



# Data Analytics for IT Networks

Developing Innovative Use Cases

[ciscopress.com](http://ciscopress.com)

**John Garrett**  
CCIE No. 6204 Emeritus  
MS Predictive Analytics

FREE SAMPLE CHAPTER

SHARE WITH OTHERS



# **Data Analytics for IT Networks**

---

## **Developing Innovative Use Cases**

John Garrett CCIE Emeritus No. 6204, MSPA

**Cisco Press**

# Data Analytics for IT Networks

## Developing Innovative Use Cases

Copyright © 2019 Cisco Systems, Inc.

Published by:

Cisco Press

All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without written permission from the publisher, except for the inclusion of brief quotations in a review.

First Printing 1 18

Library of Congress Control Number: 2018949183

ISBN-13: 978-1-58714-513-1

ISBN-10: 1-58714-513-8

## Warning and Disclaimer

This book is designed to provide information about Developing Analytics use cases. It is intended to be a guideline for the networking professional, written by a networking professional, toward understanding Data Science and Analytics as it applies to the networking domain. Every effort has been made to make this book as complete and as accurate as possible, but no warranty or fitness is implied.

The information is provided on an “as is” basis. The authors, Cisco Press, and Cisco Systems, Inc. shall have neither liability nor responsibility to any person or entity with respect to any loss or damages arising from the information contained in this book or from the use of the discs or programs that may accompany it.

The opinions expressed in this book belong to the author and are not necessarily those of Cisco Systems, Inc.

MICROSOFT AND/OR ITS RESPECTIVE SUPPLIERS MAKE NO REPRESENTATIONS ABOUT THE SUITABILITY OF THE INFORMATION CONTAINED IN THE DOCUMENTS AND RELATED GRAPHICS PUBLISHED AS PART OF THE SERVICES FOR ANY PURPOSE. ALL SUCH DOCUMENTS AND RELATED GRAPHICS ARE PROVIDED “AS IS”

WITHOUT WARRANTY OF ANY KIND. MICROSOFT AND/OR ITS RESPECTIVE SUPPLIERS HEREBY DISCLAIM ALL WARRANTIES AND CONDITIONS WITH REGARD TO THIS INFORMATION, INCLUDING ALL WARRANTIES AND CONDITIONS OF MERCHANTABILITY, WHETHER EXPRESS, IMPLIED OR STATUTORY, FITNESS FOR A PARTICULAR PURPOSE, TITLE AND NON-INFRINGEMENT. IN NO EVENT SHALL MICROSOFT AND/OR ITS RESPECTIVE

SUPPLIERS BE LIABLE FOR ANY SPECIAL, INDIRECT OR CONSEQUENTIAL DAMAGES OR ANY DAMAGES WHATSOEVER RESULTING FROM LOSS OF USE, DATA OR PROFITS, WHETHER IN AN ACTION OF CONTRACT, NEGLIGENCE OR OTHER TORTIOUS ACTION, ARISING OUT OF OR IN CONNECTION WITH THE USE OR PERFORMANCE OF INFORMATION AVAILABLE FROM THE SERVICES.

THE DOCUMENTS AND RELATED GRAPHICS CONTAINED HEREIN COULD INCLUDE TECHNICAL INACCURACIES OR TYPOGRAPHICAL ERRORS. CHANGES ARE PERIODICALLY ADDED TO THE INFORMATION HEREIN. MICROSOFT AND/OR ITS RESPECTIVE SUPPLIERS MAY MAKE IMPROVEMENTS AND/OR CHANGES IN THE PRODUCT(S) AND/OR THE PROGRAM(S) DESCRIBED HEREIN AT ANY TIME. PARTIAL SCREEN SHOTS MAY BE VIEWED IN FULL WITHIN THE SOFTWARE VERSION SPECIFIED.

## **Trademark Acknowledgments**

All terms mentioned in this book that are known to be trademarks or service marks have been appropriately capitalized. Cisco Press or Cisco Systems, Inc., cannot attest to the accuracy of this information. Use of a term in this book should not be regarded as affecting the validity of any trademark or service mark.

MICROSOFT® WINDOWS®, AND MICROSOFT OFFICE® ARE REGISTERED TRADEMARKS OF THE MICROSOFT CORPORATION IN THE U.S.A. AND OTHER COUNTRIES. THIS BOOK IS NOT SPONSORED OR ENDORSED BY OR AFFILIATED WITH THE MICROSOFT CORPORATION.

## About the Author

**John Garrett** is CCIE Emeritus (6204) and Splunk Certified. He earned an M.S. in predictive analytics from Northwestern University, and has a patent pending related to analysis of network devices with data science techniques. John has architected, designed, and implemented LAN, WAN, wireless, and data center solutions for some of the largest Cisco customers. As a secondary role, John has worked with teams in the Cisco Services organization to innovate on some of the most widely used tools and methodologies at Customer Experience over the past 12 years.

For the past 7 years, John's journey has moved through server virtualization, network virtualization, OpenStack and cloud, network functions virtualization (NFV), service assurance, and data science. The realization that analytics and data science play roles in all these brought John full circle back to developing innovative tools and techniques for Cisco Services. John's most recent role is as an Analytics Technical Lead, developing use cases to benefit Cisco Services customers as part of Business Critical Services for Cisco. John lives with his wife and children in Raleigh, North Carolina.

## About the Technical Reviewers

**Dr. Ammar Rayes** is a Distinguished Engineer at Advance Services Technology Office for Cisco, focusing on network analytics, IoT, and machine learning. He has authored 3 books and more than 100 publications in refereed journals and conferences on advances in software- and networking-related technologies, and he holds more than 25 patents. He is the founding president and board member of the International Society of Service Innovation Professionals ([www.issip.org](http://www.issip.org)), editor-in-chief of the journal *Advancements in Internet of Things* and an editorial board member of the European Alliance for Innovation—Industrial Networks and Intelligent Systems. He has served as associate editor on the journals *ACM Transactions on Internet Technology* and *Wireless Communications and Mobile Computing* and as guest editor on multiple journals and several *IEEE Communications Magazine* issues. He has co-chaired the Frontiers in Service conference and appeared as keynote speaker at several IEEE and industry conferences.

At Cisco, Ammar is the founding chair of Cisco Services Research and the Cisco Services Patent Council. He received the Cisco Chairman's Choice Award for IoT Excellent Innovation and Execution.

He received B.S. and M.S. degrees in electrical engineering from the University of Illinois at Urbana and a Ph.D. in electrical engineering from Washington University in St. Louis, Missouri, where he received the Outstanding Graduate Student Award in Telecommunications.

**Nidhi Kao** is a Data Scientist at Cisco Systems who develops advanced analytic solutions for Cisco Advanced Services. She received a B.S. in biochemistry from North Carolina State University and an M.B.A. from the University of North Carolina Kenan Flagler Business School. Prior to working at Cisco Systems, she held analytic chemist and research positions in industry and nonprofit laboratories.

## Dedications

This book is dedicated to my wife, Veronica, and my children, Lexy, Trevor, and Mason. Thank you for making it possible for me to follow my passions through your unending support.

## Acknowledgments

I would like to thank my manager, Ulf Vinneras, for supporting my efforts toward writing this book and creating an innovative culture where Cisco Services incubation teams can thrive and grow.

To that end, thanks go out to all the people in these incubation teams in Cisco Services for their constant sharing of ideas and perspectives. Your insightful questions, challenges, and solutions have led me to work in interesting roles that make me look forward to coming to work every day. This includes the people who are tasked with incubation, as well as the people from the field who do it because they want to make Cisco better for both employees and customers.

Thank you, Nidhi Kao and Ammar Rayes, for your technical expertise and your time spent reviewing this book. I value your expertise and appreciate your time. Your recommendations and guidance were spot-on for improving the book.

Finally, thanks to the Pearson team for helping me make this career goal a reality. There are many areas of publishing that were new to me, and you made the process and the experience very easy and enjoyable.

