



Machine Learning Using R

With Time Series and Industry-Based
Use Cases in R

—
Second Edition
—

Karthik Ramasubramanian
Abhishek Singh



Apress®

Machine Learning Using R

**With Time Series and Industry-Based
Use Cases in R**

Second Edition

Karthik Ramasubramanian

Abhishek Singh

Apress®

Machine Learning Using R: With Time Series and Industry-Based Use Cases in R

Karthik Ramasubramanian
New Delhi, Delhi, India

Abhishek Singh
New Delhi, Delhi, India

ISBN-13 (pbk): 978-1-4842-4214-8
<https://doi.org/10.1007/978-1-4842-4215-5>

ISBN-13 (electronic): 978-1-4842-4215-5

Library of Congress Control Number: 2018965407

Copyright © 2019 by Karthik Ramasubramanian and Abhishek Singh

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

Trademarked names, logos, and images may appear in this book. Rather than use a trademark symbol with every occurrence of a trademarked name, logo, or image we use the names, logos, and images only in an editorial fashion and to the benefit of the trademark owner, with no intention of infringement of the trademark.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Managing Director, Apress Media LLC: Welmoed Spahr
Acquisitions Editor: Celestin Suresh John
Development Editor: Matthew Moodie
Coordinating Editor: Aditee Mirashi

Cover designed by eStudioCalamar

Cover image designed by Freepik (www.freepik.com)

Distributed to the book trade worldwide by Springer Science+Business Media New York, 233 Spring Street, 6th Floor, New York, NY 10013. Phone 1-800-SPRINGER, fax (201) 348-4505, e-mail orders-ny@springer-sbm.com, or visit www.springeronline.com. Apress Media, LLC is a California LLC and the sole member (owner) is Springer Science + Business Media Finance Inc (SSBM Finance Inc). SSBM Finance Inc is a **Delaware** corporation.

For information on translations, please e-mail rights@apress.com, or visit <http://www.apress.com/rights-permissions>.

Apress titles may be purchased in bulk for academic, corporate, or promotional use. eBook versions and licenses are also available for most titles. For more information, reference our Print and eBook Bulk Sales web page at <http://www.apress.com/bulk-sales>.

Any source code or other supplementary material referenced by the author in this book is available to readers on GitHub via the book's product page, located at www.apress.com/978-1-4842-4214-8. For more detailed information, please visit <http://www.apress.com/source-code>.

Printed on acid-free paper

*To our parents for being the guiding light and
a strong pillar of support.
And to our long friendship.*

Table of Contents

About the Authors.....xvii

About the Technical Reviewerxix

Acknowledgmentsxxi

Introductionxxiii

Chapter 1: Introduction to Machine Learning and R..... 1

1.1 Understanding the Evolution 2

1.1.1 Statistical Learning..... 2

1.1.2 Machine Learning (ML)..... 3

1.1.3 Artificial Intelligence (AI)..... 4

1.1.4 Data Mining 5

1.1.5 Data Science 5

1.2 Probability and Statistics 7

1.2.1 Counting and Probability Definition 8

1.2.2 Events and Relationships 10

1.2.3 Randomness, Probability, and Distributions 13

1.2.4 Confidence Interval and Hypothesis Testing..... 15

1.3 Getting Started with R..... 21

1.3.1 Basic Building Blocks 21

1.3.2 Data Structures in R 22

1.3.3 Subsetting 24

1.3.4 Functions and the Apply Family..... 27

1.4 Machine Learning Process Flow 30

1.4.1 Plan..... 31

1.4.2 Explore..... 31

TABLE OF CONTENTS

- 1.4.3 Build 31
 - 1.4.4 Evaluate..... 32
- 1.5 Other Technologies..... 33
- 1.6 Summary..... 33
- Chapter 2: Data Preparation and Exploration 35**
 - 2.1 Planning the Gathering of Data 36
 - 2.1.1 Variables Types 36
 - 2.1.2 Data Formats 38
 - 2.1.3 Types of Data Sources 46
 - 2.2 Initial Data Analysis (IDA) 47
 - 2.2.1 Discerning a First Look..... 48
 - 2.2.2 Organizing Multiple Sources of Data into One..... 50
 - 2.2.3 Cleaning the Data 54
 - 2.2.4 Supplementing with More Information..... 58
 - 2.2.5 Reshaping..... 58
 - 2.3 Exploratory Data Analysis 60
 - 2.3.1 Summary Statistics 61
 - 2.3.2 Moment 65
 - 2.4 Case Study: Credit Card Fraud 72
 - 2.4.1 Data Import..... 72
 - 2.4.2 Data Transformation 73
 - 2.4.3 Data Exploration 74
 - 2.5 Summary..... 77
- Chapter 3: Sampling and Resampling Techniques 79**
 - 3.1 Introduction to Sampling..... 80
 - 3.2 Sampling Terminology..... 81
 - 3.2.1 Sample 82
 - 3.2.2 Sampling Distribution 82
 - 3.2.3 Population Mean and Variance 82
 - 3.2.4 Sample Mean and Variance 83

