search strategies for ensemble feature selection. Information Fusion, 6, 83–98.

Turnbaugh, P. J. , Ley, R. E. , Hamady, M. , Fraser-Liggett, C. M. , Knight, R. , & Gordon, J. I. (2007). The human microbiome project. Nature, 449, 804–810.
Uchiyama, T. , Abe, T. , Ikemura, T. , & Watanabe, K. (2005). Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes. Nature Biotechnology 23, 88–93.
Vakhlu, J. , Sudan, A. K. , & Johri, B. (2008). Metagenomics: future of microbial gene mining. Indian Journal of Microbiology, 48, 202–215.
Vicente, M. R. , López-Menéndez, A. J. , & Pérez, R. (2015). Forecasting unemployment with internet search data: Does it help to improve predictions when job destruction is skyrocketing? Technological Forecasting and Social Change, 92, 132–139.
Voget, S. , Steele, H. , & Streit, W. (2006). Characterization of a metagenome-derived halotolerant cellulase. Journal of Biotechnology, 126, 26–36.
Walsh, A. M. , Crispie, F. , O'Sullivan, O. , Finnegan, L. , Claesson, M. J. , & Cotter, P. D. (2018). Species classifier choice is a key consideration when analysing low-complexity food microbiome data. Microbiome, 6, 50.
Waschkowitz, T. , Rockstroh, S. , & Daniel, R. (2009). Isolation and characterization of metalloproteases with a novel domain structure by construction and screening of metagenomic libraries. Applied and Environmental Microbiology, 75, 2506–2516.
West, M. , Ginsburg, G. S. , Huang, A. T. , & Nevins, J. R. (2006). Embracing the complexity of genomic data for personalized medicine. Genome Research, 16, 559–566.
Zaharia, M. , Xin, R. S. , Wendell, P. , Das, T. , Armbrust, M. , Dave, A. , Meng, X. , Rosen, J. , Venkataraman, S. , & Franklin, M. J. (2016). Apache spark: A unified engine for big data processing. Communications of the ACM, 59, 56–65.

How to Implement Data Lake for Large Enterprises

AWS Services . Retrieved April 15, 2021, from https://aws.amazon.com/big-data/datalakes-and-analytics/
Azure Cloud Services . Retrieved April 10, 2021, from https://azure.microsoft.com/en-in/services/data-lake-analytics
Google Cloud . Retrieved April 12, 2021, from https://cloud.google.com/solutions/smart-analytics

A Novel Application of Data Mining Techniques for Satellite Performance Analysis

Barua, A. , & Khorasani, K. (2011). Hierarchical fault diagnosis and health monitoring in satellites formation flight. IEEE Transactions on Systems, Man, and Cybernetics—Part C:Applications and Reviews, 41(2), 223–239. 10.1109/TSMCC.2010.2049994
Džeroski, S. , & Lavrac, N. (2001). Relational Data Mining. Dzeroski, S. , & Lavrac, N. , Ed. ISBN: 3540422897, September 2001. 10.1007/978-3-662-04599-2
Fayyad, U. , Piatetshy-Shapiro, G. , Smyth, P. , & Uthurusamy, R. (1996). Advances in knowledge discovery and data mining. CA, United States: AAAI/MIT Press, 445 Burgess Drive Menlo Park.
Gao, Y. , Yang, T. , Xing, N. , & Xu, M. (2012). Fault Detection and Diagnosis for Spacecraft using Principal Component Analysis and Support Vector Machines, *7th IEEE Conference on Industrial Electronics and Applications (ICIEA)* , 978-1-4577-2119-9/12/$26.00 ©2011 IEEE, pp. 1984–1988.
Gao, Y. , Yang, T. , Xu, M. , & Xing, N. (2012). An Unsupervised Anomaly Detection Approach for Spacecraft Based on Normal Behavior Clustering, *2012 Fifth International Conference on Intelligent Computation Technology and Automation* , 978-0-7695-4637-7/12 $26.00 © 2012 IEEE, DOI 10.1109/ICICTA.2012.126, pp. 478–481.