## Contents at a Glance

Chapter 1	Getting Started with Analytics	1
Chapter 2	Approaches for Analytics and Data Science	13
Chapter 3	Understanding Networking Data Sources	35
Chapter 4	Accessing Data from Network Components	55
Chapter 5	Mental Models and Cognitive Bias	97
Chapter 6	Innovative Thinking Techniques	127
Chapter 7	Analytics Use Cases and the Intuition Behind Them	147
Chapter 8	Analytics Algorithms and the Intuition Behind Them	217
Chapter 9	Building Analytics Use Cases	273
Chapter 10	Developing Real Use Cases: The Power of Statistics	285
Chapter 11	Developing Real Use Cases: Network Infrastructure Analytics	323
Chapter 12	Developing Real Use Cases: Control Plane Analytics Using Syslog Telemetry	355
Chapter 13	Developing Real Use Cases: Data Plane Analytics	389
Chapter 14	Cisco Analytics	425
Chapter 15	Book Summary	435
Appendix A	Function for Parsing Packets from pcap Files	443
	Index	445



## Contents

	Foreword	xvii
	Introduction: Your future is in your hands!	xviii
<b>Chapter 1</b>	<b>Getting Started with Analytics</b>	<b>1</b>
	What This Chapter Covers	1
	Data: You as the SME	2
	Use-Case Development with Bias and Mental Models	2
	Data Science: Algorithms and Their Purposes	3
	What This Book Does <i>Not</i> Cover	4
	Building a Big Data Architecture	4
	Microservices Architectures and Open Source Software	5
	R Versus Python Versus SAS Versus Stata	6
	Databases and Data Storage	6
	Cisco Products in Detail	6
	Analytics and Literary Perspectives	7
	Analytics Maturity	7
	Knowledge Management	8
	Gartner Analytics	8
	Strategic Thinking	9
	Striving for “Up and to the Right”	9
	Moving Your Perspective	10
	Hot Topics in the Literature	11
	Summary	12
<b>Chapter 2</b>	<b>Approaches for Analytics and Data Science</b>	<b>13</b>
	Model Building and Model Deployment	14
	Analytics Methodology and Approach	15
	Common Approach Walkthrough	16
	Distinction Between the Use Case and the Solution	18
	Logical Models for Data Science and Data	19
	Analytics as an Overlay	20
	Analytics Infrastructure Model	22
	Summary	33
<b>Chapter 3</b>	<b>Understanding Networking Data Sources</b>	<b>35</b>
	Planes of Operation on IT Networks	36
	Review of the Planes	40

Data and the Planes of Operation 42

Planes Data Examples 44

A Wider Rabbit Hole 49

A Deeper Rabbit Hole 51

Summary 53

## **Chapter 4 Accessing Data from Network Components 55**

Methods of Networking Data Access 55

Pull Data Availability 57

Push Data Availability 61

Control Plane Data 67

Data Plane Traffic Capture 68

Packet Data 70

Other Data Access Methods 74

Data Types and Measurement Considerations 76

Numbers and Text 77

Data Structure 82

Data Manipulation 84

Other Data Considerations 87

External Data for Context 89

Data Transport Methods 89

Transport Considerations for Network Data Sources 90

Summary 96

## **Chapter 5 Mental Models and Cognitive Bias 97**

Changing How You Think 98

Domain Expertise, Mental Models, and Intuition 99

Mental Models 99

Daniel Kahneman's System 1 and System 2 102

Intuition 103

Opening Your Mind to Cognitive Bias 104

Changing Perspective, Using Bias for Good 105

Your Bias and Your Solutions 106

How You Think: Anchoring, Focalism, Narrative Fallacy, Framing,  
and Priming 107

How Others Think: Mirroring 110

What Just Happened? Availability, Recency, Correlation, Clustering,  
and Illusion of Truth 111

Enter the Boss: HIPPO and Authority Bias	113
What You Know: Confirmation, Expectation, Ambiguity, Context, and Frequency Illusion	114
What You Don't Know: Base Rates, Small Numbers, Group Attribution, and Survivorship	117
Your Skills and Expertise: Curse of Knowledge, Group Bias, and Dunning-Kruger	119
We Don't Need a New System: IKEA, Not Invented Here, Pro-Innovation, Endowment, Status Quo, Sunk Cost, Zero Price, and Empathy	121
I Knew It Would Happen: Hindsight, Halo Effect, and Outcome Bias	123
Summary	124

## **Chapter 6 Innovative Thinking Techniques 127**

Acting Like an Innovator and Mindfulness	128
Innovation Tips and Techniques	129
Developing Analytics for Your Company	140
Defocusing, Breaking Anchors, and Unpriming	140
Lean Thinking	142
Cognitive Trickery	143
Quick Innovation Wins	143
Summary	144

## **Chapter 7 Analytics Use Cases and the Intuition Behind Them 147**

Analytics Definitions	150
How to Use the Information from This Chapter	151
Priming and Framing Effects	151
Analytics Rube Goldberg Machines	151
Popular Analytics Use Cases	152
Machine Learning and Statistics Use Cases	153
Common IT Analytics Use Cases	170
Broadly Applicable Use Cases	199
Some Final Notes on Use Cases	214
Summary	214

## **Chapter 8 Analytics Algorithms and the Intuition Behind Them 217**

About the Algorithms	217
Algorithms and Assumptions	218
Additional Background	219

Data and Statistics	221
Statistics	221
Correlation	224
Longitudinal Data	225
ANOVA	227
Probability	228
Bayes' Theorem	228
Feature Selection	230
Data-Encoding Methods	232
Dimensionality Reduction	233
Unsupervised Learning	234
Clustering	234
Association Rules	240
Sequential Pattern Mining	243
Collaborative Filtering	244
Supervised Learning	246
Regression Analysis	246
Classification Algorithms	248
Decision Trees	249
Random Forest	250
Gradient Boosting Methods	251
Neural Networks	252
Support Vector Machines	258
Time Series Analysis	259
Text and Document Analysis	262
Natural Language Processing (NLP)	262
Information Retrieval	263
Topic Modeling	265
Sentiment Analysis	266
Other Analytics Concepts	267
Artificial Intelligence	267
Confusion Matrix and Contingency Tables	267
Cumulative Gains and Lift	269
Simulation	271
Summary	271

## **Chapter 9 Building Analytics Use Cases 273**

- Designing Your Analytics Solutions 274
- Using the Analytics Infrastructure Model 275
- About the Upcoming Use Cases 276
  - The Data 276
  - The Data Science 278
  - The Code 280
- Operationalizing Solutions as Use Cases 281
  - Understanding and Designing Workflows 282
- Tips for Setting Up an Environment to Do Your Own Analysis 282
- Summary 284

## **Chapter 10 Developing Real Use Cases: The Power of Statistics 285**

- Loading and Exploring Data 286
- Base Rate Statistics for Platform Crashes 288
- Base Rate Statistics for Software Crashes 299
- ANOVA 305
- Data Transformation 310
  - Tests for Normality 311
  - Examining Variance 313
- Statistical Anomaly Detection 318
- Summary 321

## **Chapter 11 Developing Real Use Cases: Network Infrastructure Analytics 323**

- Human DNA and Fingerprinting 324
- Building Search Capability 325
  - Loading Data and Setting Up the Environment 325
  - Encoding Data for Algorithmic Use 328
  - Search Challenges and Solutions 331
- Other Uses of Encoded Data 336
- Dimensionality Reduction 337
- Data Visualization 340
- K-Means Clustering 344
- Machine Learning Guided Troubleshooting 350
- Summary 353

**Chapter 12 Developing Real Use Cases: Control Plane Analytics Using Syslog Telemetry 355**

- Data for This Chapter 356
- OSPF Routing Protocols 357
- Non-Machine Learning Log Analysis Using pandas 357
  - Noise Reduction 360
  - Finding the Hotspots 362
- Machine Learning–Based Log Evaluation 366
  - Data Visualization 367
  - Cleaning and Encoding Data 369
  - Clustering 373
  - More Data Visualization 375
  - Transaction Analysis 379
- Task List 386
- Summary 387

**Chapter 13 Developing Real Use Cases: Data Plane Analytics 389**

- The Data 390
- SME Analysis 394
- SME Port Clustering 407
- Machine Learning: Creating Full Port Profiles 413
- Machine Learning: Creating Source Port Profiles 419
- Asset Discovery 422
- Investigation Task List 423
- Summary 424

**Chapter 14 Cisco Analytics 425**

- Architecture and Advisory Services for Analytics 426
- Stealthwatch 427
- Digital Network Architecture (DNA) 428
- AppDynamics 428
- Tetration 430
- Crosswork Automation 431
- IoT Analytics 432
- Analytics Platforms and Partnerships 433
- Cisco Open Source Platform 433
- Summary 434

**Chapter 15 Book Summary 435**

Analytics Introduction and Methodology 436

All About Networking Data 438

Using Bias and Innovation to Discover Solutions 439

Analytics Use Cases and Algorithms 439

Building Real Analytics Use Cases 440

Cisco Services and Solutions 442

In Closing 442

**Appendix A Function for Parsing Packets from pcap Files 443**

**Index 445**

## Icons Used in This Book



Cloud-line



Laptop



Router



Relational  
Database

## Command Syntax Conventions

The conventions used to present command syntax in this book are the same conventions used in the IOS Command Reference. The Command Reference describes these conventions as follows:

- **Boldface** indicates commands and keywords that are entered literally as shown. In actual configuration examples and output (not general command syntax), boldface indicates commands that are manually input by the user (such as a **show** command).
- *Italic* indicates arguments for which you supply actual values.
- Vertical bars (|) separate alternative, mutually exclusive elements.
- Square brackets ([ ]) indicate an optional element.
- Braces ({ }) indicate a required choice.
- Braces within brackets ([{ }]) indicate a required choice within an optional element.