3.2.5 Pooled Mean and Variance	83
3.2.6 Sample Point	84
3.2.7 Sampling Error.....	84
3.2.8 Sampling Fraction	84
3.2.9 Sampling Bias.....	85
3.2.10 Sampling Without Replacement (SWOR)	85
3.2.11 Sampling with Replacement (SWR).....	85
3.3 Credit Card Fraud: Population Statistics	86
3.4 Data Description	86
3.5 Population Mean	88
3.6 Population Variance	88
3.7 Pooled Mean and Variance.....	88
3.8 Business Implications of Sampling.....	93
3.8.1 Shortcomings of Sampling	94
3.9 Probability and Non-Probability Sampling	94
3.9.1 Types of Non-Probability Sampling.....	95
3.10 Statistical Theory on Sampling Distributions	96
3.10.1 Law of Large Numbers: LLN	97
3.10.2 Central Limit Theorem	100
3.11 Probability Sampling Techniques	106
3.11.1 Population Statistics.....	106
3.11.2 Simple Random Sampling	110
3.11.3 Systematic Random Sampling	118
3.11.4 Stratified Random Sampling.....	123
3.11.5 Cluster Sampling	131
3.11.6 Bootstrap Sampling	139
3.12 Monte Carlo Method: Acceptance-Rejection Method	147
3.13 Summary.....	150

Chapter 4: Data Visualization in R 151

4.1 Introduction to the ggplot2 Package 152

4.2 World Development Indicators 153

4.3 Line Chart..... 153

4.4 Stacked Column Charts..... 160

4.5 Scatterplots..... 167

4.6 Boxplots 168

4.7 Histograms and Density Plots 172

4.8 Pie Charts..... 177

4.9 Correlation Plots..... 180

4.10 Heatmaps..... 182

4.11 Bubble Charts 184

4.12 Waterfall Charts 189

4.13 Dendrogram 192

4.14 Wordclouds 195

4.15 Sankey Plots 197

4.16 Time Series Graphs 198

4.17 Cohort Diagrams 201

4.18 Spatial Maps 203

4.19 Summary..... 208

Chapter 5: Feature Engineering 211

5.1 Introduction to Feature Engineering 212

5.2 Understanding the Data 213

5.2.1 Data Summary..... 215

5.2.2 Properties of Dependent Variable 215

5.2.3 Features Availability: Continuous or Categorical 219

5.2.4 Setting Up Data Assumptions 221

5.3 Feature Ranking..... 221

5.4 Variable Subset Selection	226
5.4.1 Filter Method	227
5.4.2 Wrapper Methods	231
5.4.3 Embedded Methods.....	240
5.5 Principal Component Analysis.....	245
5.6 Summary.....	251
Chapter 6: Machine Learning Theory and Practice	253
6.1 Machine Learning Types	256
6.1.1 Supervised Learning.....	257
6.1.2 Unsupervised Learning.....	257
6.1.3 Semi-Supervised Learning	257
6.1.4 Reinforcement Learning	258
6.2 Groups of Machine Learning Algorithms.....	258
6.3 Real-World Datasets	264
6.3.1 House Sale Prices.....	264
6.3.2 Purchase Preference	265
6.3.3 Twitter Feeds and Article	266
6.3.4 Breast Cancer	266
6.3.5 Market Basket	267
6.3.6 Amazon Food Reviews.....	268
6.4 Regression Analysis	268
6.5 Correlation Analysis	271
6.5.1 Linear Regression.....	273
6.5.2 Simple Linear Regression.....	275
6.5.3 Multiple Linear Regression.....	279
6.5.4 Model Diagnostics: Linear Regression	283
6.5.5 Polynomial Regression	297
6.5.6 Logistic Regression	302
6.5.7 Logit Transformation.....	303
6.5.8 Odds Ratio	304

TABLE OF CONTENTS

6.5.9 Model Diagnostics: Logistic Regression	313
6.5.10 Multinomial Logistic Regression	326
6.5.11 Generalized Linear Models	330
6.5.12 Conclusion	332
6.6 Support Vector Machine SVM.....	332
6.6.1 Linear SVM	334
6.6.2 Binary SVM Classifier	335
6.6.3 Multi-Class SVM	338
6.6.4 Conclusion	340
6.7 Decision Trees	340
6.7.1 Types of Decision Trees	342
6.7.2 Decision Measures	344
6.7.3 Decision Tree Learning Methods	346
6.7.4 Ensemble Trees	367
6.7.5 Conclusion	376
6.8 The Naive Bayes Method.....	376
6.8.1 Conditional Probability.....	376
6.8.2 Bayes Theorem	377
6.8.3 Prior Probability	377
6.8.4 Posterior Probability	378
6.8.5 Likelihood and Marginal Likelihood.....	378
6.8.6 Naïve Bayes Methods	378
6.8.7 Conclusion	385
6.9 Cluster Analysis.....	385
6.9.1 Introduction to Clustering	386
6.9.2 Clustering Algorithms	387
6.9.3 Internal Evaluation.....	401
6.9.4 External Evaluation.....	403
6.9.5 Conclusion	405

