

LEARNING MADE EASY



Blockchain Data Analytics

for
dummies[®]
A Wiley Brand

Discover how blockchains
are disrupting the data world

Build models that classify,
predict, and analyze data

Use analytics models to
solve business problems

Michael G. Solomon, PhD

Blockchain Data Analytics

for
dummies[®]
A Wiley Brand



Blockchain Data Analytics

by Michael Solomon

for
dummies[®]
A Wiley Brand

Blockchain Data Analytics For Dummies®

Published by: **John Wiley & Sons, Inc.**, 111 River Street, Hoboken, NJ 07030-5774, www.wiley.com

Copyright © 2020 by John Wiley & Sons, Inc., Hoboken, New Jersey

Published simultaneously in Canada

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without the prior written permission of the Publisher. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permissions>.

Trademarks: Wiley, For Dummies, the Dummies Man logo, Dummies.com, Making Everything Easier, and related trade dress are trademarks or registered trademarks of John Wiley & Sons, Inc. and may not be used without written permission. All other trademarks are the property of their respective owners. John Wiley & Sons, Inc. is not associated with any product or vendor mentioned in this book.

LIMIT OF LIABILITY/DISCLAIMER OF WARRANTY: THE PUBLISHER AND THE AUTHOR MAKE NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE ACCURACY OR COMPLETENESS OF THE CONTENTS OF THIS WORK AND SPECIFICALLY DISCLAIM ALL WARRANTIES, INCLUDING WITHOUT LIMITATION WARRANTIES OF FITNESS FOR A PARTICULAR PURPOSE. NO WARRANTY MAY BE CREATED OR EXTENDED BY SALES OR PROMOTIONAL MATERIALS. THE ADVICE AND STRATEGIES CONTAINED HEREIN MAY NOT BE SUITABLE FOR EVERY SITUATION. THIS WORK IS SOLD WITH THE UNDERSTANDING THAT THE PUBLISHER IS NOT ENGAGED IN RENDERING LEGAL, ACCOUNTING, OR OTHER PROFESSIONAL SERVICES. IF PROFESSIONAL ASSISTANCE IS REQUIRED, THE SERVICES OF A COMPETENT PROFESSIONAL PERSON SHOULD BE SOUGHT. NEITHER THE PUBLISHER NOR THE AUTHOR SHALL BE LIABLE FOR DAMAGES ARISING HEREFROM. THE FACT THAT AN ORGANIZATION OR WEBSITE IS REFERRED TO IN THIS WORK AS A CITATION AND/OR A POTENTIAL SOURCE OF FURTHER INFORMATION DOES NOT MEAN THAT THE AUTHOR OR THE PUBLISHER ENDORSES THE INFORMATION THE ORGANIZATION OR WEBSITE MAY PROVIDE OR RECOMMENDATIONS IT MAY MAKE. FURTHER, READERS SHOULD BE AWARE THAT INTERNET WEBSITES LISTED IN THIS WORK MAY HAVE CHANGED OR DISAPPEARED BETWEEN WHEN THIS WORK WAS WRITTEN AND WHEN IT IS READ.

For general information on our other products and services, please contact our Customer Care Department within the U.S. at 877-762-2974, outside the U.S. at 317-572-3993, or fax 317-572-4002. For technical support, please visit <https://hub.wiley.com/community/support/dummies>.

Wiley publishes in a variety of print and electronic formats and by print-on-demand. Some material included with standard print versions of this book may not be included in e-books or in print-on-demand. If this book refers to media such as a CD or DVD that is not included in the version you purchased, you may download this material at <http://booksupport.wiley.com>. For more information about Wiley products, visit www.wiley.com.

Library of Congress Control Number: 2020937204

ISBN 978-1-119-65177-2 (pbk); ISBN 978-1-119-65175-8 (ebk); ISBN 978-1-119-65178-9 (ebk)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

Contents at a Glance

Introduction	1
Part 1: Intro to Analytics and Blockchain	5
CHAPTER 1: Driving Business with Data and Analytics	7
CHAPTER 2: Digging into Blockchain Technology	19
CHAPTER 3: Identifying Blockchain Data with Value	39
CHAPTER 4: Implementing Blockchain Analytics in Business	57
CHAPTER 5: Interacting with Blockchain Data	79
Part 2: Fetching Blockchain Chain	105
CHAPTER 6: Parsing Blockchain Data and Building the Analysis Dataset	107
CHAPTER 7: Building Basic Blockchain Analysis Models	121
CHAPTER 8: Leveraging Advanced Blockchain Analysis Models	139
Part 3: Analyzing and Visualizing Blockchain Analysis Data	151
CHAPTER 9: Identifying Clustered and Related Data	153
CHAPTER 10: Classifying Blockchain Data	171
CHAPTER 11: Predicting the Future with Regression	189
CHAPTER 12: Analyzing Blockchain Data over Time	209
Part 4: Implementing Blockchain Analysis Models	223
CHAPTER 13: Writing Models from Scratch	225
CHAPTER 14: Calling on Existing Frameworks	239
CHAPTER 15: Using Third-Party Toolsets and Frameworks	251
CHAPTER 16: Putting It All Together	267
Part 5: The Part of Tens	279
CHAPTER 17: Ten Tools for Developing Blockchain Analytics Models	281
CHAPTER 18: Ten Tips for Visualizing Data	295
CHAPTER 19: Ten Uses for Blockchain Analytics	305
Index	315