## Foreword

What's the future of network engineers? This is a question haunting many of us. In the past, it was somewhat easy; study for your networking certification, have the CCIE or CCDE as the ultimate goal, and your future was secured.

In my job as a General Manager within the Cisco Professional Services organization, working with Fortune 1000 clients from around the world, I meet a lot of people with opinions in this matter, with views ranging from “we just need software programmers in the future” to “data scientist is the way to go as we will automate everything.” Is either of these views correct?

My simple answer to this is, “no,” the long answer is a little more complicated.

The changes in the networking industry are to a large extent the same as the automotive industry; today most cars are computerized. Imagine though, if a car was built by people that only knew software programming, and didn't know anything about the car design, the engine, or security. The “architect” of a car needs to be an in-depth expert on car design, and at the same time know enough about software capabilities, and what can be achieved, in a way that still keeps the “soul” of the car and enhances the overall result.

When it comes to the future of networking, it is very much the same. If we replaced skilled network engineers with data science engineers, the result would be mediocre. At the same time, there is no doubt that the future of networking will be built on data science.

In my view, the ideal structure of any IT team is a core of very knowledgeable network engineers, working very closely together with skilled data scientists. The network engineers that take the time to learn the basics of data science, and start to expand into that area will automatically be the bridge to the data science, and these engineers will soon become the most critical asset in that IT department.

The author of this book, John Garrett, is a true example of someone that has made this journey. With many years of experience working with the largest Cisco clients around the world, as one of our more senior network and data center technical leads, John saw the movement of data science approaching, and decided to invest himself in learning this new discipline. I would say he did not only learn it but instead mastered the art.

In this book, John helps the reader along the journey of learning data analytics in a very practical and applied way, providing the tools to almost immediately provide value to your organization.

At the end of the day, career progress is very linked to providing unique value. If you have decided to invest in yourself, and build data science skills on top of your telecommunication, datacenter, security, or IT knowledge, this book is the perfect start.

I would argue that John is a proof point to this matter, moving from a tech lead consultant to now being part of a small core team focusing on innovation to create the future of professional services from Cisco. A confirmation of this is also the number of patent submissions that John has pending in the area, as networking skills combined with data science opened up entirely new avenues of capabilities and solutions.

By Ulf Vinneras, Cisco General Manager Customer Experience/Cross Architecture

## Introduction: Your future is in your hands!

Analytics and data science are everywhere. Everything today is connected by networks. In the past networking and data science were distinct career paths, but this is no longer the case. Network and information technology (IT) specialists can benefit from understanding analytics, and data scientists can benefit from understanding how computer networks operate and produce data. People in both roles are responsible for building analytics solutions and use cases that improve the business.

This book provides the following:

- An introduction to data science methodologies and algorithms for network and IT professionals
- An understanding of computer network data that is available from these networks for data scientists
- Techniques for uncovering innovative use cases that combine the data science algorithms with network data
- Hands-on use-case development in Python and deep exploration of how to combine the networking data and data science techniques to find meaningful insights

After reading this book, data scientists will experience more success interacting with IT networking experts, and IT networking experts will be able to aid in developing complete analytics solutions. Experts from either area will learn how to develop networking use cases independently.

## My Story

I am a network engineer by trade. Prior to learning anything about analytics, I was an engineer working in data networking. Thanks to my many years of experience, I could design most network architectures that used any electronics to move any kind of data—business critical or not—in support of world-class applications. I thought I knew everything I needed to know about networking.

Then digital transformation happened. The software revolution happened. Everything went software defined. Everything is “virtual” and “containerized” now. Analytics is everywhere. With all these changes, I found that I didn’t know as much as I once thought I did.

If this sounds like your story, then you have enough experience to realize that you need to understand the next big thing if you want to remain relevant in a networking-related role—and analytics applied in your networking domain of expertise is the next big thing for you. If yours is like many organizations today, you have tons of data, and you have analytics tools and software to dive into it, but you just do not really know what to do with it. How can your skills be relevant here? How do you make the connection from these buckets, pockets, and piles of data to solving problems for your company? How

can you develop use cases that solve both business and technical problems? Which use cases provide some real value, and which ones are a waste of your time?

Looking for that next big thing was exactly the situation I found myself in about 10 years ago. I was experienced when it came to network design. I was a 5 year CCIE, and I had transitioned my skill set from campus design to wireless to the data center. I was working in one of the forward-looking areas of Cisco Services, Cisco Advanced Services. One of our many charters was “proactive customer support,” with a goal of helping customers avoid costly outages and downtime by preventing problems from happening in the first place. While it was not called *analytics* back then, the work done by Cisco Advanced Services could fall into a bucket known today as *prescriptive analytics*.

If you are an engineer looking for that next step in your career, many of my experiences will resonate with you. Many years ago, I was a senior technical practitioner deciding what was next for developing my skill set. My son was taking Cisco networking classes in high school, and the writing was on the wall that being only a network engineer was not going to be a viable alternative in the long term. I needed to level up my skills in order to maintain a senior-level position in a networking-related field, or I was looking at a role change or a career change in the future.

Why analytics? I was learning through my many customer interactions that we needed to do more with the data and expertise that we had in Cisco Services. The domain of coverage in networking was small enough back then that you could identify where things were “just not right” based on experience and intuition. At Cisco, we know how to use our collected data, our knowledge about data on existing systems, and our intuition to develop “mental models” that we regularly apply to our customer network environments.

What are mental models? Captain Sully on US Airways flight 1549 used mental models when he made an emergency landing on the Hudson River in 2009. Given all of the airplane telemetry data, Captain Sully knew best what he needed to do in order to land the plane safely and protect the lives of hundreds of passengers. Like experienced airplane pilots, experienced network engineers like you know how to avoid catastrophic failures. Mental models are powerful, and in this book, I tell you how to use mental models and innovation techniques to develop insightful analytics use cases for the networking domain.

The Services teams at Cisco had excellent collection and reporting. Expert analysis in the middle was our secret sauce. In many cases, the anonymized data from these systems became feeds to our internal tools that we developed as “digital implementations” of our mental models. We built awesome collection mechanisms, data repositories, proprietary rule-matching systems, machine reasoning systems, and automated reporting that we could use to summarize all the data in our findings for Cisco Services customers. We were finding insights but not actively looking for them using analytics and machine learning.

My primary interest as a futurist thinker was seeking to understand what was coming next for Cisco Advanced Services and myself. What was the “next big thing” for which we needed to be prepared? In this pursuit, I explored a wide array of new technology areas over the course of 10 years. I spent some years learning and designing VMware,

OpenStack, network functions virtualization (NFV), and the associated virtual network functions (VNFs) solutions on top of OpenStack. I then pivoted to analytics and applied those concepts to my virtualization knowledge area.

After several years working on this cutting edge of virtualized software infrastructure design and analytics, I learned that whether the infrastructure is physical or virtual, whether the applications are local or in the cloud, the importance of being able to find insights within the data that we get from our networking environments is critical to the success of these environments. I also learned that the growth of data science and the availability of computer resources to munge through the data make analytics and data science very attainable for any networking professional who wishes to pivot in this direction.

