It’s All Analytics!

The Foundations of AI, Big Data, and Data Science Landscape for Professionals in Healthcare, Business, and Government

First edition published 2020
by CRC Press
6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742

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

Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

ISBN: 978-0-367-35968-3 (hbk)
ISBN: 978-0-367-49379-0 (pbk)
ISBN: 978-0-429-34398-8 (ebk)

Typeset in ITC Garamond STD
by Deanta Global Publishing Services, Chennai, India

Contents

Foreword Number One

The applications of computational methods in machine learning and artificial intelligence are rapidly changing the world that we work and live in. Many traditional industries and professions are being fundamentally reimagined as AI industries. It is becoming imperative for those at every level in companies and organizations (not to mention the general public) to understand both “what will AI do FOR me?” and “what will AI do TO me?”.

The rapid acceleration in the development and deployment of these technologies is creating an increasing gap in understanding. Many who need to know don’t even know what they don’t know. This, coupled with hyperbolic news releases on some new AI application-of-the-moment, leaves the nontechnical observer with no easy solution to bridging this gap.

Fortunately, Scott Burk and Gary Miner have astutely recognized this gap in understanding and offer a starting point for bridging this gap in *It&$3pos;s All Analytics!* This volume provides a “20,000 foot overview” of these technologies and serves as an easily-grasped read for beginning the journey to deeper understanding or broadening one’s knowledge base. While it is geared towards those with little or no understanding of AI and machine learning, it is a valuable resource for those working in these areas who may have a siloed view of the fields.

The authors are uniquely qualified to deliver this overview as they are both not only industry practitioners of these technologies, but also educators skilled at making these topics accessible to the neophyte. They have obviously paid great attention to readability and organized the material in a way that provides a memorable framework for pinning the reader&$3pos;s newly gathered knowledge. Additionally, the book is richly referenced with additional resources for taking a deeper dive into specific subject matter.

I&$3pos;d like to be among the first to congratulate the authors on this timely, engaging, useful, and highly informative read.

John W. Cromwell, M.D., FACS, FASCRS
Associate Chief Medical Officer | Director of Surgical Quality and Safety University of Iowa Hospitals & Clinics
Director, Division of Gastrointestinal, Minimally Invasive, and Bariatric Surgery Clinical Associate Professor
University of Iowa Carver College of Medicine Faculty, Interdisciplinary Graduate Program in Informatics
University of Iowa Graduate College Iowa City, Iowa

Foreword Number Two

Written with focus on the underlying concept of the entire series of books

This book seeks to reduce the “sea of terms” in Data Science to a systematic terminology to describe general aspects of AI and Data Science. This system of terms will permit multiple stakeholders in an organization to speak the same “language” across the enterprise. This common language will permit close integration between analytics and those functions in the organization that precede analytics (e.g. database design and management) and those deployment functions that follow it (e.g. marketing campaigns).

This book is not a complete expression of the subject, yet it is comprehensive in terms of the scope of each part without being comprehensive in detail. This book is not designed to show the learner how to build an analytical model with a given tool. Rather, it shows learners how to organize their thinking about the plethora of terms and concepts in Data Science to provide sufficient insight to permit practitioners to do it properly later on. Many books are available to show precise sequences of steps with a given tool for building analytical models

This book is Part I of a three-part series, each one of which will be covered in a separate book. Part II will describe: (1) the design and management of the database functions necessary to support analytics properly; (2) The structure of the analytics process (e.g. the CRISP-DM analytics process model), and; (3) The general structure of a modeling application (e.g. the major steps in building a predictive analytics model). Part III will present a survey of analytics applications in business and industry. Neither of these books should be considered as a stand-alone resource in the overall process of applying analytics in an organization. All 3 parts should be combined to compose the design of any analytics application in an organization. Development of models without the proper deployment system may relegate the models to the “shelf”, because they can’t be deployed efficiently in the organization.

Each book (and indeed each chapter) is written to be independent of information in previous chapters, as far as possible. This degree of independence is facilitated by repeat of relevant terms and principles introduced in previous chapters, which are required to understand the information in the current chapter. As such, these chapters are learning “objects”, which are more or less self-contained. These learning objects must be combined together to form a seamless solution during execution.