Table of Contents

INTRODUCTION	1
About This Book	1
Foolish Assumptions	2
Icons Used in This Book	2
Beyond the Book	2
Where to Go from Here	3
PART 1: INTRO TO ANALYTICS AND BLOCKCHAIN	5
CHAPTER 1: Driving Business with Data and Analytics	7
Deriving Value from Data	8
Monetizing data	8
Exchanging data	9
Verifying data	10
Understanding and Satisfying Regulatory Requirements	11
Classifying individuals	11
Identifying criminals	11
Examining common privacy laws	12
Predicting Future Outcomes with Data	13
Classifying entities	13
Predicting behavior	14
Making decisions based on models	16
Changing Business Practices to Create Desired Outcomes	16
Defining the desired outcome	17
Building models for simulation	17
Aligning operations and assessing results	18
CHAPTER 2: Digging into Blockchain Technology	19
Exploring the Blockchain Landscape	20
Managing ownership transfer	20
Doing more with blockchain	21
Understanding blockchain technology	21
Reviewing blockchain's family tree	22
Fitting blockchain into today's businesses	25
Understanding Primary Blockchain Types	27
Categorizing blockchain implementations	27
Describing basic blockchain type features	29
Contrasting popular enterprise blockchain implementations	30
Aligning Blockchain Features with Business Requirements	31
Reviewing blockchain core features	31
Examining primary common business requirements	33
Matching blockchain features to business requirements	34

Examining Blockchain Use Cases	35
Managing physical items in cyberspace.....	35
Handling sensitive information	36
Conducting financial transactions.....	37
CHAPTER 3: Identifying Blockchain Data with Value	39
Exploring Blockchain Data	40
Understanding what's stored in blockchain blocks	40
Recording transaction data	41
Dissecting the parts of a block.....	43
Decoding block data	47
Categorizing Common Data in a Blockchain	49
Serializing transaction data	49
Logging events on the blockchain.....	50
Storing value with smart contracts.....	52
Examining Types of Blockchain Data for Value.....	52
Exploring basic transaction data.....	53
Associating real-world meaning to events.....	53
Aligning Blockchain Data with Real-World Processes	54
Understanding smart contract functions.....	55
Assessing smart contract event logs.....	55
Ranking transaction and event data by its effect	55
CHAPTER 4: Implementing Blockchain Analytics in Business... 57	
Aligning Analytics with Business Goals.....	58
Leveraging newly accessible decentralized tools	58
Monetizing data.....	59
Exchanging and integrating data effectively	59
Surveying Options for Your Analytics Lab	60
Installing the Blockchain Client	61
Installing the Test Blockchain	65
Installing the Testing Environment.....	68
Getting ready to install Truffle.....	69
Downloading and installing Truffle.....	72
Installing the IDE	74
CHAPTER 5: Interacting with Blockchain Data	79
Exploring the Blockchain Analytics Ecosystem	80
Reviewing your blockchain lab	80
Identifying analytics client options	81
Choosing the best blockchain analytics client.....	83
Adding Anaconda and Web3.js to Your Lab	84
Verifying platform prerequisites	84
Installing the Anaconda platform	86