6.10 Association Rule Mining	405
6.10.1 Introduction to Association Concepts	406
6.10.2 Rule-Mining Algorithms	408
6.10.3 Recommendation Algorithms	417
6.10.4 Conclusion	426
6.11 Artificial Neural Networks	426
6.11.1 Human Cognitive Learning	427
6.11.2 Perceptron	429
6.11.3 Sigmoid Neuron	432
6.11.4 Neural Network Architecture	433
6.11.5 Supervised versus Unsupervised Neural Nets	435
6.11.6 Neural Network Learning Algorithms	436
6.11.7 Feed-Forward Back-Propagation	439
6.11.8 Conclusion	447
6.12 Text-Mining Approaches	448
6.12.1 Introduction to Text Mining	449
6.12.2 Text Summarization	451
6.12.3 TF-IDF	453
6.12.4 Part-of-Speech (POS) Tagging	455
6.12.5 Word Cloud	460
6.12.6 Text Analysis: Microsoft Cognitive Services	462
6.12.7 Conclusion	473
6.13 Online Machine Learning Algorithms	473
6.13.1 Fuzzy C-Means Clustering	475
6.13.2 Conclusion	479
6.14 Model Building Checklist	479
6.15 Summary	481

Chapter 7: Machine Learning Model Evaluation 483

7.1 Dataset..... 484

7.1.1 House Sale Prices..... 484

7.1.2 Purchase Preference 487

7.2 Introduction to Model Performance and Evaluation..... 489

7.3 Objectives of Model Performance Evaluation 491

7.4 Population Stability Index 492

7.5 Model Evaluation for Continuous Output..... 498

7.5.1 Mean Absolute Error 500

7.5.2 Root Mean Square Error 503

7.5.3 R-Square 504

7.6 Model Evaluation for Discrete Output 508

7.6.1 Classification Matrix 509

7.6.2 Sensitivity and Specificity 515

7.6.3 Area Under ROC Curve..... 517

7.7 Probabilistic Techniques 520

7.7.1 K-Fold Cross-Validation..... 521

7.7.2 Bootstrap Sampling 524

7.8 The Kappa Error Metric 525

7.9 Summary..... 529

Chapter 8: Model Performance Improvement..... 533

8.1 Overview of the Caret Package..... 535

8.2 Introduction to Hyper-Parameters..... 537

8.3 Hyper-Parameter Optimization..... 541

8.3.1 Manual Search..... 543

8.3.2 Manual Grid Search 545

8.3.3 Automatic Grid Search..... 548

8.3.4 Optimal Search 550

8.3.5 Random Search 553

8.3.6 Custom Searching 555

8.4 The Bias and Variance Tradeoff	559
8.5 Introduction to Ensemble Learning	564
8.5.1 Voting Ensembles	565
8.5.2 Advanced Methods in Ensemble Learning	567
8.6 Ensemble Techniques Illustration in R	570
8.6.1 Bagging Trees	571
8.6.2 Gradient Boosting with a Decision Tree	573
8.6.3 Blending KNN and Rpart	578
8.6.4 Stacking Using caretEnsemble	580
8.7 Advanced Topic: Bayesian Optimization of Machine Learning Models	586
8.8 Summary	592
Chapter 9: Time Series Modeling	595
9.1 Components of Time Series	596
9.2 Test of Stationarity	600
9.3 ACF and AR Model	604
9.4 PACF and MA Model	608
9.5 ARIMA Model	612
9.5.1 Box-Jenkins Approach	613
9.6 Linear Regression with AR Errors	621
9.7 Summary	626
Chapter 10: Scalable Machine Learning and Related Technologies	629
10.1 Distributed Processing and Storage	630
10.1.1 Google File System (GFS)	631
10.1.2 MapReduce	632
10.1.3 Parallel Execution in R	633
10.2 The Hadoop Ecosystem	638
10.2.1 MapReduce	639
10.2.2 Hive	644
10.2.3 Apache Pig	648