Kantardzic, M. (2003). Data mining – Concepts, models, methods and algorithms. New York: Wiley-Interscience, John Wiley & Sons Inc. pp. 91 -193.
Khalid, B. , & Abdelwahab, N. 2016. A comparative study of various data mining techniques: Statistics, decision trees and neural networks. International Journal of Computer Applications Technology and Research, 5(3), 172–175.
Liu, H. , & Motoda, H. (2000). Feature selection for knowledge discovery and data mining (2nd Printing). Boston, MA: Kluwer Academic Publisher.
Liu, H. , & Motoda, H. (2001). Instance selection and construction for data mining. Boston, MA: Kluwer Academic Publishers.
Liu, H. M. (1998). Feature extraction, construction and selection: A data mining perspective. Boston, MA: Kluwer Academic Publishers.
Liu, X. , Kwan, B. W. , & Foo, S. Y. (2008). Time series prediction based on fuzzy principles, *Proc. Huntsville Simulation Conference* , Department of Electrical & Computer Engineering FAMU-FSU College of Engineering, Florida State University.
Ma, L. , Liu, H. , & Feng, Z. (2012). An equipment failure prediction accuracy improvement method based on the gray GM (I, I) model, artificial intelligence and computational intelligence (pp. 294–300). Berlin Heidelberg: Springer.
Maimon, O. , & Rokach, L. (2005). Top-down induction of decision trees classifiers—A survey. IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews), 35, 476–487. 10.1109/tsmcc.2004.843247
Maimon, O. , & Rokach, L. (2009). Data mining and knowledge discovery handbook (pp. 165–187, 277–293, 353–373). New Delhi: Springer Science + Business Media Inc.
Maini, A. K. , & Agrawal, V. (2007). Satellite technology – Principles and applications (pp. 303–466). New York: John Wiley & Sons Ltd.
Mitchell, T. (1997). Machine learning (pp. 16–25). New York, NY: McGraw Hill.
Ömer, A. (2013). A rule induction algorithm for knowledge discovery and classification. Turkish Journal of Electrical Engineering and Computer Sciences, 1–38. 10.3906/elk-1202-27
Pham, D. T. , & Afify, A. A. (2005). RULES-6: A simple rule induction algorithm for supporting decision-making. IEEE. 0-7803-9252-3/05/$20.00.
Poojari, A. K. (2003). Data mining techniques (pp. 54–87). Hyderabad: Universities Press.
Ramaswamy, S. , Rastogi, R. , & Shim, K. (2000). Efficient Algorithms for Mining Outliers from Large Datasets, *Proc. SIGMOD2000* , ACM Press, pp. 162–172.
Takehisa, Y. , Kawahara, Y. , & Takata, N. (2010). Spacecraft Telemetry Data Monitoring by Dimensionality Reduction Techniques, *SICE Annual Conference 2010 August 18–21, 2010* , Taipei, Taiwan, pp. 1230–1234.
Weiss, S. M. , & Indurkhya, N. (1998) Predictive Data Mining: a Practical Guide, *Morgan Kaufman, Sun Francisco*.
Yairi, T. , Kawahara, Y. , Fujimaki, R. , Sato, Y. , & Machida, K. (2006). Telemetry-mining: A Machine Learning Approach to Anomaly Detection and Fault Diagnosis for Space Systems, *Proc. of the 2nd IEEE International Conference on Space Mission Challenges for Information Technology (SMC-IT 2006)* , July 17–20, Pasadena, California, USA.
Yang, T. , Chen, B. , Gao, Y. , Feng, J. , Zhang, H. , & Wang, X. (2013). Data Mining-Based Fault Detection and Prediction Methods for In-Orbit Satellite, *2013 2nd International Conference on Measurement, Information and Control, Harbin, China,978-1-4799-1392-3/13/$31.00 m013 IEEE* , pp.805–808.