Installing the Web3.py library	89
Setting up your blockchain analytics project.....	90
Writing a Python Script to Access a Blockchain.....	92
Interfacing with smart contracts.....	93
Finding a smart contract's ABI	94
Building a Local Blockchain to Analyze.....	100
Connecting to your blockchain	101
Invoking smart contract functions	101
Fetching blockchain data	102
PART 2: FETCHING BLOCKCHAIN CHAIN	105
CHAPTER 6: Parsing Blockchain Data and Building the Analysis Dataset	107
Comparing On-Chain and External Analysis Options.....	108
Considering access speed	108
Comparing one-off versus repeated analysis	109
Assessing data completeness	110
Integrating External Data	111
Determining what data you need	112
Extending identities to off-chain data.....	113
Finding external data	114
Identifying Features	115
Describing how features affect outcomes.....	116
Comparing filtering and wrapping methods.....	116
Building an Analysis Dataset	117
Connecting to multiple data sources	118
Building a cross-referenced dataset.....	118
Cleaning your data	118
CHAPTER 7: Building Basic Blockchain Analysis Models	121
Identifying Related Data.....	122
Grouping data based on features (attributes).....	123
Determining group membership	126
Discovering relationships among items.....	129
Making Predictions of Future Outcomes	130
Selecting features that affect outcome	131
Beating the best guess	133
Building confidence.....	134
Analyzing Time-Series Data	135
Exploring growth and maturity	137
Identifying seasonal trends	138
Describing cycles of results	138

CHAPTER 8:	Leveraging Advanced Blockchain Analysis Models	139
	Identifying Participation Incentive Mechanisms	140
	Complying with mandates	141
	Playing games with partners	141
	Rewarding and punishing participants	142
	Managing Deployment and Maintenance Costs	143
	Lowering the cost of admission	143
	Leveraging participation value	145
	Aligning ROI with analytics currency	146
	Collaborating to Create Better Models	147
	Collecting data from a cohort	148
	Building models collaboratively	148
	Assessing model quality as a team	149
	PART 3: ANALYZING AND VISUALIZING BLOCKCHAIN ANALYSIS DATA	151
CHAPTER 9:	Identifying Clustered and Related Data	153
	Analyzing Data Clustering Using Popular Models	154
	Delivering valuable knowledge with cluster analysis	154
	Examining popular clustering techniques	155
	Understanding k-means analysis	155
	Evaluating model effectiveness with diagnostics	160
	Implementing Blockchain Data Clustering Algorithms in Python	160
	Discovering Association Rules in Data	163
	Delivering valuable knowledge with association rules analysis	163
	Describing the apriori association rules algorithm	164
	Evaluating model effectiveness with diagnostics	167
	Determining When to Use Clustering and Association Rules	168
CHAPTER 10:	Classifying Blockchain Data	171
	Analyzing Data Classification Using Popular Models	172
	Delivering valuable knowledge with classification analysis	172
	Examining popular classification techniques	173
	Understanding how the decision tree algorithm works	173
	Understanding how the naïve Bayes algorithm works	176
	Evaluating model effectiveness with diagnostics	178
	Implementing Blockchain Classification Algorithms in Python	179
	Defining model input data requirements	179
	Building your classification model dataset	181
	Developing your classification model code	184
	Determining When Classification Fits Your Analytics Needs	188

CHAPTER 11: Predicting the Future with Regression	189
Analyzing Predictions and Relationships Using Popular Models	190
Delivering valuable knowledge with regression analysis	190
Examining popular regression techniques	191
Describing how linear regression works	195
Describing how logistic regression works	198
Evaluating model effectiveness with diagnostics	201
Implementing Regression Algorithms in Python	203
Defining model input data requirements	203
Building your regression model dataset	203
Developing your regression model code	204
Determining When Regression Fits Your Analytics Needs	207
CHAPTER 12: Analyzing Blockchain Data over Time	209
Analyzing Time Series Data Using Popular Models	210
Delivering valuable knowledge with time series analysis	211
Examining popular time series techniques	211
Visualizing time series results	214
Implementing Time Series Algorithms in Python	216
Defining model input data requirements	217
Developing your time series model code	219
Determining When Time Series Fits Your Analytics Needs	221
PART 4: IMPLEMENTING BLOCKCHAIN ANALYSIS MODELS	223
CHAPTER 13: Writing Models from Scratch	225
Interacting with Blockchains	226
Connecting to a Blockchain	226
Using an application programming interface to interact with a blockchain	228
Reading from a blockchain	230
Updating previously read blockchain data	234
Examining Blockchain Client Languages and Approaches	236
Introducing popular blockchain client programming languages	237
Comparing popular language pros and cons	238
Deciding on the right language	238
CHAPTER 14: Calling on Existing Frameworks	239
Benefitting from Standardization	240
Easing the burden of compliance	240
Avoiding inefficient code	242
Raising the bar on quality	244