Given this insight, I spent 3 years of time outside work, including many evenings, weekends, and all of my available vacation time in order to earn a master's degree in predictive analytics from Northwestern University. Around that same time I began reading (or listening to) hundreds of books, articles, and papers about analytics topics. I also consumed interesting writings about algorithms, data science, innovation, innovative techniques, brain chemistry, bias, and other topics related to turning data into value by using creative thinking techniques. You are an engineer, so you can associate this to learning that next new platform, software, or architecture. You go all in.

Another driver for me was that I am work centered, driven to succeed, and competitive by nature. Maybe you are, too. My customers who had purchased Cisco services were challenging us to do better. It was no longer good enough to say that everything is connected, traffic is moving just fine across your network, and if there is a problem, the network protocols will heal themselves. Our customers wanted more than that.

Cisco Advanced Services customers are highly skilled, and they wanted more than simple reporting. They wanted visibility and insights across many domains. My customers wanted data, and they wanted dashboards that shared data with them so they could determine what was wrong on their own. One customer (we will call him Dave because that was his name) wanted to be able to use his own algorithms, his own machines, and his own people to determine what was happening at the lower levels of his infrastructure. He wanted to correlate this network data with his applications and his business metrics. For me, as a very senior network and data center engineer, I felt like I was not getting the job done. I could not do the analytics. I did not have a solution that I could propose for his purpose. There was a new space in networking that I had not yet conquered. Dave wanted actionable intelligence derived from the data that he was providing to Cisco. Dave wanted real analytics insights. Challenge accepted.

That was the start of my journey into analytics and into making the transition from being a network engineer to being a data scientist with enough ability to bridge the gap between IT networking engineers and those mathematical wizards who do the hard-core data science. This book is a knowledge share of what I have learned over the past years as I have transitioned from being an enterprise-focused campus, WAN, and data center networking engineer to being a learning data scientist. I realized that it was not necessary to get to the Ph.D. level to use data science and predictive analytics. For my transition,

I wanted to be someone who can use enough data science principles to find use cases in the wild and apply them to common IT networking problems to find useful, relevant, and actionable insights for my customers.

I hope you enjoy reading about what I have learned on this journey as much as I have enjoyed learning it. I am still working at it, so you will get the very latest. I hope that my learning and experiences in data, data science, innovation, and analytics use cases can help you in your career.

## How This Book Is Organized

Chapter 1, “Getting Started with Analytics,” defines some further details about what is explored in this book, as well as the current analytics landscape in the media. You cannot open your laptop or a social media application on your phone without seeing something related to analytics.

Chapter 2, “Approaches for Analytics and Data Science,” explores methodologies and approaches that will help you find success as a data scientist in your area of expertise. The simple models and diagrams that I have developed for internal Cisco trainings can help with your own solution framing activities.

Chapter 3, “Understanding Networking Data Sources,” begins by looking at network data and the planes of operation in networks that source this data. Virtualized solutions such as OpenStack and network functions virtualization (NFV) create additional complexities with sourcing data for analysis. Most network devices can perform multiple functions with the same hardware. This chapter will help you understand how they all fit together so you can get the right data for your solutions.

Chapter 4, “Accessing Data from Network Components,” introduces networking data details. Networking environments produce many different types of data, and there are multiple ways to get at it. This chapter provides overviews of the most common data access methods in networking. You cannot be a data scientist without data! If you are a seasoned networking engineer, you may only need to skim this chapter.

Chapter 5, “Mental Models and Cognitive Bias,” shifts gears toward innovation by spending time in the area of mental models, cognitive science, and bias. I am not a psychology expert or an authority in this space, but in this chapter I share common biases that you may experience in yourself, your users, and your stakeholders. This cognitive science is where things diverge from a standard networking book—but in a fascinating way. Understanding your audience is key to building successful use cases for them.

Chapter 6, “Innovative Thinking Techniques,” introduces innovative techniques and interesting tricks that I have used to uncover use cases in my role with Cisco. Understanding bias from Chapter 5 coupled with innovation techniques from this chapter will prepare you to maximize the benefit of the use cases and algorithms you learn in the upcoming chapters.

Chapter 7, “Analytics Use Cases and the Intuition Behind Them,” has you use your new knowledge of innovation to walk through analytics use cases across many industries. I have learned that combining the understanding of data with new and creative—and sometimes biased—thinking results in new understanding and new perspective.

Chapter 8, “Analytics Algorithms and the Intuition Behind Them,” walks through many common industry algorithms from the use cases in Chapter 7 and examines the intuition behind them. Whereas Chapter 7 looks at use cases from a top-down perspective, this chapter looks at algorithms to give you an inside-out view. If you know the problems you want to solve, this is your toolbox.

Chapter 9, “Building Analytics Use Cases,” brings back the models and methodologies from Chapter 2 and reviews how to turn your newfound ideas and algorithms into solutions. The use cases and data for the next four chapters are outlined here.

Chapter 10, “Developing Real Use Cases: The Power of Statistics,” moves from the abstract to the concrete and explores some real Cisco Services use cases built around statistics. There is still a very powerful role for statistics in our fancy data science world.

Chapter 11, “Developing Real Use Cases: Network Infrastructure Analytics,” looks at actual solutions that have been built using the feature information about your network infrastructure. A detailed look at Cisco Advanced Services fingerprinting, and other infrastructure-related capabilities is available here.

Chapter 12, “Developing Real Use Cases: Control Plane Analytics Using Syslog Telemetry,” shows how to build solutions that use network event telemetry data. The popularity of pushing data from devices is growing, and you can build use cases by using such data. Familiar algorithms from previous chapters are combined with new data in this chapter to provide new insight.

Chapter 13, “Developing Real Use Cases: Data Plane Analytics,” introduces solutions built for making sense of data plane traffic. This involves analysis of the packets flowing across your network devices. Familiar algorithms are used again to show how you can use the same analytics algorithms in many ways on many different types of data to find different insights.

Chapter 14, “Cisco Analytics,” runs through major Cisco product highlights in the analytics space. Any of these products can function as data collectors, sources, or engines, and they can provide you with additional analytics and visualization capabilities to use for solutions that extend the capabilities and base offerings of these platforms. Think of them as “starter kits” that help you get a working product in place that you can build on in the future.

Chapter 15, “Book Summary,” closes the book by providing a complete wrap-up of what I hope you learned as you read this book.

## Credits

- Stephen R. Covey, *The 7 Habits of Highly Effective People: Powerful Lessons in Personal Change*, 2004, Simon and Schuster.
- ITU Annual Regional Human Capacity Building Workshop for Sub-Saharan Countries in Africa Mauritius, 28–30 June 2017
- Empirical Model-Building and Response Surfaces*, 1987, George box, John Wiley.
- Predictably Irrational: The Hidden Forces that Shape Our Decisions*, Dan Ariely, HarperCollins.
- Thinking, Fast and Slow*, Daniel Kahneman, Macmillan Publishers
- Abraham Wald
- Thinking, Fast and Slow*, Daniel Kahneman, Macmillan Publishers
- Thinking, Fast and Slow*, Daniel Kahneman, Macmillan Publishers
- Thinking, Fast and Slow*, Daniel Kahneman, Macmillan Publishers
- Charles Duhigg
- De, B. E. (1985). Six thinking hats. Boston: Little, Browne and Company.
- Henry Ford
- Ries, E. (2011). *The lean startup: How constant innovation to creates radically successful businesses*. Penguin Books
- The Post-Algorithmic Era Has Arrived By Bill Franks*, Dec 14, 2017.

## Figure Credits

- Figure 8-13 Scikit-learn
- Figure 8-32 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 8-33 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 8-34 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 10-07 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 10-08 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 10-18 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 10-22 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 10-23 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 10-24 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 10-26 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 10-27 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 10-30 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 10-31 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 10-32 Screenshot of Jupyter Notebook © 2018 Project Jupyter

- Figure 10-34 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 10-37 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 10-38 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 10-39 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 10-40 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 10-47 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 10-49 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 10-51 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 10-53 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 10-54 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 10-61 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 10-62 Screenshot of Excel © Microsoft
- Figure 11-22 Screenshot of Business Critical Insights © 2018 Cisco Systems, Inc.
- Figure 11-32 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 11-34 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 11-38 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 11-41 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 11-51 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 13-10 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 13-12 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 13-13 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 13-14 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 13-15 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 13-35 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 12-03 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 12-04 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 12-05 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 12-07 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 12-08 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 12-09 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 12-10 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 12-11 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 12-12 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 12-15 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 12-18 Screenshot of Jupyter Notebook © 2018 Project Jupyter
- Figure 12-42 Screenshot of Jupyter Notebook © 2018 Project Jupyter



# Approaches for Analytics and Data Science

This chapter examines a simple methodology and approach for developing analytics solutions. When I first started analyzing networking data, I used many spreadsheets, and I had a lot of data access, but I did not have a good methodology to approach the problems. You can only sort, filter, pivot, and script so much when working with a single data set in a spreadsheet. You can spend hours, days, or weeks diving into the data, slicing and dicing, pivoting this way and that...only to find that the best you can do is show the biggest and the smallest data points. You end up with no real insights. When you share your findings to glassy-eyed managers, the rows and columns of data are a lot more interesting to you than they are to them. I have learned through experience that you need more.