Big Data Analytics: A Text Mining Perspective and Applications in Biomedicine and Healthcare

Abaya, S. A. (2012). Association rule mining based on Apriori algorithm in minimizing candidate generation. International Journal of Scientific & Engineering Research, 3(7), 1–4.
Abdulkadhar, S. , Murugesan, G. , & Natarajan, J. (2020). Classifying protein-protein interaction articles from biomedical literature using many relevant features and context-free grammar. Journal of King Saud University-Computer and Information Sciences, 32(5) 553-560.

AbuZeina, D. , & Al-Anzi, F. S. (2018). Employing fisher discriminant analysis for Arabic text classification. Computers & Electrical Engineering, 66, 474–486.
Adnan, K. , & Akbar, R. (2019). An analytical study of information extraction from unstructured and multidimensional big data. Journal of Big Data, 6(1), 91.
Agerri, R. , Artola, X. , Beloki, Z. , Rigau, G. , & Soroa, A. (2015). Big data for natural language processing: A streaming approach. Knowledge-Based Systems, 79, 36–42.
Aggarwal, C. C. , & Cheng, X.Z. (2012). “A survey of text clustering algorithms.” Mining text data (pp. 77–128). Boston, MA: Springer.
Aggarwal, C. C. , & Zhai, C. (2012). An introduction to text mining. In Mining text data (pp. 1–10). Boston, MA: Springer.
Akter, S. , & Wamba, S. F. (2016). Big data analytics in E-commerce: A systematic review and agenda for future research. Electronic Markets, 26(2), 173–194.
Amado, A. , Cortez, P. , Rita, P. , & Moro, S. (2018). Research trends on Big Data in Marketing: A text mining and topic modeling based literature analysis. European Research on Management and Business Economics, 24(1), 1–7.
Arshad, O. , Gallo, I. , Nawaz, S. , & Calefati, A. (2019). Aiding intra-text representations with visual context for multimodal named entity recognition. arXiv preprint arXiv:1904.01356.
Ayre, K. K. , Caldwell, C. A. , Stinson, J. , & Landis, W.G. (2014). Analysis of regional scale risk of whirling disease in populations of Colorado and Rio Grande cutthroat trout using a Bayesian belief network model. Risk Analysis, 34(9), 1589–1605.
Azzag, H. , Guinot, C. , & Venturini, G. (2006). Data and text mining with hierarchical clustering ants. In Swarm intelligence in data mining (pp. 153–189). Berlin, Heidelberg: Springer.
Baro, E. , Degoul, S. , Beuscart, R. , & Chazard, E. (2015). Toward a literature-driven definition of big data in healthcare. BioMed Research International, 1–9. 639021.
Bashir, S. , Qamar, U. , & Khan, F. H. (2015). BagMOOV: A novel ensemble for heart disease prediction bootstrap aggregation with multi-objective optimized voting. Australasian Physical & Engineering Sciences in Medicine, 38(2), 305–323.
Batbaatar, E. , & Ryu, K. H. (2019). Ontology-based healthcare named entity recognition from Twitter messages using a recurrent neural network approach. International Journal of Environmental Research and Public Health, 16(19), 3628.
Bhasuran, B. , Murugesan, G. , Abdulkadhar, S. , & Natarajan, J. (2016). Stacked ensemble combined with fuzzy matching for biomedical named entity recognition of diseases. Journal of Biomedical Informatics, 64, 1–9.
Bhasuran, B. , & Natarajan, J. (2018). Automatic extraction of gene-disease associations from literature using joint ensemble learning. PloS One, 13(7), e0200699.
Bhasuran, B. , & Natarajan, J. (2019). Distant supervision for large-scale extraction of gene–disease associations from literature using DeepDive. In International Conference on Innovative Computing and Communications (pp. 367–374). Springer, Singapore.
Bhasuran, B. , Subramanian, D. , & Natarajan, J. (2018). Text mining and network analysis to find functional associations of genes in high altitude diseases. Computational Biology and Chemistry, 75, 101–110.
Björne, J. , Heimonen, J. , Ginter, F. , Airola, A. , Pahikkala, T. , & Salakoski, T. (2011). Extracting contextualized complex biological events with rich graph-based feature sets. Computational Intelligence, 27(4), 541–557.
Boytcheva, S. , Angelova, G. , Angelov, Z. , & Tcharaktchiev, D. (2015). Text mining and big data analytics for retrospective analysis of clinical texts from outpatient care. Cybernetics and Information Technologies, 15(4), 58–77.
Chen, H. C. , Chen, Z. Y. , Huang, S. Y. , Ku, L. W. , Chiu, Y. S. , & Yang, W. J. (2018, July). Relation extraction in knowledge base question answering: From general-domain to the catering industry. In International Conference on HCI in Business, Government, and Organizations (pp. 26–41). Springer, Cham.
Choi, E. , Bahadori, M. T. , Schuetz, A. , Stewart, W. F. , & Sun, J. (2016a, December). Doctor AI: Predicting clinical events via recurrent neural networks. In Machine Learning for Healthcare Conference (pp. 301–318).
Choi, E. , Bahadori, M. T. , Searles, E. , Coffey, C. , Thompson, M. , Bost, J. , Tejedor-Sojo, J. , & Sun, J. (2016b, August). Multi-layer representation learning for medical concepts. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1495–1504).