Focusing on Analytics, Not Utilities	245
Avoiding feature bloat	245
Setting granular goals	246
Managing post-operational models	247
Leveraging the Efforts of Others	248
Deciding between make or buy	248
Scoping your testing efforts	249
Aligning personnel expertise with tasks	250
CHAPTER 15: Using Third-Party Toolsets and Frameworks	251
Surveying Toolsets and Frameworks	252
Describing TensorFlow	253
Examining Keras	255
Looking at PyTorch	256
Supercharging PyTorch with fast.ai	258
Presenting Apache MXNet	260
Introducing Caffe	261
Describing Deeplearning4j	262
Comparing Toolsets and Frameworks	264
CHAPTER 16: Putting It All Together	267
Assessing Your Analytics Needs	268
Describing the project's purpose	268
Defining the process	270
Taking inventory of resources	271
Choosing the Best Fit	273
Understanding personnel skills and affinity	273
Leveraging infrastructure	275
Integrating into organizational culture	276
Embracing iteration	276
Managing the Blockchain Project	277
PART 5: THE PART OF TENS	279
CHAPTER 17: Ten Tools for Developing Blockchain Analytics Models	281
Developing Analytics Models with Anaconda	282
Writing Code in Visual Studio Code	283
Prototyping Analytics Models with Jupyter	284
Developing Models in the R Language with RStudio	285
Interacting with Blockchain Data with web3.py	287
Extract Blockchain Data to a Database	288
Extracting blockchain data with EthereumDB	288
Storing blockchain data in a database using Ethereum-etl	288

Accessing Ethereum Networks at Scale with Infura	289
Analyzing Very Large Datasets in Python with Vaex.	290
Examining Blockchain Data	291
Exploring Ethereum with Etherscan.io	291
Perusing multiple blockchains with Blockchain.com	292
Viewing cryptocurrency details with ColossusXT	293
Preserving Privacy in Blockchain Analytics with MADANA	293
CHAPTER 18: Ten Tips for Visualizing Data	295
Checking the Landscape around You	296
Leveraging the Community	297
Making Friends with Network Visualizations.	298
Recognizing Subjectivity	299
Using Scale, Text, and the Information You Need	300
Considering Frequent Updates for Volatile Blockchain Data	301
Getting Ready for Big Data.	302
Protecting Privacy	302
Telling Your Story	303
Challenging Yourself!	303
CHAPTER 19: Ten Uses for Blockchain Analytics	305
Accessing Public Financial Transaction Data	306
Connecting with the Internet of Things (IoT).	307
Ensuring Data and Document Authenticity	308
Controlling Secure Document Integrity	308
Tracking Supply Chain Items	310
Empowering Predictive Analytics	310
Analyzing Real-Time Data.	311
Supercharging Business Strategy	312
Managing Data Sharing	312
Standardizing Collaboration Forms	312
INDEX	315

Introduction

Data is the driver of today's organizations. Ignore the vast amounts of data available to you about your products, services, customers, and even competitors, and you'll quickly fall behind. But if you embrace data and mine it like it contains valuable jewels, you could find the edge to stay ahead of your competition and keep your customers happy.

And the potential value you can find in data gets even more enticing when you incorporate blockchain technology into your organization. Blockchain is a fast-growing innovation that maintains untold pieces of information you could use to decrease costs and increase revenue. Realizing blockchain data value depends on understanding how blockchain stores data and how to get to it.

Blockchain Data Analytics For Dummies introduces readers to blockchain technology, how it stores data, how to identify and get to interesting data, and how to analyze that data to find meaningful information. You learn how to set up your own blockchain analytics lab and local blockchain to learn and practice blockchain analytics techniques. After you set up your analytics lab, you find out how to extract blockchain data and build popular analytics models to uncover your data's hidden information.

About This Book

Blockchain technology is often described as the most important and disruptive technology of our generation. At its core, blockchain technology provides a novel way to add data to a ledger of transactions that is shared by other users whom you do not trust. Blockchain technology has the potential to change the way we conduct business at every level. And, while managing transactions between any two or more parties, any data related to the transaction gets stored on the shared ledger that can never be changed or deleted. The availability of unmodified history of transactions can be a huge advantage for organizations of all types.

Unlocking the trends or lessons in these blockchain transactions is the focus of this book. *Blockchain Data Analytics For Dummies* gives you the foundation of blockchain technology data storage and techniques to analyze blockchain-based data. You learn — in clear language — how to build analytics models and populate them with blockchain data.

Foolish Assumptions

I don't make many assumptions about your experience with blockchain technology, application programming, or cryptography, but I do assume the following:

- » You have a computer and access to the Internet.
- » You know the basics of using your computer and the Internet, as well as how to download and install programs.
- » You know how to find files on your computer's disk and how to create folders.
- » You're new to blockchain and you aren't an experienced software developer.
- » You're new to building data analytics models.

Icons Used in This Book



TIP

The Tip icon marks tips (duh!) and shortcuts you can use to extract blockchain data and build analytics models.



REMEMBER

Remember icons mark the information that's especially important to know.



TECHNICAL
STUFF

The Technical Stuff icon marks information of a highly technical nature that you can normally skip over.