TABLE OF CONTENTS

10.2.4 HBase	652
10.2.5 Spark	654
10.3 Machine Learning in R with Spark.....	655
10.3.1 Setting the Environment Variable	656
10.3.2 Initializing the Spark Session	656
10.3.3 Loading Data and the Running Preprocess	657
10.3.4 Creating SparkDataFrame	658
10.3.5 Building the ML Model.....	659
10.3.6 Predicting the Test Data.....	660
10.3.7 Stopping the SparkR Session	661
10.4 Machine Learning in R with H2O.....	661
10.4.1 Installation of Packages	663
10.4.2 Initialization of H2O Clusters	664
10.5 Summary.....	665
Chapter 11: Deep Learning Using Keras and TensorFlow	667
11.1 Introduction to Deep Learning	668
11.2 Deep Learning Architectures.....	669
11.2.1 Convolutional Neural Networks (CNN)	669
11.2.2 Recurrent Neural Networks (RNN).....	670
11.2.3 Generative Adversarial Network (GAN)	672
11.3 Deep Learning Toolset.....	674
11.3.1 High-Level Library	674
11.3.2 Backend Engine or Frameworks.....	674
11.3.3 Hardware Capability	675
11.3.4 Programming Language Choice	675
11.3.5 Cloud Infrastructure.....	675
11.4 Use Case: Identify Duplicate Questions in Quora	676
11.4.1 Environment Setup	676
11.4.2 Data Preprocessing	676
11.4.3 Benchmark Model	678

11.4.4 Siamese Recurrent Architectures	680
11.4.5 The Keras Model.....	683
11.4.6 The Model Summary.....	683
11.4.7 The Validation Sample	684
11.4.8 Train the Model.....	684
11.4.9 Save the Model.....	685
11.4.10 Model Performance	686
11.4.11 Make Predictions.....	687
11.4.12 Example Predictions	687
11.5 Summary.....	688
Index.....	689

About the Authors



Karthik Ramasubramanian has over seven years of practice and leading data science and business analytics in retail, FMCG, eCommerce, information technology, and the hospitality industry with multi-national companies and unicorn startups. Karthik is a researcher and problem solver with a diverse set of experiences in the data science lifecycle, starting from a data problem discovery to creating data science PoCs and products for various industry use cases.

In his leadership roles, he has been instrumental in solving many RoI driven business problems through data science solutions. He has mentored and trained hundreds of professionals and students around the world through various online platforms and university engagement programs in data science.

On the descriptive side of data science, he has designed, developed, and spearheaded many A/B experiment frameworks for improving product features, conceptualized funnel analysis for understanding user interactions and identifying the friction points within a product, and designed statistically robust metrics. On the predictive side, he has developed intelligent chatbots based on deep learning models that understand human-like interactions, customer segmentation models, recommendation systems, and many Natural Language Processing models.

His current areas of interest include ROI-driven data product development, advanced machine learning algorithms, data product frameworks, Internet of Things (IoT), scalable data platforms, and model deployment frameworks.

Karthik Completed his M.Sc. in Theoretical Computer Science from PSG College of Technology, Coimbatore (Affiliated to Anna University, Chennai), where he pioneered the application of machine learning, data mining, and fuzzy logic in his research work on computer and network security.

ABOUT THE AUTHORS



Abhishek Singh is on a mission to profess the de facto language of this millennium, the numbers. He is on a journey to bring machines closer to humans, for a better and more beautiful world by generating opportunities with artificial intelligence and machine learning. He leads a team of data science professionals who are solving pressing problems in food security, cyber security, natural disasters, healthcare, and many more areas, all with the help of data and technology. Abhishek is in the process of bringing smart IoT devices to smaller cities in India so people can leverage technology for the betterment of life.

He has worked with colleagues from many parts of the United States, Europe, and Asia, and strives to work with more people from various backgrounds. In a span of six years at big corporations, he has stress-tested the assets of U.S. banks, solved insurance pricing models, and made telecom experiences easier for customers. He is now creating data science opportunities with his team of young minds.

He actively participates in analytics-related thought leadership, authoring, public speaking, meetups, and training in data science. He is a staunch supporter of responsible use of AI to remove biases and fair use for a better society.

Abhishek completed his MBA from IIM Bangalore, a B.Tech. in Mathematics and Computing from IIT Guwahati, and a PG Diploma in Cyber Law from NALSAR University, Hyderabad.

About the Technical Reviewer



Taweh Beysolow II is a data scientist and author currently based in San Francisco, California. He has a Bachelor's of Science degree in economics from St. Johns University and a Master of Science in applied statistics from Fordham University. His professional experience has included working at Booz Allen Hamilton as a consultant and in various startups as a data scientist, specifically focusing on machine learning. He has applied machine learning to the Federal Consulting, Financial Services, and Agricultural sectors.