Cohen, A. M. , & Hersh, W. R. (2005). A survey of current work in biomedical text mining. Briefings in Bioinformatics, 6(1), 57–71.
Cohen, K. B. , & Hunter L. (2008). Getting started in text mining. PloS Comput Biol, 4, e20.
Crone, S. F. , & Koeppel, C. (2014, March). Predicting exchange rates with sentiment indicators: An empirical evaluation using text mining and multilayer perceptrons. In 2014 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr) (pp. 114–121). IEEE.
Dai, H. J. , Wu, J. C. Y. , Lin, W. S. , Reyes, A. J. F. , Syed-Abdul, S. , Tsai, R. T. H. , & Hsu, W. L. (2014). LiverCancerMarkerRIF: A liver cancer biomarker interactive curation system combining text mining and expert annotations. Database, 2014.
De Mauro, A. , Greco, M. , & Grimaldi, M. (2015, February). What is big data? A consensual definition and a review of key research topics. In AIP conference proceedings (Vol. 1644, No. 1, pp. 97–104). American Institute of Physics.
Deekshatulu, B. L. , & Chandra, P. (2013). Classification of heart disease using k-nearest neighbor and genetic algorithm. Procedia Technology, 10, 85–94.
Dimitrov, D. V. (2016). Medical internet of things and big data in healthcare. Healthcare Informatics Research, 22(3), 156–163.
Edo-Osagie, O. , Smith, G. , Lake, I. , Edeghere, O. , & De La Iglesia, B. (2019). Twitter mining using semi-supervised classification for relevance filtering in syndromic surveillance. PloS One, 14(7), e0210689.
Feldman, R. , & Sanger, J. (2007). The text mining handbook: Advanced approaches in analyzing unstructured data. Cambridge: Cambridge University Press.
FerreiraMello, R. , André, M. , Pinheiro, A. , Costa, E. , & Romero, C. (2019). Text mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(6), e1332.
Fiorini, N. , Leaman, R. , Lipman, D. J. , & Lu, Z. (2018). How user intelligence is improving PubMed. Nature Biotechnology, 36(10), 937–945.
Fonarow, G. C. , Adams, K. F. , Abraham, W. T. , Yancy, C. W. , Boscardin, W. J. and ADHERE Scientific Advisory Committee (2005). Risk stratification for in-hospital mortality in acutely decompensated heart failure: Classification and regression tree analysis. JAMA, 293(5), 572–580.
Fraley, C. , & Hesterberg, T. (2009). Least angle regression and LASSO for large datasets. Statistical Analysis and Data Mining: The ASA Data Science Journal, 1(4), 251–259.
Francis, S. , Van Landeghem, J. , & Moens, M. F. (2019). Transfer learning for named entity recognition in financial and biomedical documents. Information, 10(8), 248.
Georgescu, T. M. (2020). Natural language processing model for automatic analysis of cybersecurity-related documents. Symmetry, 12(3), 354.
Gök, A. , Waterworth, A. , & Shapira, P. (2015). Use of web mining in studying innovation. Scientometrics, 102(1), 653–671.
Gotz, D. , Stavropoulos, H. , Sun, J. , & Wang, F. (2012). ICDA: a platform for intelligent care delivery analytics. In AMIA annual symposium proceedings (Vol. 2012, p. 264). American Medical Informatics Association.
Gunn, S. R. (1998). Support vector machines for classification and regression. ISIS Technical Report, 14(1), 5–16.
Hanif, S.M. , & Prevost, L. (2009, July). Text detection and localization in complex scene images using constrained adaboost algorithm. In 2009 10th international conference on document analysis and recognition (pp. 1–5). IEEE.
Hashem, I. A. T. , Yaqoob, I. , Anuar, N. B. , Mokhtar, S. , Gani, A. , & Khan, S.U. (2015). The rise of “big data” on cloud computing: Review and open research issues. Information Systems, 47, 98–115.
Hassani, H. , Beneki, C. , Unger, S. , Mazinani, M. T. , & Yeganegi, M. R. (2020). Text mining in Big Data analytics. Big Data and Cognitive Computing, 4(1), 1.
He, K. , Hong, N. , Lapalme-Remis, S. , Lan, Y. , Huang, M. , Li, C. , & Yao, L. (2019). Understanding the patient perspective of epilepsy treatment through text mining of online patient support groups. Epilepsy & Behavior, 94, 65–71.
Hearst, M. A. (1999, June). Untangling text data mining. In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics (pp. 3–10). Association for Computational Linguistics.