WARNING

The Warning icon tells you to watch out! It marks important information that may save you headaches when writing your own blockchain applications.

Beyond the Book

In addition to the material in the print or e-book you're reading right now, this product also comes with some access-anywhere goodies on the web. Check out the free cheat sheet for more on blockchain technology and data analytics at www.dummies.com/cheatsheet/blockchaindataanalyticsfd.

You'll find summary information about blockchain technology, data analytics models, and extracting blockchain data. The cheat sheet is a reference to use over and over as you gain experience in extracting blockchain data and building data analytics models.

In addition, if you'd rather download the code you see in this book instead of typing it, go to <http://www.dummies.com/go/blockchaindataanalyticsfd>. You can download zip files for each of the projects you'll create to develop and test data access and analytics scripts.

Where to Go from Here

The *Dummies* series tells you what you need to know and how to do the things you need to do to get the results you want. Readers don't have to read the entire book to just learn about some topics. For example, if you just want to learn about extracting blockchain data, you can jump right to Chapters 5 and 6. On the other hand, if you need to set up your own blockchain analytics lab, read Chapter 4, which tells you how to do that with clear, step-by-step instructions.

1

Intro to Analytics and Blockchain

IN THIS PART . . .

Using data analytics to drive strategic decisions

Exploring blockchain technology and popular use cases

Examining blockchain data to identify data of value

Building a blockchain analytics lab

Populating a local blockchain with data to analyze

- » Discovering the value of data
- » Complying with regulations
- » Protecting customer privacy
- » Predicting expected actions with data
- » Changing plans to control outcome

Chapter 1

Driving Business with Data and Analytics

In the twenty-first century, personalization is king — and data makes personalization possible. A good friend can pick out a much more personal gift for you than a stranger because that friend knows what you like and dislike. Marketers have known for decades that establishing a connection with someone can dramatically increase the chances that the person will become a customer. Organizations' desire to attract customers and increase sales drives the pursuit of meeting consumers' needs.

Consumers demand personal attention and have come to expect a high level of individualized customer service, online or when physically shopping in a bricks-and-mortar store. Due to advances in consumer interaction sophistication, the bar is high for all types of organizations. For example, it isn't good enough for web searches to return a general list of responses. Consumers expect their searches to be personalized and filtered based on their preferences. Today's search engines, and most shopping sites, suggest responses before you even finish typing. It's almost as if the search function knows you and what you're about to ask.

The capability to guess what a user is likely to ask or find interesting is based on data. Humans are creatures of habit and most processes (and even natural events) tend to be cyclic. The repetitive nature of behavior means that if you have enough historical data, you should be able to predict what comes next. Expending effort to

collect, maintain, and analyze data related to your organization's operation can help to reduce costs, limit exposure to fines and lawsuits, and lead to increased revenue.

In short, learning how to use your data helps you learn how to make your organization more profitable. In this chapter, you learn about ways that data can provide value to organizations.

Deriving Value from Data

The increased trend toward personalized offerings both depends on data and exposes data's importance to business operations. Data is no longer simply a consequence of engaging in transactions — data is necessary to increase the volume of transactions. Organizations are learning how valuable data is to their capability to conduct and expand operations. If you want to stay competitive in today's economy, you'll have to provide an experience that's responsive and personal. Data from previous transactions makes it possible to anticipate subsequent activity and tailor offerings to customer and partner preferences.

For example, the items you've bought online in the past give online shopping sites such as Amazon.com enough of your background to be able to make suggestions for additional purchases. Using past data to recommend future purchase or actions is a common way to derive value from data. In this section, I introduce three ways organizations can identify data with the greatest potential value.

Monetizing data

Over the past two decades, many organizations have come to view data as the primary fuel of the information age. Since the dawn of the twenty-first century, many organizations with data as their central business driver either started or expanded rapidly. Amazon relies on customer data to make additional purchase suggestions, while companies such as Facebook and Google rely on data as their primary product to drive advertising revenue. All these organizations found ways to turn data into revenue.

As data becomes more directly associated with revenue, data giants Google, Facebook, and Amazon control a growing demand for access to that data. Users have long been encouraged to share their personal data and activities, with little or no compensation. In the beginning, the perception was that sharing personal data was harmless and had little value.

However, a growing number of consumers and business partners realize that their data has value. Legislative bodies have recognized the importance of personal data and are passing new levels of privacy protection legislation each year. Data not only has value in and of itself but, when linked to other related personal data, can also provide valuable insight into personal behavior.

The realization that personal data has value has resulted in a game of sorts. Organizations that value consumer data attempt to acquire as much data as possible, while consumers are becoming more willing to deny free access to their personal data or demand compensation. Compensation often takes the form not of a direct monetary payment but of other perks or discounts.