Analytics solutions look at data to uncover stories about what is happening now or what will be happening in the future. In order to be effective in a data science role, you must step up your storytelling game. You can show the same results in different ways—sometimes many different ways—and to be successful, you must get the audience to see what you are seeing. As you will learn in Chapter 5, “Mental Models and Cognitive Bias,” people have biases that impact how they receive your results, and you need to find a way to make your results relevant to each of them—or at least make your results relevant to the stakeholders who matter.

You have two tasks here. First, you need to find a way to make your findings interesting to nontechnical people. You can make data more interesting to nontechnical people with statistics, top-*n* reporting, visualization, and a good storyline. I always call this the “BI/BA of analytics,” or the simple descriptive analytics. Business intelligence (BI)/business analytics (BA) dashboards are a useful form of data presentation, but they typically rely on the viewer to find insight. This has value and is useful to some extent but generally tops out at cool visualizations that I call “*Sesame Street* analytics.”

If you are from my era, you grew up with the *Sesame Street* PBS show, which had a segment that taught children to recognize differences in images and had the musical tagline “One of these things is not like the others.” Visualizations with anomalies identified in

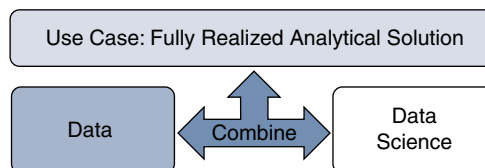
contrasting colors immediately help the audience see how “one of these things is not like the others,” and you do not need a story if you have shown this properly. People look at your visualization or infographic and just see it.

Your second task is to make the data interesting to the technical people, your new data science friends, your peers. You do this with models and analytics, and your visualizing and storytelling must be at a completely new level. If you present “*Sesame Street* analytics” to a technical audience, you are likely to hear “That’s just visualization; I want to know *why* is it an outlier.” You need to do more—with real algorithms and analytics—to impress this audience. This chapter starts your journey toward impressing both audiences.

## Model Building and Model Deployment

As mentioned in Chapter 1, “Getting Started with Analytics,” when it comes to analytics models, people often overlook a very important distinction between *developing and building* and *implementing and deploying* models. The ability for your model to be usable outside your own computer is a critical success factor, and you need to know how to both build and deploy your analytics use cases. It is often the case that you build models centrally then deploy them at the edge of a network or at many edges of corporate or service provider networks. Where do you think the speech recognition models on your mobile phone were built? Where are they ultimately deployed? If your model is going to have impact in your organization, you need to develop workflows that use your model to benefit the business in some tangible way.

Many models are developed or built from batches of test data, perhaps with data from a lab or a big data cluster, built on users’ machines or inside an analytics package of data science algorithms. This data is readily available, cleaned, and standardized, and they have no missing values. Experienced data science people can easily run through a bunch of algorithms to visualize and analyze the data in different ways to glean new and interesting findings. With this captive data, you can sometimes run through hundreds of algorithms with different parameters, treating your model like a black box, and only viewing the results. Sometimes you get very cool-looking results that are relevant. In the eyes of management or people who do not understand the challenges in data science, such development activity looks like the simple layout in Figure 2-1, where data is simply combined with data science to develop a solution. Say hello to your nontechnical audience. This is not a disparaging remark; some people—maybe even most people—prefer to just get to the point, and nothing gets to the point better than results. These people do not care about the details that you needed to learn in order to provide solutions at this level of simplicity.



**Figure 2-1** *Simplified View of Data Science*

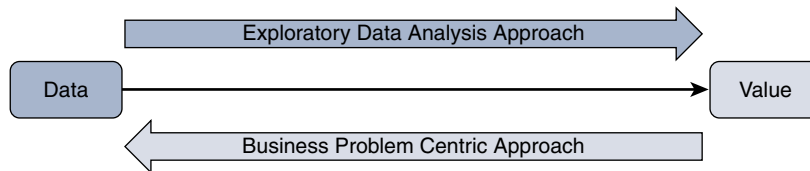
Once you find a model, you bring in more data to further test and validate that the model's findings are useful. You need to prove beyond any reasonable doubt that the model you have on your laptop shows value. Fantastic. Then what? How can you bring all data across your company to your computer so that you can run it through the model you built?

At some point in the process, you will deploy your analytics to a production system, with real data, meaning that an automated system is set up to run new data, in batches or streaming, against your new model. This often involves working with a development team, whose members may or may not be experts in analytics. In some cases, you do not need to deploy into production at all because the insight is learned, and no further understanding is required. In either case, you then need to use your model against new batches of data to extend the value beyond the data you originally used to build and test it.

Because I am often the one with models on my computer, and I have learned how to deploy those models as part of useful applications, I share my experiences in turning models into useful tools in later chapters of this book, as we go through actual use cases.

## Analytics Methodology and Approach

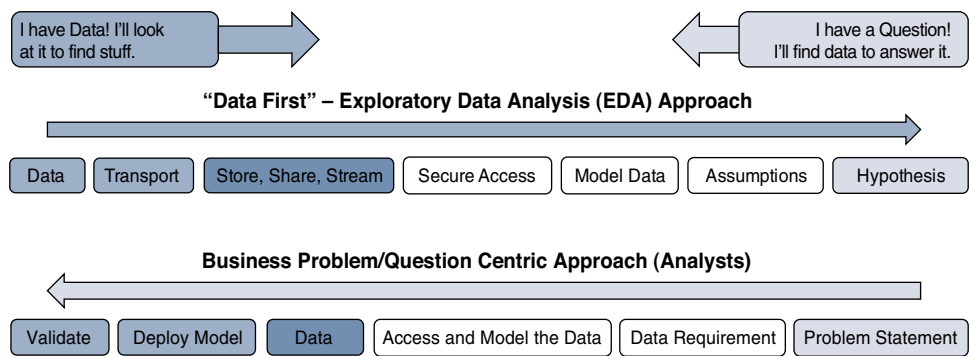
How you approach an analytics problem is one of the factors that determine how successful your solution will be in solving the problem. In the case of analytics problems, you can use two broad approaches, or methodologies, to get to insightful solutions. Depending on your background, you will have some predetermined bias in terms of how you want to approach problems. The ultimate goal is to convert data to value for your company. You get to that value by finding insights that solve technical or business problems. The two broad approaches, shown in Figure 2-2, are the “explore the data” approach, and the “solve the business problem” approach.



**Figure 2-2** *Two Approaches to Developing Analytics Solutions*

These are the two main approaches that I use, and there is literature about many granular, systematic methodologies that support some variation of each of these approaches. Most analytics literature guides you to the problem-centric approach. If you are strongly aware of the data that you have but not sure how to use it to solve problems, you may find yourself starting in the statistically centered exploratory data analysis (EDA) space that is most closely associated with statistician John Tukey. This approach often has some quick wins along the way in finding statistical value in the data rollups and visualizations used to explore the data.

Most domain data experts tend to start with EDA because it helps you understand the data and get the quick wins that allow you to throw a bone to the stakeholders while digging into the more time-consuming part of the analysis. Your stakeholders often have hypotheses (and some biases) related to the data. Early findings from this side often sound like “You can see that issue X is highly correlated with condition Y in the environment; therefore, you should address condition Y to reduce the number of times you see issue X.” Most of my early successes in developing tools and applications for Cisco Advanced Services were absolutely data first and based on statistical findings instead of analytics models. There were no heavy algorithms involved, there was no machine learning, and there was no real data science. Sometimes, statistics are just as effective at telling interesting stories. Figure 2-3 shows how to view these processes as a comparison. There is no right or wrong side on which to start; depending on your analysis goals, either direction or approach is valid. Note that this model includes data acquisition, data transport, data storage, sharing, or streaming, and secure access to that data, all of which are things to consider if the model is to be implemented on a production data flow—or “operation-alized.” The previous, simpler model that shows a simple data and data science combination (refer to Figure 2-1) still applies for exploring a static data set or stream that you can play back and analyze using offline tools.