Herland, M. , Khoshgoftaar, T. M. , & Wald, R. (2014). A review of data mining using big data in health informatics. Journal of Big Data, 1(1), 1–35.
Hilbert, M. (2016). Big data for development: A review of promises and challenges. Development Policy Review, 34(1), 135–174.
Hirschberg, J. , & Manning, C.D. (2015). Advances in natural language processing. Science, 349(6245), 261–266.
Hogenboom, F. , Frasincar, F. , Kaymak, U. , & De Jong, F. (2011, October). An overview of event extraction from text. In DeRiVE@ ISWC (pp. 48–57).
Hsiao, Y. W. , & Lu, T. P. (2019). Text-mining in cancer research may help identify effective treatments. Translational Lung Cancer Research, 8(Suppl 4), S460.
Inmon, W. H. , & Linstedt, D. (2014). Data architecture: a primer for the data scientist: big data, data warehouse and data vault. Morgan Kaufmann.
Isa, D. , Kallimani, V. P. , & Lee, L. H. (2009). Using the self organizing map for clustering of text documents. Expert Systems with Applications, 36(5), 9584–9591.
Jahromi, A. H. , & Taheri, M. (2017, October). A non-parametric mixture of Gaussian naive Bayes classifiers based on local independent features. In 2017 Artificial Intelligence and Signal Processing Conference (AISP) (pp. 209–212). IEEE.
Jelodar, H. , Wang, Y. , Orji, R. , & Huang, H. (2020). Deep sentiment classification and topic discovery on novel coronavirus or covid-19 online discussions: Nlp using lstm recurrent neural network approach. arXiv preprint arXiv:2004.11695.
Jensen, L. J. , Saric, J. , & Bork, P. (2006). Literature mining for the biologist: from information retrieval to biological discovery. Nature reviews genetics, 7(2), 119–129.
Jiang, M. , Liang, Y. , Feng, X. , Fan, X. , Pei, Z. , Xue, Y. , & Guan, R. (2018). Text classification based on deep belief network and softmax regression. Neural Computing and Applications, 29(1), 61–70.
Jing, L. , Ng, M. K. , Xu, J. , & Huang, J. Z. (2005, May). Subspace clustering of text documents with feature weighting k-means algorithm. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 802–812). Springer, Berlin, Heidelberg.
Jurca, G. , Addam, O. , Aksac, A. , Gao, S. , Özyer, T. , Demetrick, D. , & Alhajj, R. (2016). Integrating text mining, data mining, and network analysis for identifying genetic breast cancer trends. BMC Research Notes, 9(1), 236.
Kang, M. , Ahn, J. , & Lee, K. (2018). Opinion mining using ensemble text hidden Markov models for text classification. Expert Systems with Applications, 94, 218–227.
Kang, N. , van Mulligen, E. M. , & Kors, J. A. (2011). Comparing and combining chunkers of biomedical text. Journal of Biomedical Informatics, 44, 354–360.
Kaur, S. , & Rashid, E. M. (2016). Web news mining using Back Propagation Neural Network and clustering using K-Means algorithm in big data. Indian Journal of Science and Technology, 9(41), 1–8.
Kawashima, K. , Bai, W. , & Quan, C. (2017, June). Text mining and pattern clustering for relation extraction of breast cancer and related genes. In 2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD) (pp. 59–63). IEEE.
Kimball, R. (2011). The evolving role of the enterprise data warehouse in the era of big data analytics. Whitepaper, Kimball Group, April.
Korhonen, A. , Séaghdha, D. Ó. , Silins, I. , Sun, L. , Högberg, J. , & Stenius, U. (2012). Text mining for literature review and knowledge discovery in cancer risk assessment and research. PloS One, 7(4), e33427.
Larsen, J. , Szymkowiak, A. , & Hansen, L. K. (2002). Probabilistic hierarchical clustering with labeled and unlabeled data. International Journal of Knowledge Based Intelligent Engineering Systems, 6(1), 56–63.
Lee, J. , Yoon, W. , Kim, S. , Kim, D. , Kim, S. , So, C. H. , & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234–1240.
Liu, J. , Zhao, S. , & Wang, G. (2018). SSEL-ADE: a semi-supervised ensemble learning framework for extracting adverse drug events from social media. Artificial Intelligence in Medicine, 84, 34–49.
Liu, M. , Jiang, L. , & Hu, H. (2017). Automatic extraction and visualization of semantic relations between medical entities from medicine instructions. Multimedia Tools and Applications, 76(8),