Exchanging data

As organizations realize the increasing value of consumer and partner data, the more they explore ways to leverage that value. When consumers interact with any organization, or organizations interact with partners, a trail of data artifacts is left behind. Artifacts that document transaction timing and contents, as well as any changes to data, describe how entities interact with organizations. As more interactions with all types of organizations become more automated, the quantity and frequency of data artifacts increases.

Organizations that collect data artifacts find that not all are useful — at least not to that organization. However, as data becomes more and more valuable, many organizations have expanded the scope of data they collect with the intention of selling that data to other organizations. As data becomes a source of both direct and indirect revenue, data collection and management moves from a supporting role to a strategic planning concern.

For example, political campaigns routinely spend large sums of money to purchase demographic information on customers who have purchased specific types of products. Political candidates who strongly support environmental issues find value in identifying people who purchase green products because these customers are likely potential supporters. The identities can then be used to solicit campaign donations.



TIP

The overuse of data selling has led to concern and frustration over personal privacy. Most people come to the eventual realization that online activity has consequences. Every time you provide your email address or telephone number to anyone, your data will likely end up being used by some other organization (or probably multiple organizations). Always be careful about what data you allow others to use.

Sharing and exchanging data isn't always bad. In some cases, you want your data to be shared among businesses and organizations. For example, sharing the

complete service history for your car could make getting service easier and more reliable. With shared service data, you could take your car to any service provider and not have to remember the last time you had the oil changed or tires rotated. Techniques that support beneficial and responsible data sharing among organizations can be valuable to business and consumers.

Verifying data

One of the obstacles to realizing the full value of data is the dependence on its quality. Quality data is valuable, while incomplete or untrusted data is often worthless. What's worse, low-quality data may require more budget to clean than it will potentially generate in revenue. The only way to realize data's true value is to ensure that the data is valid and represents entities in the real world.

Verifying data has long been one of the highest costs associated with collecting and using data. Campaigns that depend on physical or email addresses will have little effect if the target addresses are largely incorrect. Bad data can come from many sources, including mischievous data submission, sloppy data collection, or even malicious data modification. An important aspect of relying on data is putting controls in place that verify the source of any collected data, along with that data's adherence to collection requirements.

A simple approach to verifying data in a distributed environment is to carry out a simple validation at the source and again at the server as the data is stored in a repository. While validating data at least twice may seem excessive, the practice makes user errors easier to catch and ensures that data received by the server is clean.



TIP

Validating data twice makes it possible for client applications to quickly catch errors, such as too many digits in a phone number or a missing field, while the server handles more complex validation tasks. A server may need access to other related data to ensure that data is valid before storing it in a repository. Server validation could include things such as verifying that order quantities are available in a warehouse and that data wasn't changed by a malicious agent during transmission from the client.

One of the reasons data verification is so important is that organizations are relying more and more on their data to direct business efforts. Aligning business activities with expectations based on faulty data leads to undesirable results. In other words, decisions are only as good as the data on which those decisions are based. The "garbage in, garbage out" adage still holds true.

Understanding and Satisfying Regulatory Requirements

The information age offers many new opportunities and just as many (if not more) challenges. The vast amount of data available to organizations of all types empowers advanced decision-making and raises new questions of privacy and ethics. Consumer protection groups have long been voicing concerns about how personal data is being used. In response to discovered abuses and the recognition of potential future abuses, governing bodies around the world have passed regulations and legislation to limit how data is collected and used.

Although collecting a few pieces of information about a customer may seem innocent, it doesn't take long for accumulated data to paint a picture of an individual's personal characteristics and behavior. Knowing the past behavior of someone makes it relatively easy to predict the person's future actions and choices. Predicting actions has value for marketing but also poses a danger to an individual's privacy.

Classifying individuals

The concern is that personal data has been, and will continue to be, used to classify individuals based on their past behavior. Classifying individuals can be great for marketing and sales purposes. For example, any retailer that can identify engaged couples can target them with ads and coupons for wedding-related items. This type of targeted advertising is generally more productive than general marketing. Advertising budget can be focused on target markets that provide the greatest ROI.

On the other hand, knowing too much about individuals may violate a person's privacy. One instance of a privacy violation was a result of the Target Corporation's astute data analysis. Target's analysts were able to identify expectant mothers early in their pregnancy based on their changing purchasing habits. When a new expectant mother was identified, Target would send unsolicited coupons for baby-related items. In one case, the coupons arrived in the mail before the mother had shared that she was pregnant; her family found out about the pregnancy from a retailer. Privacy is such a difficult issue because legitimate actions can violate a person's privacy.