**Figure 2-3** *Exploratory Data Versus Problem Approach Comparison*

### Common Approach Walkthrough

While many believe that analytics is done only by math PhDs and statisticians, general analysts and industry subject matter experts (SMEs) now commonly use software to explore, predict, and preempt business and technical problems in their areas of expertise. You and other “citizen data scientists” can use a variety of software packages available today to find interesting insights and build useful models. You can start from either side when you understand the validity of both approaches. The important thing to understand is that many of the people you work with may be starting at the other end of the spectrum, and you need to be aware of this as you start sharing your insights with a wider audience. When either audience asks, “What problem does this solve for us?” you can present relevant findings.

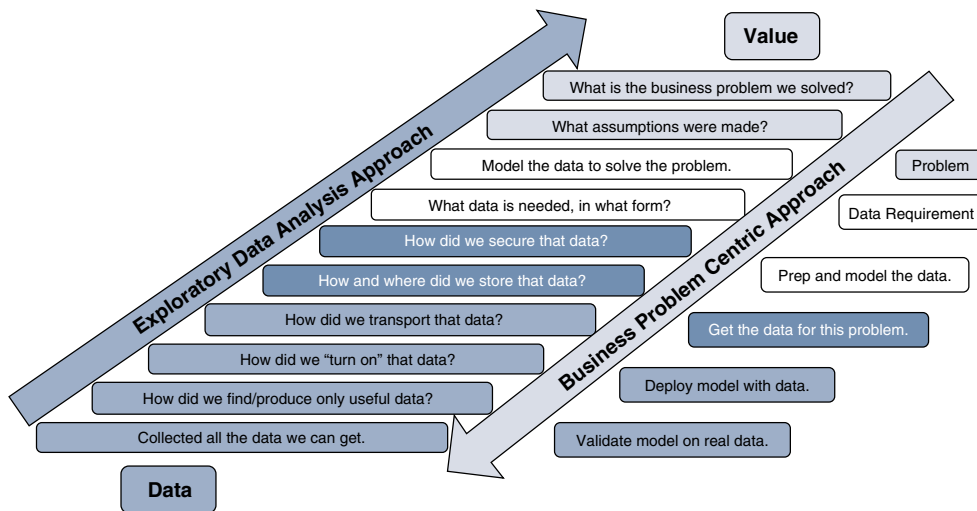
Let's begin on the data side. During model *building*, you skip over the transport, store, and secure phases as you grab a batch of useful data, based on your assumptions, and try to test some hypothesis about it. Perhaps through some grouping and clustering of your trouble ticket data, you have seen excessive issues on your network routers with some specific version of software. In this case, you can create an analysis that proves your hypothesis that the problems are indeed related to the version of software that is running on the suspect network routers. For the data first approach, you need to determine the problems you want to solve, and you are also using the data to guide you to what is possible, given your knowledge of the environment.

What do you need in this suspect routers example? Obviously, you must get data about the network routers when they showed the issue, as well as data about the same types of routers that have not had the issue. You need both of these types of information in order to find the underlying factors that may or may not have contributed to the issue you are researching. Finding these factors is a form of inference, as you would like to infer something about all of your routers, based on comparisons of differences in a set of devices that exhibit the issue and a set of devices that do not. You will later use the same analytics model for prediction.

You can commonly skip the “production data” acquisition and transport parts of the model building phase. Although in this case you have a data set to work with for your analysis, consider here how to automate the acquisition of data, how to transport it, and where it will live if you plan to put your model into a fully automated production state so it can notify you of devices in the network that meet these criteria. On the other hand, full production state is not always necessary. Sometimes you can just grab a batch of data and run it against something on your own machine to find insights; this is valid and common. Sometimes you can collect enough data about a problem to solve that problem, and you can gain insight without having to implement a full production system.

Starting at the other end of this spectrum, a common analyst approach is to start with a known problem and figure out what data is required to solve that problem. You often need to seek things that you don't know to look for. Consider this example: Perhaps you have customers with service-level agreements (SLAs), and you find that you are giving them discounts because they are having voice issues over the network and you are not meeting the SLAs. This is costing your company money. You research what you need to analyze in order to understand why this happens, perhaps using voice drop and latency data from your environment. When you finally get these data, you build a proposed model that identifies that higher latency with specific versions of software on network routers is common on devices in the network path for customers who are asking for refunds. Then you deploy the model to flag these “SLA suckers” in your production systems and then validate that the model is effective as the SLA issues have gone away. In this case, *deploy* means that your model is watching your daily inventory data and looking for a device that matches the parameters that you have seen are problematic. What may have been a very complex model has a simple deployment.

Whether starting at data or at a business problem, ultimately solving the problem represents the value to your company and to you as an analyst. Both of these approaches follow many of the same steps on the analytics journey, but they often use different terminology. They are both about turning data into value, regardless of starting point, direction, or approach. Figure 2-4 provides a more detailed perspective that illustrates that these two approaches can work in the same environment on the same data and the very same problem statement. Simply put, all of the work and due diligence needs to be done to have a fully operational (with models built, tested, and deployed), end-to-end use case that provides real, continuous value.



**Figure 2-4** Detailed Comparison of Data Versus Problem Approaches

There are a wide variety of detailed approaches and frameworks available in industry today, such as CRISP-DM (cross-industry standard process for data mining) and SEMMA (Sample Explore, Modify, Model, and Assess), and they all generally follow these same principles. Pick something that fits your style and roll with it. Regardless of your approach, the primary goal is to create useful solutions in your problem space by combining the data you have with data science techniques to develop use cases that bring insights to the forefront.

## Distinction Between the Use Case and the Solution

Let's slow down a bit and clarify a few terms. Basically, a *use case* is simply a description of a problem that you solve by combining data and data science and applying analytics. The underlying algorithms and models comprise the actual analytics solution. In the case of Amazon, for example, the use case is getting you to spend more money. Amazon does this by showing you what other people have also bought in addition to buying the same item that are purchasing. The intuition behind this is that you will buy more things because

other people like you needed those things when they purchased the same item that you did. The model is there to uncover that and remind you that you may also need to purchase those other things. Very helpful, right?

From the exploratory data approach, Amazon might want to do something with the data it has about what people are buying online. It can then collect the high patterns of common sets of purchases. Then, for patterns that are close but missing just a few items, Amazon may assume that those people just “forgot” to purchase something they needed because everyone else purchased the entire “item set” found in the data. Amazon might then use software implementation to find the people who “forgot” and remind them that they might need the other common items. Then Amazon can validate the effectiveness by tracking purchases of items that the model suggested.

From a business problem approach, Amazon might look at wanting to increase sales, and it might assume (or find research which suggests) that, if reminded, people often purchase common companion items to what they are currently viewing or have in their shopping carts. In order to implement this, Amazon might collect buying pattern data to determine these companion items. The company might then suggest that people may also want to purchase these items. Amazon can then validate the effectiveness by tracking purchases of suggested items.

Do you see how both of these approaches reach the same final solution?

The Amazon case is about increasing sales of items. In predictive analytics, the use case may be about predicting home values or car values. More simply, the use case may be the ability to predict a continuous number from historical numbers. No matter the use case, you can view analytics as simply the application of data and data science to the problem domain. You can choose how you approach finding and building the solutions either by using the data as a guide or by dissecting the stated problem.

## Logical Models for Data Science and Data

This section discusses analytics solutions that you model and build for the purpose of deployment to your environment. When I was working with Cisco customers in the early days of analytics, it became clear that setting up the entire data and data science pipeline as a working application on a production network was a bit confusing to many customers, as well as to traditional Cisco engineers.