10555–10573.
Lu, Z. (2011). PubMed and beyond: a survey of web tools for searching biomedical literature. Database (Oxford); 2011:baq036.
Lv, X. , Guan, Y. , Yang, J. , & Wu, J. (2016). Clinical relation extraction with deep learning. International Journal of Hybrid Information Technology, 9(7), 237–248.
Maroli, N. , Kalagatur, N. K. , Bhasuran, B. , Jayakrishnan, A. , Manoharan, R. R. , Kolandaivel, P. , Natarajan, J. , & Kadirvelu, K. (2019). Molecular mechanism of T-2 toxin-induced cerebral edema by aquaporin-4 blocking and permeation. Journal of Chemical Information and Modeling, 59(11), 4942–4958.
McAullay, D. , Williams, G. , Chen, J. , Jin, H. , He, H. , Sparks, R. , & Kelman, C. (2005, January). A delivery framework for health data mining and analytics. In Proceedings of the Twenty-eighth Australasian conference on Computer Science-Volume 38 (pp. 381–387). Australian Computer Society, Inc.
McRoy, S. , Rastegar-Mojarad, M. , Wang, Y. , Ruddy, K. J. , Haddad, T. C. , & Liu, H. (2018). Assessing unmet information needs of breast cancer survivors: Exploratory study of online health forums using text classification and retrieval. JMIR Cancer, 4(1), e10.
Miotto, R. , Li, L. , Kidd, B. A. , & Dudley, J. T. (2016). Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. Scientific Reports, 6(1), 1–10.
Mirza, B. , Wang, W. , Wang, J. , Choi, H. , Chung, N. C. , & Ping, P. (2019). Machine learning and integrative analysis of biomedical big data. Genes, 10(2), 87.
Mukherjee, A. , & Zhu, J. (2011). Reduced rank ridge regression and its kernel extensions. Statistical Analysis and Data Mining: The ASA Data Science Journal, 4(6), 612–622.
Murphy, S. N. , Weber, G. , Mendis, M. , Gainer, V. , Chueh, H. C. , Churchill, S. , & Kohane, I. (2010). Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). Journal of the American Medical Informatics Association, 17(2), 124–130.
Murugesan, G. , Abdulkadhar, S. , Bhasuran, B. , & Natarajan, J. (2017a). BCC-NER: bidirectional, contextual clues named entity tagger for gene/protein mention recognition. EURASIP Journal on Bioinformatics and Systems Biology, 2017(1), 7.
Murugesan, G. , Abdulkadhar, S. , & Natarajan, J. (2017b). Distributed smoothed tree kernel for protein-protein interaction extraction from the biomedical literature. PLOS One, 12, e0187379.
Ohno-Machado, L. , Bafna, V. , Boxwala, A. A. , Chapman, B. E. , Chapman, W. W. , Chaudhuri, K. , Day, M. E. , Farcas, C. , Heintzman, N. D. , Jiang, X. , & Kim, H. (2012). iDASH: Integrating data for analysis, anonymization, and sharing. Journal of the American Medical Informatics Association, 19(2), 196–201
Omar, Y. (2015). Al-Jarrah, Paul D. Yoo, Sami Muhaidat, George K. Karagiannidis, Kamal Taha, Efficient Machine Learning for Big Data. Big Data Research, 2(3), 87–93.
Palangi, H. , Deng, L. , Shen, Y. , Gao, J. , He, X. , Chen, J. , Song, X. , & Ward, R. (2016). Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24(4), 694–707.
Paul, M. J. , & Dredze, M. (2014). Discovering health topics in social media using topic models. PloS One, 9(8), e103408.
Pham, T. , Tran, T. , Phung, D. , & Venkatesh, S. (2016, April). Deepcare: A deep dynamic memory model for predictive medicine. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 30–41). Springer, Cham.
Poria, S. , Cambria, E. , & Gelbukh, A. (2015, September). Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In Proceedings of the 2015 conference on empirical methods in natural language processing (pp. 2539–2544).
Pouyanfar, S. , Yang, Y. , Chen, S. C. , Shyu, M. L. , & Iyengar, S. S. (2018). Multimedia big data analytics: A survey. ACM Computing Surveys (CSUR), 51(1), 1–34.
Qiang, L. I. (2006). A Comparative Study on Algorithms of Constructing Decision Trees——ID3, C4. 5 and C5. 0 (J). Journal of Gansu Sciences, 4, 84–87.
Rademaker, A. (2018). Challenges for Information Extraction in the Oil and Gas Domain. In ONTOBRAS (pp. 11–25).
Raghupathi, W. , & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. Health information science and systems, 2(1), 3.