Both this book and the entire series present a “layered” approach to learning the practice of Data Science. Each layer is designed to be as functionally independent as possible, yet easily related to previous layers and subsequent layers. This layered learning approach follows the Layered Learning Practice Model (LLPM) shown to be very effective in training learners to provide specific clinical or patient services in the practice of Oncology.

Robert Nisbet, Ph.D.
Goleta, California

Foreword Number Three

Almost 30 years ago I began using the term "Information Democracy" to describe a world where everyone has timely, relevant, and actionable insights to carry out the tasks associated with their role – and align them with the overarching strategy of the organization.

Since that time, we've made some progress, but not nearly as much as we would have hoped. In fact, based on our most recent research, a majority of organizations report that less than half of their users have such access.

Further to this point, only 43% indicate that they consistently use data in their decision-making process. And, only a third of organizations claim high or extremely high data literacy.

Further clouding things is the constant barrage of technologies, techniques, and buzzwords that bombard us each day – including artificial intelligence, data science, machine learning, IoT, edge computing, etc.

The only way that we can make real progress is through education about the importance and value of business intelligence and analytics, increasing data literacy and establishing a solid understanding of all relevant approaches.

To that end, Drs. Burk and Miner have created what is an excellent addition to the growing body of work available on the subject. In contrast to the many volumes on the subject, their approach has made many of these topics readily accessible to the novice or manager seeking a basic understanding as well as to the data science professional seeking a well-organized reference.

The future of an information democracy is highly dependent upon knowledgeable management and skilled users. Accordingly, as organizations strive to increase their data literacy, and leverage data science, this book should be required reading.

Howard Dresner
Chief Research Officer
Dresner Advisory Services
www.dresneradvisory.com

Preface

The Basis for This Book and the Series

The authors have been collaborating on the ideas in this book for years. We met while working together in the software and technology sector. Each of us has backgrounds in statistics, machine learning, analytics, healthcare, and business. We noticed that while the products and solutions in the artificial intelligence (AI), data science and analytics space offered tremendous value, there was a great deal of misunderstanding of the terminology in and around this space. Furthermore, we noticed a lot of "reinvention" of methods that existed for years, meaning that in this "reinvention" many of the same concepts acquired "new names," thus adding to the confusion, especially when attempting to communicate among different disciplines.

This book will dive into many different domains: AI, machine learning, visual business intelligence (BI), analytics, and more. For brevity, when we are not explicitly writing about a particular domain (AI, data science), we choose to use the term "analytics" as a broad and general term for the overarching domain. In the end, you shall see, *It's All Analytics!* We have witnessed companies of all sizes and in multiple industries gain tremendous value from applying analytic methods and technologies. We know there are companies of all shapes and sizes that are beginning their journey. We know many others that have deep roots, but in an applied area and cannot see the "big picture" and know of technologies outside their immediate application area. We know there are those that struggle with a constantly changing sea of terms and technology. We know some are scared of what they don't know, what they are not doing, what they should be doing. Everyone is moving at a million miles an hour and companies are worried they might lose their advantage in a competitive market and need to do something quickly with analytics or expand with some nascent technology. *However, it is time to take a step back: to survey the landscape and synthesize it*. With a pause, we can view the analytics domain holistically.

Even after years of conversations and numerous emails, it was difficult for us to arrive at a title. First, how do you cover a sea of changing terminology, what is hot what is not – even if it is the same thing? Second, we wanted a wide applicability to a variety of industries. Some are at differing levels of maturity and adoption. Thus, the name for this book is a combination of three of the most commonly used domains and the application space where these methods have the most opportunity to advance society.

Additionally, we determined that our ambitions were beyond the scope of a single book. Therefore we are breaking our goals into three books. ***The first***, which you are reading, is a book with a goal of synthesizing disciplines across many educational disciplines and practiced fields at an executive or professional level. ***The second book*** will present design considerations for your analytics architecture across the continuum of your company's program goals – organization, data architecture, analytics architecture. ***The third book*** will provide examples of applications across a very broad variety of business and policy goals.

Professionals Need This Book

Why do you need this book? No other book has a comprehensive view of the landscape. Yes, you can get a lot of information from the Internet, but it is not curated or validated. Is it someone's blog post who is trying to promote themselves as an expert? Anyone can write a blog and there are many analytics websites which are platforms for marketecture (promotion of product or service over capability).