Many customers thought that they could simply buy network analytics software and install it onto the network as they would any other application—and they would have fully insightful analytics. This, of course, is not the case. Analytics packages integrate into the very same networks for which you build models to run. We can use this situation to introduce the concept of an *overlay*, which is a very important concept for understanding network data (covered in Chapter 3, “Understanding Networking Data Sources”). Analytics packages installed on computers that sit on networks can *build* the models as discussed earlier, but when it is time to *deploy* the models that include data feeds from network environments, the analytics packages often have tendrils that reach deep into

the network and IT systems. Further, these solutions can interface with business and customer data systems that exist elsewhere in the network. Designing such a system can be daunting because most applications on a network do not interact with the underlying hardware. A second important term you should understand is the *underlay*.

## **Analytics as an Overlay**

So how do data and analytics applications fit within network architectures? In this context, you need to know the systems and software that consume the data, and you need to use data science to provide solutions as general applications. If you are using some data science packages or platforms today, then this idea should be familiar to you. These applications take data from the infrastructure (perhaps through a central data store) and combine it with other applications data from systems that reside within the IT infrastructure.

This means the solution is analyzing the very same infrastructure in which it resides, along with a whole host of other applications. In networking, an *overlay* is a solution that is abstracted from the underlying physical infrastructure in some way. Networking purists may not use the term *overlay* for applications, but it is used here because it is an important distinction needed to set up the data discussion in the next chapter. Your model, when implemented in production on a live network, is just an overlay instance of an application, much like other overlay application instances riding on the same network.

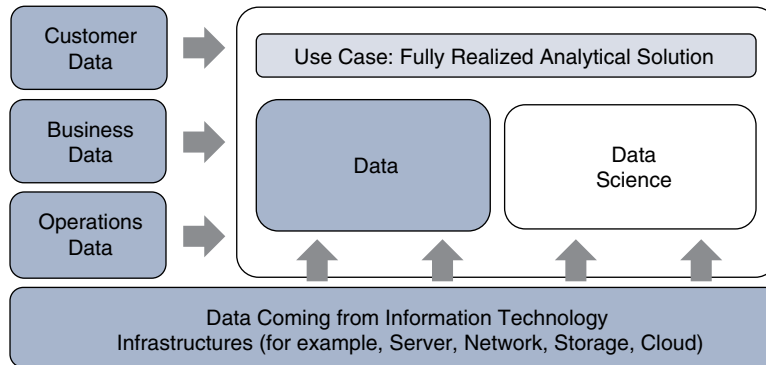
This concept of network layers and overlay/underlay is why networking is often blamed for fault or outage—because the network underlays all applications (and other network instances, as discussed in the next chapter). Most applications, if looked at from an application-centric view, are simply overlays onto the underlying network infrastructure. New networking solutions such as Cisco Application Centric Infrastructure (ACI) and common software-defined wide area networks (SD-WANs) such as Cisco iWAN+Viptela take overlay networking to a completely new level by adding additional layers of policy and network segmentation. In case you have not yet surmised, you probably should have a rock-solid underlay network if you want to run all these overlay applications, virtual private networks (VPNs), and analytics solutions on it.

Let's look at an example here to explain overlays. Consider your very own driving patterns (or walking patterns, if you are urban) and the roads or infrastructure that you use to get around. You are one overlay on the world around you. Your neighbor traveling is another overlay. Perhaps your overlay is “going to work,” and your neighbor's overlay for the day is “going shopping.” You are both using the same infrastructure but doing your own things, based on your interactions with the underlay (walkways, roads, bridges, home, offices, stores, and anything else that you interact with). Each of us is an individual “instance” using the underlay, much as applications are instances on networks. There could be hundreds or even thousands of these applications—or millions of people using the roadway system. The underlay itself has lots of possible “layers,” such as the physical roads and intersections and the controls such as signs and lights. Unseen to you, and therefore “virtual,” is probably some satellite layer where GPS is making decisions about how another application overlay (a delivery truck) should be using the underlay (roads).



This concept of overlays and layers, both physical and virtual, for applications as well as networks, was a big epiphany for me when I finally got it. The very networks themselves have layers and planes of operations. I recall it just clicking one day that the packets (routing protocol packets) that were being used to “set up” packet forwarding for a path in my network were using the same infrastructure that they were actually setting up. That is like me controlling the stoplights and walk signs as I go to work, while I am trying to get there. We’ll talk more about this “control plane” later. For now, let’s focus on what is involved with an analytics infrastructure overlay model.

By now, I hope that I have convinced you that this concept of some virtual overlay of functionality on a physical set of gear is very common in networking today. Let’s now look at an analytics infrastructure overlay diagram to illustrate that the data and data science come together to form the use cases of always-on models running in your IT environment. Note in Figure 2-5 how other data, such as customer, business, or operations data, is exported from other application overlays and imported into yours.



**Figure 2-5** *Analytics Solution Overlay*

In today’s digital environment, consider that all the data you need for analysis is produced by some system that is reachable through a network. Since everyone is connected, this is the very same network where you will use some system to collect and store this data. You will most likely deploy your favorite data science tools on this network as well. Your role as the analytics expert here is to make sure you identify how this is set up, such that you successfully set up the data sources that you need to build your analytics use case. You must ensure these data sources are available to the proper layer—your layer—of the network.

The concept of customer, business, and operations data may be new, so let’s get right to the key value. If you used analytics in your customer space, you know who your valuable customers are (and, conversely, which customers are more costly than they are worth). This adds context to findings from the network, as does the business context (which network components have the greatest impact) and operations (where you are spending

excessive time and money in the network). Bringing all these data together allows you to develop use cases with relevant context that will be noticed by business sponsors and stakeholders at higher levels in your company.

As mentioned earlier in this chapter, you can build a model with batches of data, but deploying an active model into your environment requires planning and setup of the data sources needed to “feed” your model as it runs every day in the environment. This may also include context data from other customer or business applications in the network environment. Once you have built a model and wish to operationalize it, making sure that everything properly feeds into your data pipelines is crucial—including the customer, business, operations, and other applications data.

Analytics Infrastructure Model

This section moves away from the overlays and network data to focus entirely on building an analytics solution. (We revisit the concepts of layers and overlays in the next chapter, when we dive deeper into the data sources in the networking domain.) In the case of IT networking, there are many types of deep technical data sources coming up from the environment, and you may need to combine them with data coming from business or operations systems in a common environment in order to provide relevance to the business. You use this data in the data science space with maturity levels of usage, as discussed in Chapter 1. So how can you think about data that is just “out there in the ether” in such a way that you can get to actual analytics use cases? All this is data that you define or create. This is just one component of a model that looks at the required data and components of the analytics use cases.

Figure 2-6 is a simple model for thinking about the flow of data for building deployable, operationalized models that provide analytics solutions. We can call this a simple model for analytics infrastructure, and, as shown in the figure, we can contrast this model with a problem-centric approach used by a traditional business analyst.

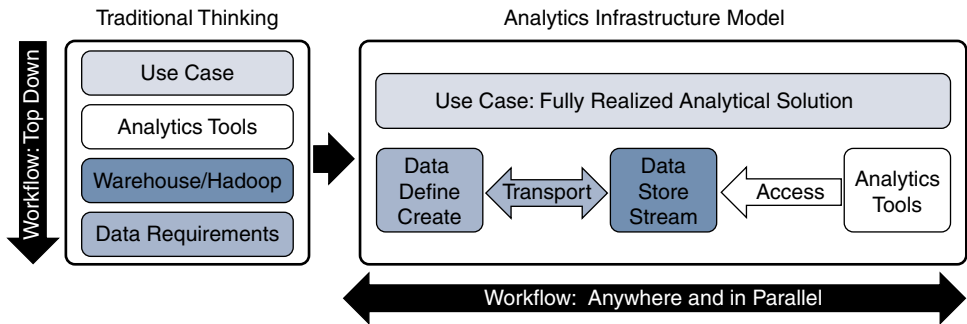


Figure 2-6 Traditional Analyst Thinking Versus Analytics Infrastructure Model

No, analytics infrastructure is not artificial intelligence. Due to the focus on the lower levels of infrastructure data for analytics usage, this analytics infrastructure name fits best. The goal is to identify how to build analytics solutions much the same way you have

built LAN, WAN, wireless, and data center network infrastructures for years. Assembling a full architecture to extract value from data to solve a business problem is an infrastructure in itself. This is very much like an end-to-end application design or an end-to-end networking design, but with a focus on analytics solutions only.