Rebholz-Schuhmann, D. , Oellrich, A. , Hoehndorf, R. (2012). Textmining solutions for biomedical research: enabling integrative biology. Nat Rev Genet, 13, 829–839.
Rifaie, M. , Kianmehr, K. , Alhajj, R. , & Ridley, M. J. (2008, July). Data warehouse architecture and design. In 2008 IEEE International Conference on Information Reuse and Integration (pp. 58–63). IEEE.
Rim, K. , Lynch, K. , & Pustejovsky, J. (2019, June). Computational Linguistics Applications for Multimedia Services. In Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (pp. 91–97).
Rong, J. , Michalska, S. , Subramani, S. , Du, J. , & Wang, H. (2019). Deep learning for pollen allergy surveillance from twitter in Australia. BMC medical informatics and decision making, 19(1), 208.
Sagiroglu, S. , & Sinanc, D. (2013, May). Big data: A review. In 2013 international conference on collaboration technologies and systems (CTS) (pp. 42–47). IEEE.
Sampathkumar, H. , Chen, X. W. , & Luo, B. (2014). Mining adverse drug reactions from online healthcare forums using hidden Markov model. BMC Medical Informatics and Decision Making, 14(1), 91.
Seoud, R. A. A. , & Mabrouk, M. S. (2013). TMT-HCC: A tool for text mining the biomedical literature for hepatocellular carcinoma (HCC) biomarkers identification. Computer Methods and Programs in Biomedicine, 112(3), 640–648.
Tao, D. , Yang, P. , & Feng, H. (2020). Utilization of text mining as a big data analysis tool for food science and nutrition. Comprehensive Reviews in Food Science and Food Safety, 19(2), 875–894.
Toh, K. A. , Yau, W. Y. , & Jiang, X. (2004). A reduced multivariate polynomial model for multimodal biometrics and classifiers fusion. IEEE Transactions on Circuits and Systems for Video Technology, 14(2), 224–233.
Tsuruoka, Y. , McNaught, J. , Tsujii, J. I. C. , & Ananiadou, S. (2007). Learning string similarity measures for gene/protein name dictionary look-up using logistic regression. Bioinformatics, 23(20), 2768–2774.
Uguz, H. (2011). A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. Knowledge-Based Systems, 24(7), 1024–1032.
Uramoto, N. , Matsuzawa, H. , Nagano, T. , Murakami, A. , Takeuchi, H. , & Takeda, K. (2004). A text-mining system for knowledge discovery from biomedical documents. IBM Systems Journal, 43(3), 516–533.