Gartner is an independent analysis firm that reports on the technology sector. They phrased one need for this book in several comments on their website, including the need for members of an organization to speak the same language. Kasey Panetta (Panetta 2019) stated this clearly: "[Our need is to]... champion data literacy and teach data as a second language to enable data-driven business." She continued by stating,

> Imagine an organization where the marketing department speaks French, the product designers speak German, the analytics team speaks Spanish and no one speaks a second language... That's essentially how a data-driven business functions when there is no data literacy.

She points out that this year (2020), half of all organizations will lack data literacy skills that are needed to achieve business value.

And Valerie Logan, Senior Director Analyst, Gartner, points out another important fact (see Panetta 2019), that the

> prevalence of data and analytics capabilities, including artificial intelligence, requires creators and consumers to "speak data" as a common language.... Data and analytics leaders must champion workforce data literacy as an enabler of digital business and treat information as a second language.

We will help you speak the language in this book. In fact, we will cover the dialects of AI, machine learning, analytics, data science, and statistics and show you what they have in common and what separates them.

What This Book Is and Is Not; Who Should Not Be Reading This Book

This is not an academic book. We don't talk about hyper-parameters, the Stone-Weierstrass theorem or stochastic processes. It is not a scholarly pursuit; it does not include the same rigor as an academic endeavor. However, we will provide many references as well as "Resources for the Avid Learner" throughout the chapters. This book should serve as an introductory reference book and is meant to get you thinking about the broader scope of disciplines beyond what is available in the hype cycle.

It is not meant to be a daily practitioner's step-by-step practice guide. However, it may be very useful to expose a practitioner to a wider view and provide ideas of alternative methods to their problem set. As any carpenter, artist or do-it-yourselfer knows, the right tool makes all the difference in the world.

Therefore, knowing what kinds of problems typically are solved by statistical inference versus machine learning is very helpful.

It does not discuss low-level, nuts-and-bolts recipes for producing machine learning algorithms, applying algorithms, or determining which statistical test should be used for a particular set of data. It does not talk about specific commercial software vendors, although there is a brief mention of open source platforms for illustration.

It is not complete. It would be far too big a task to write, or for that matter read, a complete compendium of over a hundred years of thought and advancement in a wide field. Our goal is to provide a book that helps make sense of this wide field, where there is uniqueness, where there is overlap, and where there is opportunity.

This is not a book on a specific area; it is not an AI book, a statistics book, or a BI book. It is a compilation of disciplines and technologies and contrasts across those disciplines. It is not a recipe book for how to write code or where to find an open source library for a specific machine learning algorithm. It does not tell you when to apply a specific statistical test of inference or the dos and don'ts of creating a BI dashboard. **It is meant to be a synthesis book (a 101 or survey course on data-driven methods if you will).** It encompasses an evolution and rebirth of fields and technologies and demonstrates how they are interrelated and how they are independently and inter-reliably useful. It sets you up for success in a career in analytics by seeing the big picture. In addition, the series will provide a larger view of analytics across the industry, how to design and operate an enterprise for success in analytics.

This book is not a history book. We do provide some historical elements to provide some interesting detail and context, but these are not intended to be comprehensive in nature. There are some good books on the history of statistics, mathematics, and AI as well as other fields. Note: most of these are written for an academic audience or an audience that desires a deep historical perspective.

How This Book Is Organized

This book contains 12 chapters. This first chapter sets the stage and outlines more specific needs and purposes of the book. Chapter 2 provides business justifications and design recommendations for creating a successful analytics program within your organization. It also explains the hype cycle, what you should pay attention to and what to avoid. Chapter 3 describes the heart and soul of data, processes as well as other fundamentals – models, algorithms, and a standard process for analytics projects. In Chapter 4 we look at "analytics" – far and wide from methodology to application. The rest of the book dives more deeply in the foundations of analytics – BI, machine learning, AI, data science, big data, statistics, and more.

One of our key objectives is to make this book as valuable to the reader as possible. One way we want to provide that value is to make it a resource that can serve as a reference where the reader can pick it up and read a freestanding chapter on its own without having to pick up at the beginning. To make this amenable, we will repeat ourselves; this was a conscious choice that we hope will not be too much of an inconvenience, but a benefit when you revisit us here.

