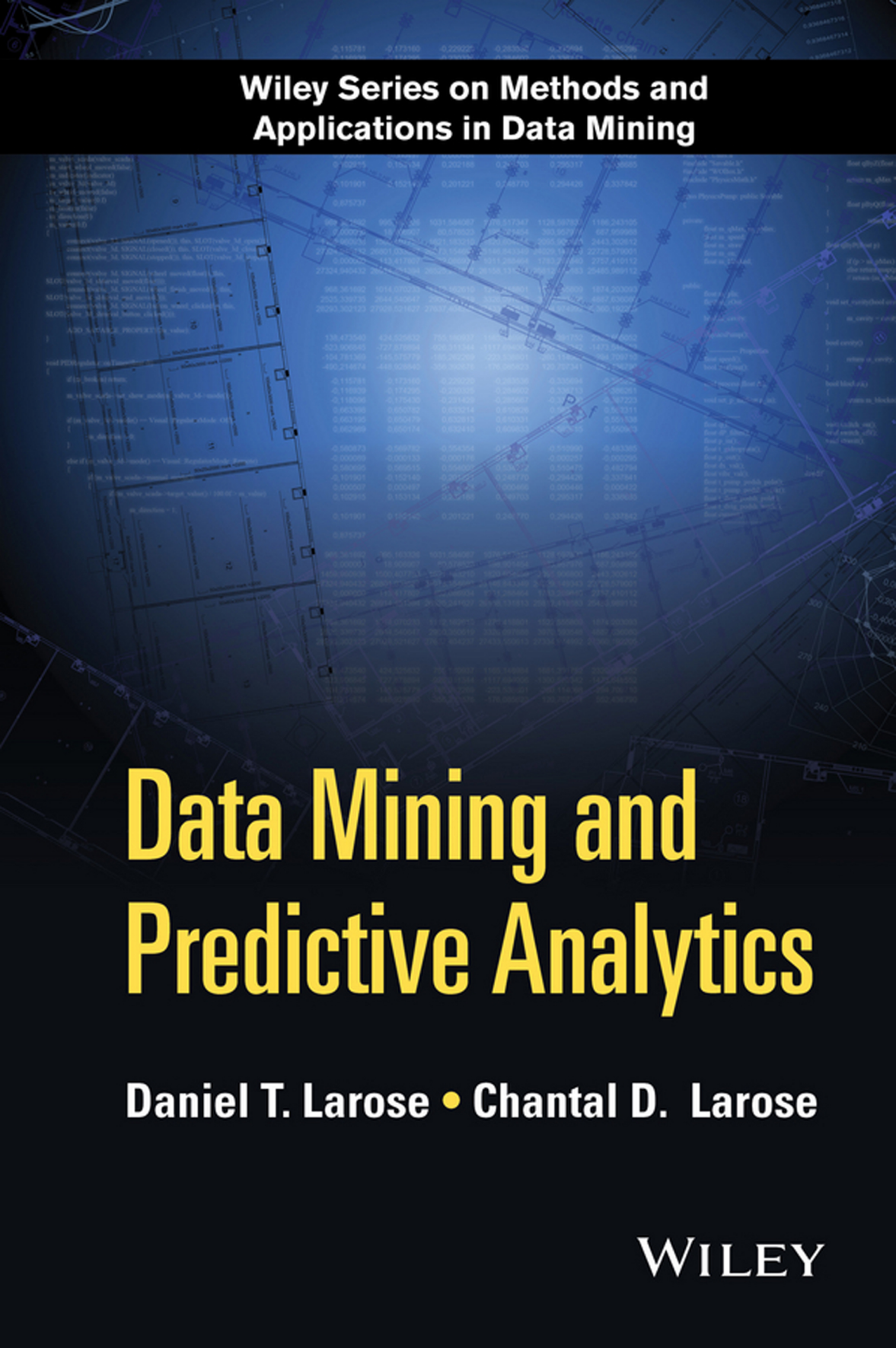

Wiley Series on Methods and Applications in Data Mining
Data Mining and Predictive Analytics
Daniel T. Larose • Chantal D. Larose
WILEY

DATA MINING AND PREDICTIVE ANALYTICS

WILEY SERIES ON METHODS AND APPLICATIONS IN DATA MINING

Series Editor: **Daniel T. Larose**

Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition • Daniel T. Larose and Chantal D. Larose

Data Mining for Genomics and Proteomics: Analysis of Gene and Protein Expression Data • Darius M. Dziuda