Acknowledgments

We are grateful to our teachers, open source communities, and colleagues for enriching us with the knowledge and confidence to write the first edition of this book. Thanks to all our readers. You have made the second edition of the book possible. The knowledge in this book is an accumulation of several years of research work and professional experience gained at our alma mater and industry. We are grateful to Prof R. Nadarajan and Prof R. Anitha, Department of Applied Mathematics and Computational Sciences, PSG College of Technology, Coimbatore, for their continued support and encouragement for our efforts in the field of data science.

In the rapidly changing world, the field of machine learning is evolving very fast and most of the latest developments are driven by the open source platform. We thank all the developers and contributors across the globe who are freely sharing their knowledge. We also want to thank our colleagues our our past and current companies—Snapdeal, Deloitte, Hike, Prudential, Probyto, and Mahindra & Mahindr—for providing opportunities to experiment and create cutting-edge data science solutions.

Karthik especially would like to thank his father, Mr. S Ramasubramanian, for always being a source of inspiration in his life. He is immensely thankful to his supervisor, Mr. Nikhil Dwarakanath, director of the data science team at Snapdeal, for creating the opportunities to bring about the best analytics professional in him and providing the motivation to take up challenging projects.

Abhishek would like to thank his father, Mr. Charan Singh, a senior scientist in the India meteorological department, for introducing him to the power of data in weather forecasting in his formative years. On a personal front, Abhishek would like to thank his mother Jaya, sister Asweta, and brother Avilash, for their continued moral support.

We want to thank our publisher Apress, specifically Celestine, for proving us with this opportunity, Sanchita Prachi for managing the first edition of the book, and Aditee Mirashi for the second edition, Poonam and Piyush for their reviews, and everybody involved in the production team.

—Karthik Ramasubramanian

—Abhishek Singh

Introduction

In the second edition of *Machine Learning Using R*, we added a new chapter on time series modeling (Chapter 9), a traditional topic that has its genesis from statistics. The second newly added chapter is deep learning (Chapter 11), which is fast emerging as a sub-field of machine learning. Apart from these two new chapters, the overall presentation of text and code in the book is put out in a new reader-friendly format.

The new edition continues to focus on building the use cases using R, a popular statistical programming language. For topics like deep learning, it might be advised to adopt Python with frameworks like TensorFlow. However, in this new edition, we will show you how to use the R programming language with TensorFlow, hence avoiding the effort of learning Python if you are only comfortable with R.

Like in the first edition, we have kept the fine balance of theory and application of machine learning through various real-world use cases, which give the readers a truly comprehensive collection of topics in machine learning in one volume.

What you'll learn:

- Understand machine learning algorithms using R
- Master a machine learning model building a process flow
- Theoretical foundations of machine learning algorithms
- Industry focused real-world use cases
- Time series modeling in R
- Deep learning using Keras and TensorFlow in R

Who This Book is For

This book is for data scientists, data science professionals, and researchers in academia who want to understand the nuances of machine learning approaches/algorithms in practice using R. The book will also benefit readers who want to understand the technology behind implementing a scalable machine learning model using Apache Hadoop, Hive, Pig, and Spark.

INTRODUCTION

This book is a comprehensive guide for anybody who wants to understand the machine learning model building process from end to end, including:

- Practical demonstration of concepts in R
- Machine learning models using Apache Hadoop and Spark
- Time series analysis
- Introduction to deep learning models using Keras and TensorFlow using R

CHAPTER 1

Introduction to Machine Learning and R

Beginners to machine learning are often confused by the plethora of algorithms and techniques being taught in subjects like statistical learning, data mining, artificial intelligence, soft computing, and data science. It's natural to wonder how these subjects are different from one another and which is the best for solving real-world problems. There is substantial overlap in these subjects and it's hard to draw a clear Venn diagram explaining the differences. Primarily, the foundation for these subjects is derived from probability and statistics. However, many statisticians probably won't agree with machine learning giving life to statistics, giving rise to the never-ending chicken and egg conundrum kind of discussions. Fundamentally, without spending much effort in understanding the pros and cons of this discussion, it's wise to believe that the power of statistics needed a pipeline to flow across different industries with some challenging problems to be solved and machine learning simply established that high-speed and frictionless pipeline. The other subjects that evolved from statistics and machine learning are simply trying to broaden the scope of these two subjects and putting it into a bigger banner.

Except for statistical learning, which is generally offered by mathematics or statistics departments in the majority of the universities across the globe, the rest of these subjects—like machine learning, data mining, artificial intelligence, and soft computing—are taught by computer science department.

In the recent years, this separation is disappearing but the collaboration between the two departments is still not complete. Programmers are intimidated by the complex theorems and proofs and statisticians hate *talking* (read as *coding*) to machines all the time. But as more industries are becoming data- and product-driven, the need for getting the two departments to speak a common language is strongly emphasized. Roles in industry are suitably revamped to create openings like machine learning engineers, data engineers, and data scientists into a broad group being called the *data science team*.