It should be noted that we attempt to offer some consistency where it makes sense. As an example, in Chapters 1 to 3, after the preamble we begin with a section called "The Hip, the Hype, the Fears, the Intrigue, and the Reality." Our goal with this section is to present the reader with what they might find in "the hip," social media, blogs, user channels... or the "hype" from marketers, sellers, excitement journalists in press releases... and then the reality as we see it. We are attempting to help you distinguish the signal from the noise; we cover this concept in Chapter 3.

We also attempt to outline the objectives for the chapter up front in the "Key Words," "Preamble" and "Introduction" sections so that you know what to expect from the chapter. This is intended to support the goal of being a reference book down the line. Chapters conclude with postscripts, references, and "Resources for the Avid Learner" sections.

We want this book to be a pleasure to read as well as to provide valuable information. While this is not a "hands-on" or methods book, we do want you to mentally practice the concepts presented here. We offer many examples and thought exercises in gray boxes to stimulate your reading and improve attention. The book should flow without reading every gray box, but these gray boxes should make for reading that is more enjoyable.

In summary, we have noticed a changing sea of terminology. We know that much of this is generated by companies, technologists and consultants. We recognize software and technology have improved greatly over time, but much has just been relabeled. Without changing names of techniques, it is difficult to sound innovative and fresh. Marketers, product managers and consultants are pushed to be more innovative, to differentiate, and to forge new horizons. We understand. However, we want to help professionals in the field understand some of the underlying foundations, some of the relationships and some ways to think about the subject. Leaders and practitioners are confused about what is relevant and what is not. Fear of missing out (FOMO) is causing panic and mis-investment. We hope to clarify some of these misunderstandings in this book series.

Your Authors,
Scott Burk, Ph.D. Temple, Texas
Gary D. Miner, Ph.D., Tulsa, Oklahoma and Rome, Georgia

Reference

Panetta, Kasey. February 6, 2019. "A data and analytics leader's guide to data literacy," Gartner, www.gartner.com/smarterwithgartner/a-data-and-analytics-leaders-guide-to-data-literacy/.

Endorsements

Almost every company in the world now understands the critical importance of collecting, processing, analyzing, and acting upon data. The largest hurdles impeding companies in this process aren&$3pos;t caused by technical limitations or a lack of trained specialists, but by the people who need to understand how it affects them, what can be done, and how to implement and manage it within their organization, but don&$3pos;t. In this book, Burk and Miner help to solve that problem in language that is straightforward, sensible, and based on their considerable experience. If actionable analytics is a key need for your organization, and you want to minimize the struggle and confusion required to implement it, you should read this book.

Dylan Zwick
Former Director of Data Science, Overstock.com

Burk and Miner have created a map to guide anxious and overwhelmed executives through the rapidly changing and often unwieldy landscape of data and analytics techniques and technologies. Their survey cuts through the hype and hyperbole and enables data practitioners and non-practitioners to clearly communicate how to understand, optimize, and ultimately transform their business processes through analytics. Highly recommended.

Josh Wills
Former Director of Data Engineering, Slack

It's All Analytics! deserves a prominent place on executives' bookshelves. Burk and Miner have undertaken a noteworthy challenge in their synthesis of data science, machine learning, data mining, artificial intelligence, and statistics, presented at a level both useful and provocative to business leaders. The chapter on statistics particularly fills a gap in current discourse about the latest fashions in AI and Machine Learning.

Loren Williams
Former Chief Data Scientist, Big Four

The rise of artificial intelligence brings us excitement and hope, but also causes some anxiety and even fear. The internet is flooded with a sea of terminology and concepts. For anyone who is interested in learning more about AI, numerous online courses, articles, blogs are at finger tips. However, not all information has been curated, thus resulting in a tremendous amount of confusion and a great deal of misunderstanding.

I am thrilled that Scott and Gary compact several decades of history of AI, data science, analytics, an incredible amount of terminology, concepts, and a comprehensive view of the current landscape, all into this one book. With their years of experience across a broad spectrum of industry, the book offers many practical examples and thought exercises, and explains complex concepts in simple language.