Knowledge Discovery with Support Vector Machines • Lutz Hamel

Data-Mining on the Web: Uncovering Patterns in Web Content, Structure, and Usage • Zdravko Markov and Daniel T. Larose

Data Mining Methods and Models • Daniel T. Larose

Practical Text Mining with Perl • Roger Bilisoly

Data Mining and Predictive Analytics • Daniel T. Larose and Chantal D. Larose

DATA MINING AND PREDICTIVE ANALYTICS

Second Edition

DANIEL T. LAROSE
CHANTAL D. LAROSE

WILEY

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

Library of Congress Cataloging-in-Publication Data:

Larose, Daniel T.
Data mining and predictive analytics / Daniel T. Larose, Chantal D. Larose.
pages cm. – (Wiley series on methods and applications in data mining)
Includes bibliographical references and index.
ISBN 978-1-118-11619-7 (cloth)
1. Data mining. 2. Prediction theory. I. Larose, Chantal D. II. Title.
QA76.9.D343L3776 2015
006.3′12–dc23
2014043340

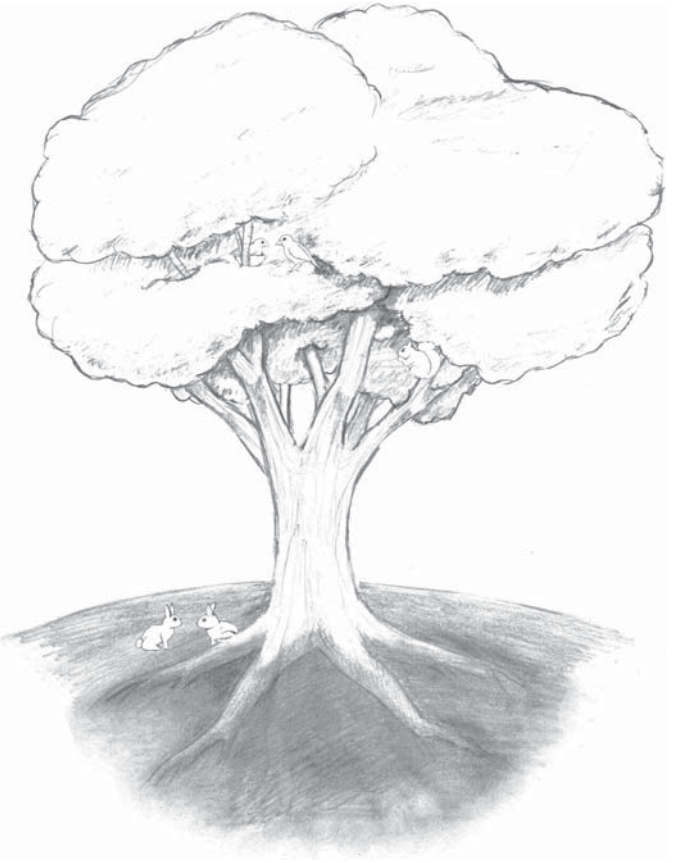
Set in 10/12pt Times by Laserwords Private Limited, Chennai, India

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

2 2015

To those who have gone before us,
And to those who come after us,
In the Family Tree of Life …



CONTENTS

PREFACE

WHAT IS DATA MINING? WHAT IS PREDICTIVE ANALYTICS?

Data mining is the process of discovering useful patterns and trends in large data sets.

Predictive analytics is the process of extracting information from large data sets in order to make predictions and estimates about future outcomes.

Data Mining and Predictive Analytics, by Daniel Larose and Chantal Larose, will enable you to become an expert in these cutting-edge, profitable fields.

WHY IS THIS BOOK NEEDED?

According to the research firm MarketsandMarkets, the global big data market is expected to grow by 26% per year from 2013 to 2018, from $14.87 billion in 2013 to $46.34 billion in 2018.[1] Corporations and institutions worldwide are learning to apply data mining and predictive analytics, in order to increase profits. Companies that do not apply these methods will be left behind in the global competition of the twenty-first-century economy.

Humans are inundated with data in most fields. Unfortunately, most of this valuable data, which cost firms millions to collect and collate, are languishing in warehouses and repositories. *The problem is that there are not enough trained human analysts available who are skilled at translating all of this data into knowledge*, and thence up the taxonomy tree into wisdom. This is why this book is needed.