The purpose of this chapter is to take one step back and demystify the terminologies as we travel through the history of machine learning and emphasize that putting the ideas from statistics and machine learning into practice by broadening the scope is critical.

At the same time, we elaborate on the importance of learning the fundamentals of machine learning with an approach inspired by the contemporary techniques from data science. We have simplified all the mathematics to as much extent as possible without compromising the fundamentals and core part of the subject. The right balance of statistics and computer science is always required for understanding machine learning, and we have made every effort for our readers to appreciate the elegance of mathematics, which at times is perceived by many to be hard and full of convoluted definitions, theories, and formulas.

1.1 Understanding the Evolution

The first challenge anybody finds when starting to understand how to build intelligent machines is how to mimic human behavior in many ways or, to put it even more appropriately, how to do things even better and more efficiently than humans. Some examples of these things performed by machines are identifying spam emails, predicting customer churn, classifying documents into respective categories, playing chess, participating in jeopardy, cleaning house, playing football, and much more. Carefully looking at these examples will reveal that humans haven't perfected these tasks to date and rely heavily on machines to help them. So, now the question remains, where do you start learning to build such intelligent machines? Often, depending on which task you want to take up, experts will point you to machine learning, artificial intelligence (AI), or many such subjects, that sound different by name but are intrinsically connected.

In this chapter, we have taken up the task to knit together this evolution and finally put forth the point that machine learning, which is the first block in this evolution, is where you should fundamentally start to later delve deeper into other subjects.

1.1.1 Statistical Learning

The whitepaper, *Discovery with Data: Leveraging Statistics with Computer Science to Transform Science and Society* by American Statistical Association (ASA) [1], published in July 2014, defines *statistics* as “the science of learning from data, and of measuring, controlling, and communicating uncertainty is the most mature of the data sciences”.

This discipline has been an essential part of the social, natural, bio-medical, and physical sciences, engineering, and business analytics, among others. Statistical thinking not only helps make scientific discoveries, but it quantifies the reliability, reproducibility, and general uncertainty associated with these discoveries. This excerpt from the whitepaper is very precise and powerful in describing the importance of statistics in data analysis.

Tom Mitchell, in his article, “The Discipline of Machine Learning [2],” appropriately points out, “Over the past 50 years, the study of machine learning has grown from the efforts of a handful of computer engineers exploring whether computers could learn to play games, and a field of statistics that largely ignored computational considerations, to a broad discipline that has produced fundamental statistical-computational theories of learning processes.”

This learning process has found its application in a variety of tasks for commercial and profitable systems like computer vision, robotics, speech recognition, and many more. At large, it’s when statistics and computational theories are fused together that machine learning emerges as a new discipline.

1.1.2 Machine Learning (ML)

The Samuel Checkers-Playing Program, which is known to be the first computer program that could learn, was developed in 1959 by Arthur Lee Samuel, one of the fathers of machine learning. Followed by Samuel, *Ryszard S. Michalski*, also deemed a father of machine learning, came out with a system for recognizing handwritten alphanumeric characters, working along with Jacek Karpinski in 1962-1970. The subject from then has evolved with many facets and led the way for various applications impacting businesses and society for the good.

Tom Mitchell defined the fundamental question machine learning seeks to answer as, “How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?” He further explains that the defining question of computer science is, “How can we build machines that solve problems, and which problems are inherently tractable/intractable?”, whereas statistics focus on answering “What can be inferred from data plus a set of modeling assumptions, with what reliability?”

This set of questions clearly shows the difference between statistics and machine learning. As mentioned earlier in the chapter, it might not even be necessary to deal with the chicken and egg conundrum, as we clearly see that one simply complements the other and is paving the path for the future. As we dive deep into the concepts of

statistics and machine learning, you will see the differences clearly emerging or at times completely disappearing. Another line of thought, in the paper “Statistical Modeling: The Two Cultures” by Leo Breiman in 2001 [3], argued that statisticians rely too heavily on data modeling, and that machine learning techniques are instead focusing on the predictive accuracy of models.

1.1.3 Artificial Intelligence (AI)

The AI world from very beginning was intrigued by games. Whether it be *checkers*, *chess*, *Jeopardy*, or the recently very popular *Go*, the AI world strives to build machines that can play against humans to beat them in these games and it has received much accolades for the same. IBM’s Watson beat the two best players of Jeopardy, a quiz game show wherein participants compete to come out with their responses as a phrase in the form of questions to some general knowledge clues in the form of answers. Considering the complexity in analyzing natural language phrases in these answers, it was considered to be very hard for machines to compete with humans. A high-level architecture of IBM’s DeepQA used in Watson looks something like in Figure 1-1.