Business executives will benefit from this book with in-depth understanding of the technical concept and capability, as well as organizational planning and strategy; people leaders will get help to build a strong team with the right talents, tooling, and capability; technical professionals will broaden their view of the data science world and have a clearer expectation of career path.

Doctoral Fellowship. During the doctoral study years, he also studied mammalian genetics at The Jackson Laboratory, Bar Harbor, ME, under a College Training Program on an NIH award; and another College Training Program at the Bermuda Biological Station, St. George's West, Bermuda in a marine developmental embryology course, on an NSF award; and a third college training program held at the University of California, San Diego at the Molecular Techniques in Developmental Biology Institute, again on an NSF award.

Following that he studied as a post-doctoral student at the University of Minnesota in Behavioral Genetics, where, along with research in schizophrenia and Alzheimer's disease (AD), he learned "how to write books" from assisting in editing two book manuscripts of his mentor, Irving Gottesman, Ph.D. (Dr. Gottesman returned the favor 41 years later by writing two tutorials for the book *Practical Text Mining*). After academic research and teaching positions, Miner did another two-year NIH post-doctoral in psychiatric epidemiology and biostatistics at the University of Iowa, where he became thoroughly immersed in studying affective disorders and Alzheimer's disease. Altogether, he spent over 30 years researching and writing papers and books on the genetics of Alzheimer's disease (Miner, G.D., Richter, R, Blass, J.P., Valentine, J.L, and Winters-Miner, Linda. *Familial Alzheimer's Disease: Molecular Genetics and Clinical Perspectives*. Dekker: New York, 1989; and Miner, G.D., Winters-Miner, Linda, Blass, J.P., Richter, R, and Valentine, J.L. *Caring for Alzheimer's Patients: A Guide for Family & Healthcare Providers*. Plenum Press Insight Books: New York, 1989).

Over the years he has held positions, including professor and chairman of a department, at various universities including the University of Kansas, the University of Minnesota, Northwest Nazarene University, Eastern Nazarene University, Southern Nazarene University, Oral Roberts University Medical School, where he was Associate Professor of Pharmacology and Director of the Alzheimer Disease & Geriatric Disorders Research Laboratories and even for a period of time in the 1990s was a visiting clinical professor of psychology for geriatrics at the Fuller Graduate School of Psychology & Fuller Theological Seminary in Pasadena, CA.

In 1985 he and his wife, Dr. Linda Winters-Miner (author of several tutorials in this book) founded The Familial Alzheimer's Disease Research Foundation (aka "The Alzheimer's Foundation"), which became a leading force in organizing both local and international scientific meetings and thus bringing together all the leaders in the field of genetics of AD from several countries, which then lead to the writing of the first scientific book on the genetics of Alzheimer's disease; this book included papers by over 100 scientists coming out of the First International Symposium on the Genetics of Alzheimer's Disease held in Tulsa, OK in October 1987. During part of this time, he was also an affiliate research scientist with the Oklahoma Medical Research Foundation located in Oklahoma City with the University of Oklahoma School of Medicine.

Miner was influential in bringing all of the world's leading scientists working on genetics of AD together at just the right time when various laboratories from Harvard to Duke University and the University of California-San Diego, to the University of Heidelberg, in Germany, and universities in Belgium, France, England and Perth, Australia were beginning to find "genes" which they thought were related to Alzheimer's disease.

During the 1990s Dr. Miner was appointed to the Oklahoma Governor's Task Force on Alzheimer's Disease, and also was Associate Editor for Alzheimer's Disease for *The Journal of Geriatric Psychiatry & Neurology*, which he still serves on to this day. By 1995 most of these dominantly inherited genes for AD had been discovered, and the one that Miner had been working on since the mid-1980s with the University of Washington in Seattle was the last of these initial 5 to be identified, this gene on Chromosome 1 of the human genome. At that time, having met the goal of finding out some of the genetics of AD, Miner decided to do something different, to find an area of the business world, and since he had been analyzing data for over 30 years, working for StatSoft, Inc. as a senior statistician and data mining consultant, which seemed a perfect "semi-retirement" career. Interestingly (as his wife had predicted), he discovered that the "business world" was much more fun than the "academic world," and at a KDD-Data Mining meeting in 1999 in San Francisco, he decided that he would specialize in "data mining." Incidentally, he first met

Index