The McKinsey Global Institute reports[2]:

> There will be a shortage of talent necessary for organizations to take advantage of big data. A significant constraint on realizing value from big data will be a shortage of talent, particularly of people with deep expertise in statistics and machine learning, and the

[1] *Big Data Market to Reach $46.34 Billion by 2018*, by Darryl K. Taft, *eWeek*, www.eweek.com/database/big-data-market-to-reach-46.34-billion-by-2018.html, posted September 1, 2013, last accessed March 23, 2014.

[2] *Big data: The next frontier for innovation, competition, and productivity*, by James Manyika *et al*., Mckinsey Global Institute, www.mckinsey.com, May, 2011. Last accessed March 16, 2014.

> managers and analysts who know how to operate companies by using insights from big data . . . . We project that demand for deep analytical positions in a big data world could exceed the supply being produced on current trends by 140,000 to 190,000 positions. … In addition, we project a need for 1.5 million additional managers and analysts in the United States who can ask the right questions and consume the results of the analysis of big data effectively.

This book is an attempt to help alleviate this critical shortage of data analysts.

Data mining is becoming more widespread every day, because it empowers companies to uncover profitable patterns and trends from their existing databases. Companies and institutions have spent millions of dollars to collect gigabytes and terabytes of data, but are not taking advantage of the valuable and actionable information hidden deep within their data repositories. However, as the practice of data mining becomes more widespread, companies that do not apply these techniques are in danger of falling behind, and losing market share, because their competitors are applying data mining, and thereby gaining the competitive edge.

WHO WILL BENEFIT FROM THIS BOOK?

In *Data Mining and Predictive Analytics*, the step-by-step hands-on solutions of real-world business problems using widely available data mining techniques applied to real-world data sets will appeal to managers, CIOs, CEOs, CFOs, data analysts, database analysts, and others who need to keep abreast of the latest methods for enhancing return on investment.

Using *Data Mining and Predictive Analytics*, you will learn what types of analysis will uncover the most profitable nuggets of knowledge from the data, while avoiding the potential pitfalls that may cost your company millions of dollars. *You will* learn *data mining and predictive analytics* by doing *data mining and predictive analytics*.

DANGER! DATA MINING IS EASY TO DO BADLY

The growth of new off-the-shelf software platforms for performing data mining has kindled a new kind of danger. The ease with which these applications can manipulate data, combined with the power of the formidable data mining algorithms embedded in the black-box software, make their misuse proportionally more hazardous.

In short, *data mining is easy to do badly*. A little knowledge is especially dangerous when it comes to applying powerful models based on huge data sets. For example, analyses carried out on unpreprocessed data can lead to erroneous conclusions, or inappropriate analysis may be applied to data sets that call for a completely different approach, or models may be derived that are built on wholly unwarranted specious assumptions. If deployed, these errors in analysis can lead to very expensive failures. *Data Mining and Predictive Analytics* will help make you a savvy analyst, who will avoid these costly pitfalls.

"WHITE-BOX" APPROACH

Understanding the Underlying Algorithmic and Model Structures

The best way to avoid costly errors stemming from a blind black-box approach to data mining and predictive analytics is to instead apply a "white-box" methodology, which emphasizes an understanding of the algorithmic and statistical model structures underlying the software.

Data Mining and Predictive Analytics applies this white-box approach by

- clearly explaining *why* a particular method or algorithm is needed;
- getting the reader acquainted with *how* a method or algorithm works, using a toy example (tiny data set), so that the reader may follow the logic step by step, and thus gain a *white-box insight* into the inner workings of the method or algorithm;
- providing an application of the method to a large, real-world data set;
- using exercises to test the reader's level of understanding of the concepts and algorithms;
- providing an opportunity for the reader to experience doing some real data mining on large data sets.

ALGORITHM WALK-THROUGHS

Data Mining Methods and Models walks the reader through the operations and nuances of the various algorithms, using small data sets, so that the reader gets a true appreciation of what is really going on inside the algorithm. For example, in Chapter 21, we follow step by step as the balanced iterative reducing and clustering using hierarchies (BIRCH) algorithm works through a tiny data set, showing precisely how BIRCH chooses the optimal clustering solution for this data, from start to finish. As far as we know, such a demonstration is unique to this book for the BIRCH algorithm. Also, in Chapter 27, we proceed step by step to find the optimal solution using the selection, crossover, and mutation operators, using a tiny data set, so that the reader may better understand the underlying processes.

Applications of the Algorithms and Models to Large Data Sets

Data Mining and Predictive Analytics provides examples of the application of data analytic methods on actual large data sets. For example, in Chapter 9, we analytically unlock the relationship between nutrition rating and cereal content using a real-world data set. In Chapter 4, we apply principal components analysis to real-world census data about California. All data sets are available from the book series web site: www.dataminingconsultant.com.