van Altena, A. J. , Moerland, P. D. , Zwinderman, A. H. , & Delgado Olabarriaga, S. (2019). Usage of the term Big Data in biomedical publications: A text mining approach. Big Data and Cognitive Computing, 3(1), 13.
Van Landeghem, S. , Björne, J. , Wei, C. H. , Hakala, K. , Pyysalo, S. , Ananiadou, S. , Kao, H. Y. , Lu, Z. , Salakoski, T. , Van de Peer, Y. , & Ginter, F. (2013). Large-scale event extraction from literature with multi-level gene normalization. PloS One, 8(4).
Wang, P. , Hao, T. , Yan, J. , & Jin, L. (2017). Large-scale extraction of drug-disease pairs from the medical literature. Journal of the Association for Information Science and Technology, 68(11), 2649–2661.
Wang, Y. , Li, Y. , Song, Y. , Rong, X. , & Zhang, S. (2017). Improvement of ID3 algorithm based on simplified information entropy and coordination degree. Algorithms, 10(4), 124.
Yang, H. , & Yang, C. C. (2013, September). Harnessing social media for drug-drug interactions detection. In 2013 IEEE International Conference on Healthcare Informatics (pp. 22–29). IEEE.
Yang, L. C. , Tan, I. K. , Selvaretnam, B. , Howg, E. K. , & Kar, L. H. (2019, May). TEXT: Traffic Entity eXtraction from Twitter. In Proceedings of the 2019 5th International Conference on Computing and Data Engineering (pp. 53–59).
Yang, Y. H. , Lin, Y. C. , Su, Y. F. , & Chen, H. H. (2007, July). Music emotion classification: A regression approach. In 2007 IEEE International Conference on Multimedia and Expo (pp. 208–211). IEEE.
Yates, A. , & Goharian, N. (2013, March). ADRTrace: detecting expected and unexpected adverse drug reactions from user reviews on social media sites. In European Conference on Information Retrieval (pp. 816–819). Springer, Berlin, Heidelberg.
Ye, Z. , Tafti, A. P. , He, K. Y. , Wang, K. , & He, M. M. (2016). Sparktext: Biomedical text mining on big data framework. PloS One, 11(9), e0162721.

Zhang, Y. F. , Xiong, Z. Y. , Geng, X. F. , & Chen, J. M. (2010). Analysis and improvement of ECLAT algorithm. Computer Engineering, 23, 28–30.

Zhu, F. , Patumcharoenpol, P. , Zhang, C. , Yang, Y. , Chan, J. , Meechai, A. , Vongsangnak, W. , & Shen, B. (2013). Biomedical text mining and its applications in cancer research. Journal of Biomedical Informatics, 46(2), 200–211.

Zweigenbaum, P. , Demner-Fushman, D. , Yu, H. , & Cohen, K. B. (2007). Frontiers of biomedical text mining: Current progress. Briefings in Bioinformatics, 8(5), 358–375.