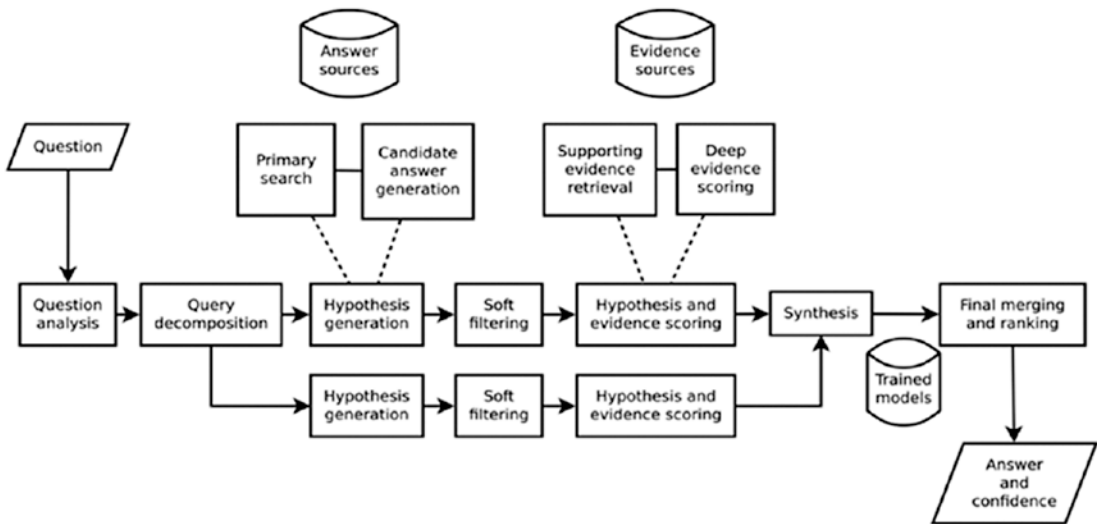


Figure 1-1. Architecture of IBM’s DeepQA

AI also sits at the core of robotics. The 1971 Turing Award winner, John McCarthy, a well known American computer scientist, was believed to have coined this term and in his article titled, “What Is Artificial Intelligence?” he defined it as “the science and

engineering of making intelligent machines [4]”. So, if you relate back to what we said about machine learning, we instantly sense a connection between the two, but AI goes the extra mile to congregate a number of sciences and professions, including linguistics, philosophy, psychology, neuroscience, mathematics, and computer science, as well as other specialized fields such as artificial psychology. It should also be pointed out that machine learning is often considered to be a subset of AI.

1.1.4 Data Mining

Knowledge Discovery and Data Mining (KDD), a premier forum for data mining, states its goal to be advancement, education, and adoption of the “science” for knowledge discovery and data mining. Data mining, like ML and AI, has emerged as a interdisciplinary subfield of computer science and for this reason, KDD commonly projects data mining methods, as the intersection of AI, ML, statistics, and database systems. Data mining techniques were integrated into many database systems and business intelligence tools, when adoption of analytic services were starting to explode in many industries.

The research paper, “WEKA Experiences with a Java open-source project”[5] (WEKA is one of the widely adapted tools for doing research and projects using data mining), published in the *Journal of Machine Learning Research*, talked about how the classic book, *Data Mining: Practical Machine Learning Tools and Techniques with Java*, [6] was originally named just *Practical Machine Learning*, and the term data mining was only added for marketing reasons. Eibe Frank and Mark A. Hall, who wrote this research paper, are the two coauthors of the book, so we have a strong rationale to believe this reason for the name change. Once again, we see fundamentally, ML being at the core of data mining.

1.1.5 Data Science

It’s not wrong to call data science a big umbrella that brought everything with a potential to show insight from data and build intelligent systems inside it. In the book, *Data Science for Business* [7], Foster Provost and Tom Fawcett introduced the notion of viewing data and data science capability as a strategic asset, which will help businesses think explicitly about the extent to which one should invest in them. In a way, data science has emphasized the importance of data more than the algorithms of learning.

It has established a well defined process flow that says, first think about doing descriptive data analysis and then later start to think about modeling. As a result of this, businesses have started to adopt this new methodology because they were able to

relate to it. Another incredible change data science has brought is around creating the synergies between various departments within a company. Every department has its own subject matter experts and data science teams have started to build their expertise in using data as a common language to communicate. This paradigm shift has witnessed the emergence of data-driven growth and many data products. Data science has given us a framework, which aims to create a conglomerate of skillsets, tools, and technologies. Drew Conway, the famous American data scientist who is known for his Venn diagram definition of data science as shown in Figure 1-2, has very rightly placed machine learning in the intersection of Hacking Skills and Math & Statistics Knowledge.