Chapter Exercises: Checking to Make Sure You Understand It

Data Mining and Predictive Analytics includes over 750 chapter exercises, which allow readers to assess their depth of understanding of the material, as well as have a little fun playing with numbers and data. These include *Clarifying the Concept* exercises, which help to clarify some of the more challenging concepts in data mining, and *Working with the Data* exercises, which challenge the reader to apply the particular data mining algorithm to a small data set, and, step by step, to arrive at a computationally sound solution. For example, in Chapter 14, readers are asked to find the *maximum a posteriori* classification for the data set and network provided in the chapter.

Hands-On Analysis: Learn Data Mining by Doing Data Mining

Most chapters provide the reader with *Hands-On Analysis* problems, representing an opportunity for the reader to apply his or her newly acquired data mining expertise to solving real problems using large data sets. Many people learn by doing. *Data Mining and Predictive Analytics* provides a framework where the reader can learn data mining by doing data mining. For example, in Chapter 13, readers are challenged to approach a real-world credit approval classification data set, and construct their best possible logistic regression model, using the methods learned in this chapter as possible, providing strong interpretive support for the model, including explanations of derived variables and indicator variables.

EXCITING NEW TOPICS

Data Mining and Predictive Analytics contains many exciting new topics, including the following:

- Cost-benefit analysis using data-driven misclassification costs.
- Cost-benefit analysis for trinary and k-nary classification models.
- Graphical evaluation of classification models.
- BIRCH clustering.
- Segmentation models.
- Ensemble methods: Bagging and boosting.
- Model voting and propensity averaging.
- Imputation of missing data.

THE R ZONE

R is a powerful, open-source language for exploring and analyzing data sets (www.r-project.org). Analysts using *R* can take advantage of many freely available packages, routines, and graphical user interfaces to tackle most data analysis

problems. In most chapters of this book, the reader will find *The R Zone*, which provides the actual *R* code needed to obtain the results shown in the chapter, along with screenshots of some of the output.

APPENDIX: DATA SUMMARIZATION AND VISUALIZATION

Some readers may be a bit rusty on some statistical and graphical concepts, usually encountered in an introductory statistics course. *Data Mining and Predictive Analytics* contains an appendix that provides a review of the most common concepts and terminology helpful for readers to hit the ground running in their understanding of the material in this book.

THE CASE STUDY: BRINGING IT ALL TOGETHER

Data Mining and Predictive Analytics culminates in a detailed Case Study. Here the reader has the opportunity to see how everything he or she has learned is brought all together to create actionable and profitable solutions. This detailed Case Study ranges over four chapters, and is as follows:

- Chapter 29: *Case Study, Part 1: Business Understanding, Data Preparation, and EDA*
- Chapter 30: *Case Study, Part 2: Clustering and Principal Components Analysis*
- Chapter 31: *Case Study, Part 3: Modeling and Evaluation for Performance and Interpretability*
- Chapter 32: *Case Study, Part 4: Modeling and Evaluation for High Performance Only*

The Case Study includes dozens of pages of graphical, exploratory data analysis (EDA), predictive modeling, customer profiling, and offers different solutions, depending on the requisites of the client. The models are evaluated using a custom-built data-driven cost-benefit table, reflecting the true costs of classification errors, rather than the usual methods such as overall error rate. Thus, the analyst can compare models using the estimated profit per customer contacted, and can predict how much money the models will earn, based on the number of customers contacted.

HOW THE BOOK IS STRUCTURED

Data Mining and Predictive Analytics is structured in a way that the reader will hopefully find logical and straightforward. There are 32 chapters, divided into eight major parts.

- Part 1, *Data Preparation*, consists of chapters on data preparation, EDA, and dimension reduction.

- Part 2, *Statistical Analysis*, provides classical statistical approaches to data analysis, including chapters on univariate and multivariate statistical analysis, simple and multiple linear regression, preparing to model the data, and model building.
- Part 3, *Classification*, contains nine chapters, making it the largest section of the book. Chapters include *k*-nearest neighbor, decision trees, neural networks, logistic regression, naïve Bayes, Bayesian networks, model evaluation techniques, cost-benefit analysis using data-driven misclassification costs, trinary and *k*-nary classification models, and graphical evaluation of classification models.
- Part 4, *Clustering*, contains chapters on hierarchical clustering, *k*-means clustering, Kohonen networks clustering, BIRCH clustering, and measuring cluster goodness.
- Part 5, *Association Rules*, consists of a single chapter covering a priori association rules and generalized rule induction.
- Part 6, *Enhancing Model Performance*, provides chapters on segmentation models, ensemble methods: bagging and boosting, model voting, and propensity averaging.
- Part 7, *Further Methods in Predictive Modeling*, contains a chapter on imputation of missing data, along with a chapter on genetic algorithms.
- Part 8, *Case Study: Predicting Response to Direct-Mail Marketing*, consists of four chapters presenting a start-to-finish detailed Case Study of how to generate the greatest profit from a direct-mail marketing campaign.