Identifying criminals

Another aspect of privacy is when criminals, or other individuals who deliberately want to operate anonymously, hide their identities from exposure. Privacy may be important to the general population, but it's a necessity for criminal activity. The ability to deny, or repudiate, some action is crucial in avoiding discovery and

capture, and to any subsequent defense. Money laundering and fraud are two activities in which privacy and anonymity are desired to obfuscate illegal activity.

On the other hand, law enforcement needs the ability to associate actions with individuals. That's why laws exist that protect the general public but allow law enforcement to conduct investigations and identify alleged perpetrators.

Protecting the privacy of law-abiding individuals while identifying criminals has become important across a spectrum of organizations. To enable law enforcement to deal with online privacy issues, legislative bodies have passed various laws to address those issues directly.

Examining common privacy laws

Here are a few of the most important privacy-related laws you'll likely encounter and may be compelled to satisfy:

- » **Children's Online Privacy Protection Act (COPPA):** Passed in 1998, COPPA requires parental or guardian consent before collecting or using private information about children under the age of 13.
- » **Health Insurance Portability and Accountability Act (HIPAA):** Passed in 1996, HIPAA modernized the flow of healthcare information and contains specific stipulations on protecting the privacy of personal health information (PHI).
- » **Family Educational Rights and Privacy Act (FERPA):** Passed in 1974, FERPA protects access to educational information, including protection for the privacy of student records.
- » **General Data Protection Regulation (GDPR):** Passed in 2016 (and implemented in 2018), GDPR is a comprehensive regulation from the European Union (EU) protecting the private data of EU citizens. Every organization, regardless of location, must comply with GDPR to conduct business with EU citizens. The EU citizen must retain control over his or her own data, its collection, and its use.
- » **California Consumer Protection Act (CCPA):** Passed in 2018, CCPA has been called "GDPR lite" to imply that it includes many of the requirements of GDPR. CCPA requires any organization that conducts business to protect consumer data privacy.
- » **Anti-Money Laundering Act (AML):** AML is a set of laws and regulations that assists law enforcement investigations by requiring financial transactions to be associated with validated identities. AML imposes requirements and

procedures on financial institutions that essentially make it very difficult to transfer money without leaving a clear audit trail.

- » **Know Your Customer (KYC):** KYC laws and regulations work with AML to ensure that businesses expend reasonable effort to verify the identity of each customer and business partner. KYC helps to discourage money laundering, bribery, and other financial-based criminal activities that rely on anonymity.

Predicting Future Outcomes with Data

Data can unlock lots of secrets. Data you collect through regular interactions with your customers and business partners can help you understand them and better meet their needs and wants. Assuming you have taken measures to protect individual privacy and have permission to collect and use the data, analyzing that data can benefit your organization and your customers (and partners, too).

A common way to use data is to build analytics models that help to explain the data, uncover hidden information, and even predict future behavior. Data analytics is all about using formal methods to unlock secrets that your data is hiding. These secrets aren't hidden on purpose — they just get lost in the mountains of data you collect. Without a structured approach to examining your data, you might miss some of its value that can lead to increased revenue.

Classifying entities

An *entity* is any object that your data describes, such as a customer, a vendor, a product, an order, or anything else that has characteristics data items can describe. In traditional database terms, an entity would correspond to a record or a row. The concept of a row maps to a spreadsheet concept as well. Think of a spreadsheet of customers. Each row would contain all the data that describes a single customer. Figure 1-1 shows a collection of customers in a table format.



TECHNICAL
STUFF

These customers are stored in a comma-separated value (CSV) text file named `customer.csv`, and displayed in Visual Studio Code using the Edit as CSV extension. To learn more about Visual Studio Code and its extensions, see Chapter 4.

Note that each customer has a set of characteristics, such as name, address, and contact, stored in separate columns. Data analytics models use these different characteristics, also called *features*, to examine how different entities are related.

FIGURE 1-1:
Customer entities
presented as a
table.