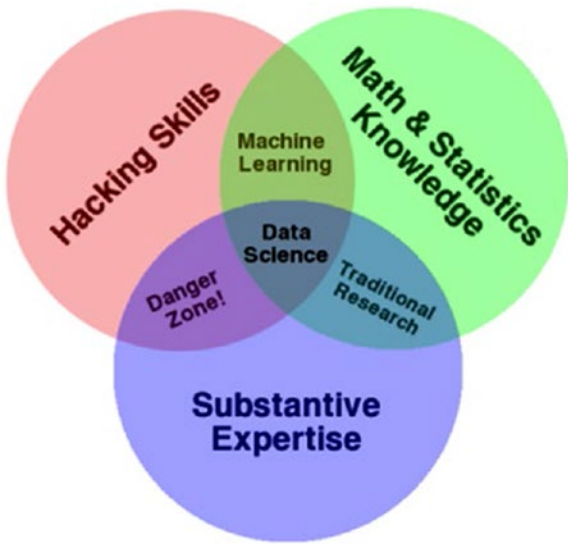


Figure 1-2. Venn diagram definition of data science

We strongly believe the fundamentals of these different fields of study are all derived from statistics and machine learning but different flavors, for reasons justifiable in its own context, were given to it, which helped the subject be molded into various systems and areas of research. This book will help trim down the number of different terminologies being used to describe the same set of algorithms and tools. It will present a simple-to-understand and coherent approach, the algorithms in machine learning and its practical use with R. Wherever it's appropriate, we will emphasize the need to go outside the scope of this book and guide our readers with the relevant materials. By doing so, we are re-emphasizing the need for mastering traditional approaches in machine learning and, at the same time, staying abreast with the latest development in tools and technologies in this space.

Our design of topics in this book are strongly influenced by data science framework but instead of wandering through the vast pool of tools and techniques you would find in the world of data science, we have kept our focus strictly on teaching practical ways of applying machine learning algorithms with R.

The rest of this chapter is organized to help readers understand the elements of probability and statistics and programming skills in R. Both of these will form the foundations for understanding and putting machine learning into practical use. The chapter ends with a discussion of technologies that apply ML to a real-world problem. Also, a generic machine learning process flow will be presented showing how to connect the dots, starting from a given problem statement to deploying ML models to working with real-world systems.

1.2 Probability and Statistics

Common sense and gut instincts play a key role for policymakers, leaders, and entrepreneurs in building nations and large enterprises. The question is, how do we change some intractable qualitative decision making into objectively understood quantitative decision making? That's where probability and statistics come in. Much of statistics is focused on analyzing existing data and drawing suitable conclusions using probability models. Though it's very common to use probabilities in many statistical modeling, we feel it's important to identify the different questions probability and statistics help us answer. An example from the book, *Learning Statistics with R: A Tutorial for Psychology Students and Other Beginners* by Daniel Navarro [8], University of Adelaide, helps us understand it much better. Consider these two pairs of questions:

1. What are the chances of a fair coin coming up heads 10 times in a row?
2. If my friend flips a coin 10 times and gets 10 heads. Is she playing a trick on me?

and

1. How likely it is that five cards drawn from a perfectly shuffled deck will all be hearts?
2. If five cards off the top of the deck are all hearts, how likely is it that the deck was shuffled?

In case of the coin toss, the first question could be answered if we know the coin is fair, there's a 50% chance that any individual coin flip will come up heads, in probability notation, $P(\text{heads}) = 0.5$. So, our probability is $P(\text{heads 10 times in a row}) = .0009765625$ (since all the 10 coin tosses are independent of each other, we can simply compute $(0.5)^{10}$ to arrive at this value). The probability value .0009765625 quantifies the chances of a fair coin coming up heads 10 times in a row.

On the other side, such a small probability would mean the occurrence of the event (heads 10 times in a row) is very rare, which helps to *infer* that my friend is playing some trick on me when she got all heads. Think about this—does tossing a coin 10 times give you strong evidence for doubting your friend? Maybe no; you may ask her to repeat the process several times. The more the data we generate, the better will be the inference. The second set of questions has the same thought process but is applied to a different problem. We encourage you to perform the calculations yourself to answer the question.

So, fundamentally, probability could be used as a tool in statistics to help us answer many such real-world questions using a model. We will explore some basics of both these worlds, and it will become evident that both converge at a point where it's hard to observe many differences between the two.

1.2.1 Counting and Probability Definition

Imagine we are conducting an experiment with coin flips, in which we will flip three coins eight times each. Each combination of heads and tails constitutes a unique outcome. For example, HHH is a unique outcome. The possible outcomes are the following: (HHH, HHT, HTH, HTT, THH, THT, TTH, and TTT). Figure 1-3 shows a basic illustration of this experiment, with three coins, a total of eight possible outcomes (HHH, HHT, HTH, HTT, THH, THT, TTH, and TTT) are present. This set is called the *sample space*.