THE SOFTWARE

The software used in this book includes the following:

- *IBM SPSS Modeler* data mining software suite
- *R* open source statistical software
- *SAS Enterprise Miner*
- *SPSS* statistical software
- *Minitab* statistical software
- *WEKA* open source data mining software.

IBM SPSS Modeler (www-01.ibm.com/software/analytics/spss/products/modeler/) is one of the most widely used data mining software suites, and is distributed by *SPSS*, whose base software is also used in this book. *SAS Enterprise Miner* is probably more powerful than *Modeler*, but the learning curve is also steeper. *SPSS* is available for download on a trial basis as well (Google "spss" download). *Minitab* is an easy-to-use statistical software package that is available for download on a trial basis from their web site at www.minitab.com.

WEKA: THE OPEN-SOURCE ALTERNATIVE

The Weka (Waikato Environment for Knowledge Analysis) machine learning workbench is open-source software issued under the GNU General Public License, which includes a collection of tools for completing many data mining tasks. *Data Mining and Predictive Modeling* presents several hands-on, step-by-step tutorial examples using Weka 3.6, along with input files available from the book's companion web site www.dataminingconsultant.com. The reader is shown how to carry out the following types of analysis, using WEKA: Logistic Regression (Chapter 13), Naïve Bayes classification (Chapter 14), Bayesian Networks classification (Chapter 14), and Genetic Algorithms (Chapter 27). For more information regarding Weka, see www.cs.waikato.ac.nz/ml/weka/. The author is deeply grateful to James Steck for providing these WEKA examples and exercises. James Steck (james_steck@comcast.net) was one of the first students to complete the master of science in data mining from Central Connecticut State University in 2005 (GPA 4.0), and received the first data mining Graduate Academic Award. James lives with his wife and son in Issaquah, WA.

THE COMPANION WEB SITE: WWW.DATAMININGCONSULTANT.COM

The reader will find supporting materials, both for this book and for the other data mining books written by Daniel Larose and Chantal Larose for *Wiley InterScience*, at the companion web site, www.dataminingconsultant.com. There one may download the many data sets used in the book, so that the reader may develop a hands-on feel for the analytic methods and models encountered throughout the book. Errata are also available, as is a comprehensive set of data mining resources, including links to data sets, data mining groups, and research papers.

However, the real power of the companion web site is available to faculty adopters of the textbook, who will have access to the following resources:

- Solutions to all the exercises, including the hands-on analyses.
- PowerPoint® presentations of each chapter, ready for deployment in the classroom.
- Sample data mining course projects, written by the author for use in his own courses, and ready to be adapted for your course.
- Real-world data sets, to be used with the course projects.
- Multiple-choice chapter quizzes.
- Chapter-by-chapter web resources.

Adopters may e-mail Daniel Larose at larosed@ccsu.edu to request access information for the adopters' resources.

DATA MINING AND PREDICTIVE ANALYTICS AS A TEXTBOOK

Data Mining and Predictive Analytics naturally fits the role of textbook for a one-semester course or two-semester sequences of courses in introductory and intermediate data mining. Instructors may appreciate

- the presentation of data mining as a *process*;
- the "white-box" approach, emphasizing an understanding of the underlying algorithmic structures;
  - — Algorithm walk-throughs with toy data sets
  - — Application of the algorithms to large real-world data sets
  - — Over 300 figures and over 275 tables
  - — Over 750 chapter exercises and hands-on analysis
- the many exciting new topics, such as cost-benefit analysis using data-driven misclassification costs;
- the detailed *Case Study*, bringing together many of the lessons learned from the earlier 28 chapters;
- the Appendix: Data Summarization and Visualization, containing a review of statistical and graphical concepts readers may be a bit rusty on;
- the companion web site, providing the array of resources for adopters detailed above.

Data Mining and Predictive Analytics is appropriate for advanced undergraduate- or graduate-level courses. An introductory statistics course would be nice, but is not required. No computer programming or database expertise is required.