	CustNum	Country	Name	Address	Address2	City	State	PostalCode	Contact
1	1	USA	Lift Tours	276 North Drive		Burlington	MA	01730	Gloria Shepley
2	2	Finland	Upton Frisbee	Rattipolku 3		Oslo	Uusimaa	45321	Urho Leppakoski
3	3	USA	Hoops	Suite 415	40 Grove St.	Atlanta	GA	02112	Michael Traister
4	4	United Kingdom	Go Fishing Ltd	Unit 2	83 Ponders End Rd	Harrow	Middlesex	HA8 3BB	Alan Froghenok
5	5	USA	Match Point Tennis	66 Homer Pl	Address 2	Boston	MA	02134	Robert Door
6	6	United Kingdom	Fanatical Athletes	20 Bicep Bridge Rd	Leyton	Montgomery	AL	48855	Andrew Strong
7	7	Finland	Aerobics valine Ky	Pelohantie 3		Tikkurila	Uusimaa	4567	Elli Imanen
8	8	USA	Game Set Match	Box 60		Deatsville	AL	85114	Tore Break
9	9	Finland	Pihlajataun Pyora	Puikontie 2		Riihimäki	Etelä-Savo	44400	Matti Vento
10	10	United Kingdom	Just Joggers Limited	Fairwind Trading Est	Shoe Lane	Ramsbottom	Sussex	BL0 9ND	George Lacey
11	11	Finland	Kellalu ja Biljardi	Vattuniemenuja 8 A		Helsinki	Uusimaa	210	Ritza Hammermaki
12	12	Finland	Surf Lautareikkoset	Venemestarinkatu 35		Salo	Varsinais-Suomi	24240	Veikko Lehtinen
13	13	Finland	Biljardi ja tennis	Uinhuuritie 1		Mantala	Uusimaa	4600	Tuomo A. Mattinen
14	14	France	Paris St Germain	113, avenue du Stade	BP 338	Paris	Seine	75012	Philippe Dargil
15	15	USA	Hoopia Basketball	87 Calumnet St		Egg Harbor	NJ	08215	Robert Johns
16	16	USA	Thundering Surf Inc.	354 Market St		Coffee City	TX	75763	Fran Horan
17	17	United Kingdom	High Tide Sailing	178 Schooner Hill	South Killingholme	Birmingham	Humber-side	DN46 3LU	Angela Tackle
18	18	Finland	Amin Testastysse	Vanhainkirkonkatu 1		Syama	Keski-Suomi	19700	Antti Oymmetta
19	19	USA	Buffalo Shuffleboard	155 Carolina Ave		Buffalo	NY	13142	George Butkus
20	20	Finland	Espoon Pallokeskus	Sinkalontie 18		Espoo	Uusimaa	1200	Pekka Maukka
21	21	USA	Pedal Power Cycles	11 Perkins St		Boston	MA	02145	Alicia Primes
22	22	Finland	Kesport Katiska	Koskikallantie 4		Forsas	Etelä-Hame	30100	Pentti Koskinen
23	24	Finland	Jazz Futsi Kauppa	Yrjokatu 10		Pori	Satakunta	28100	Markku Vaa-laeoja
24	25	France	La Boule Lyonaise	rue de la gare	Z.A. Bellecourt	Lyon	Rhone	69010	Mr Roulemousse
25	26	United Kingdom	Bulls Eye Sports	Highcore House	180 Goodshot Street	Leeds	West Yorkshire	LS17 8TS	Kevin Butler
26	27	Austria	Bumm Bumm Tennis	Weizenmarkt 17		Wien	Ost	A-1020	Hr. Baller
27	28	Finland	Luisin ja Pyora Oy	Ruutintie 7 A		Laihia	Etelä-Pohjanmaa	66400	Jarkko J. Scott
28	29	USA	Bug in a Rug by	120 Henry St		Yamhill	OR	97148	Chip Clark
29	30	USA	South Killingholme	14 Gordon St		Boston	MA	02114	Robert Door

One type of analysis is to examine the features of different entities to see if some features can help group entities or imply some relationship. For example, suppose you asked a group of people to name their favorite baseball team. You would expect that most people who answered “the Colorado Rockies” most likely live near Colorado. However, you can’t always make such simple associations. If you asked the same question in the 1990s, not everyone who answered “the Atlanta Braves” lived in Georgia. During the 1990s, cable TV was becoming popular and Turner Broadcasting System, whose owner also owned the Braves, broadcast all Braves games nationally. Many people who didn’t live in Georgia became Braves fans.

The Braves example shows that analytics models cannot be trusted unconditionally. Data analytics can provide tremendous value but also requires care and diligence to build models that return results that hold true over time.

Assuming that you invest sufficiently to build good models, classification models can help to identify entities that are similar. Similarity information helps organizations develop targeted marketing campaigns and services to give customers and partners the sense of being treated individually. You learn about several classification models in Chapter 7 and build a few in Chapter 10.

Predicting behavior

Although the capability to classify entities to identify groups of similarity can be valuable, analytics can also make predictions. Past behavior is a strong indication of future behavior. Humans tend to repeat actions and decisions, so you can use models that identify patterns to predict future actions. The capability to predict future actions can have tremendous value to organizations. If an organization can determine items that tend to be purchased together frequently, it can use that information to make additional purchase suggestions.