The analytics infrastructure model used in IT networking differs from traditional analyst thinking in that it involves always looking to build repeatable, reusable, flexible solutions and not just find a data requirement for a single problem. This means that once you set up a data source—perhaps from routers, switches, databases, third-party systems, network collectors, or network management systems—you want to use that data source for multiple applications. You may want to replicate that data pipeline across other components and devices so others in the company can use it. This is the “build once, use many” paradigm that is common in Cisco Services and in Cisco products. Solutions built on standard interfaces are connected together to form new solutions. These solutions are reused as many times as needed. Analytics infrastructure model components can be used as many times as needed.

It is important to use standards-based data acquisition technologies and perhaps secure the transport and access around the central data cleansing, sharing, and storage of any networking data. This further ensures the reusability of your work for other solutions. Many such standard data acquisition techniques for the network layer are discussed in Chapter 4, “Accessing Data from Network Components.”

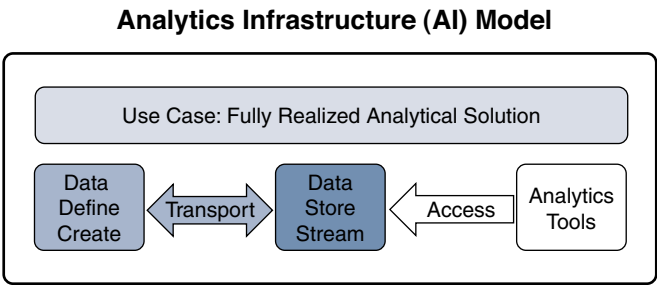
At the far right of the model in Figure 2-6, you want to use any data science tool or package you can to access and analyze your data to create new use cases. Perhaps one package builds a model that is implemented in code, and another package produces the data visualization to show what is happening. The components in the various parts of the model are pluggable so that parts (for example, a transport or a database) could be swapped out with suitable replacements. The role and functionality of a component, not the vendor or type, is what is important.

Finally, you want to be able to work this in an Agile manner and not depend on the top-down Waterfall methods used in traditional solution design. You can work in parallel in any sections of this analytics infrastructure model to help build out the components you need to enable in order to operationalize any analytics model onto any network infrastructure. When you have a team with different areas of expertise along the analytics infrastructure model components, the process is accelerated.

Later in the book, this model is referenced as an aid to solution building. The analytics infrastructure model is very much a generalized model, but it is open, flexible, and usable across many different job roles, both technical and nontechnical, and allows for discussion across silos of people with whom you need to interface. All components are equally important and should be used to aid in the design of analytics solutions.

The analytics infrastructure model (shown enlarged in in Figure 2-7) also differs from many traditional development models in that it segments functions by job roles, which allows for the aforementioned Agile parallel development work. Each of these job roles may still use specialized models within its own functions. For example, a data scientist

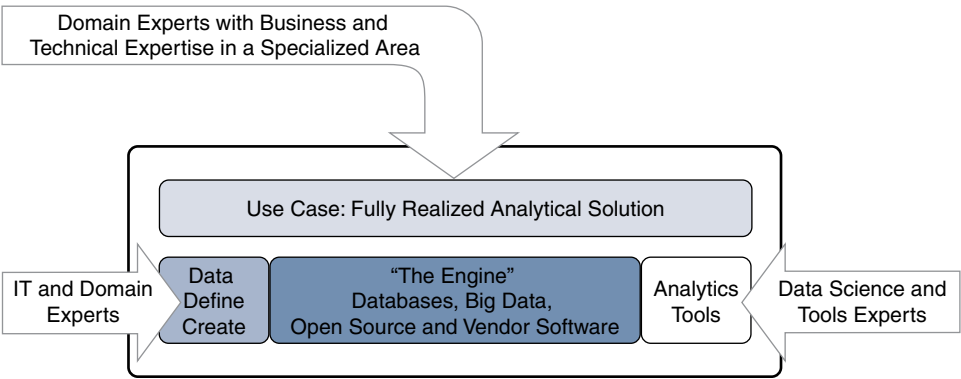
might use a preferred methodology and analytics tools to explore the data that you provided in the data storage location. As a networking professional, defining and creating data (far left) in your domain of expertise is where you play, and it is equally as important as the setup of the big data infrastructure (center of the model) or the analysis of the data using specialized tools and algorithms (far right).



**Figure 2-7** *Analytics Infrastructure Model for Developing Analytics Solutions*

Here is a simple elevator pitch for the analytics infrastructure model: “Data is defined, created, or produced in some system from which it is moved into a place where it is stored, shared, or streamed to interested users and data science consumers. Domain-specific solutions using data science tools, techniques, and methodologies provide the analysis and use cases from this data. A fully realized solution crosses all of the data, data storage, and data science components to deliver a use case that is relevant to the business.”

As mentioned in Chapter 1, this book spends little time on “the engine,” which is the center of this model, identified as the big data layer shown in Figure 2-8. When I refer to anything in this engine space, I call out the function, such as “store the data in a database” or “stream the data from the Kafka bus.” Due to the number of open source and commercial components and options in this space, there is an almost infinite combination of options and instructions readily available to build the capabilities that you need.



**Figure 2-8** *Roles and the Analytics Infrastructure Model*

It is not important that you understand how “the engine” in this car works; rather, it is important to ensure that you can use it to drive toward analytics solutions. Whether using open source big data infrastructure or packages from vendors in this space, you can readily find instructions to transport, store, share, and stream and provide access to the data on the Internet. Run a web search on “data engineering pipelines” and “big data architecture,” and you will find a vast array of information and literature in the data engineering space.

The book aims to help you understand the job roles around the common big data infrastructure, along with data, data science, and use cases. The following are some of the key roles you need to understand:

- **Data domain experts**—These experts are familiar with the data and data sources.
- **Analytics or business domain experts**—These experts are familiar with the problems that need to be solved (or questions that need to be answered).
- **Data scientists**—These experts have knowledge of the tools and techniques available to find the answers or insights desired by the business or technical experts in the company.

The analytics infrastructure model is location agnostic, which is why you see callouts for data transport and data access. This overall model approach applies regardless of technology or location. Analytics systems can be on-premises, in the cloud, or hybrid solutions, as long as all the parts are available for use. Regardless of where the analytics is used, the networking team is usually involved in ensuring that the data is in the right place for the analysis. Recall from the overlay discussion earlier in the chapter that the underlay is necessary for the overlay to work. Parts of this analysis may exist in the cloud, other parts on your laptop, and other parts on captive customer relationship management (CRM) systems on your corporate networks. You can use the analytics infrastructure model to diagram a solution flow that results in a fully realized analytics use case.

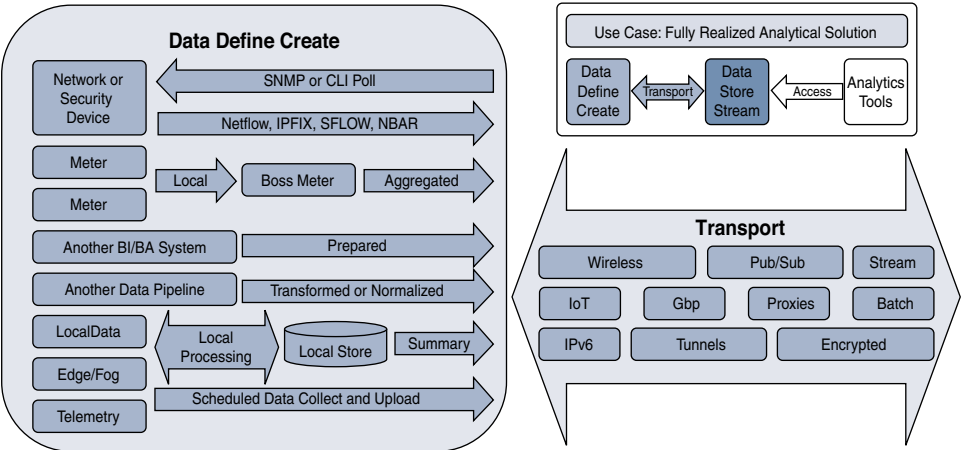
Depending on your primary role, you may be involved in gathering the data, moving the data, storing the data, sharing the data, streaming the data, archiving the data, or providing the analytics analysis. You may be ready to build the entire use case. There are many perspectives when discussing analytics solutions. Sometimes you will wear multiple hats. Sometimes you will work with many people; sometimes you will work alone if you have learned to fill all the required roles. If you decide to work alone, make sure you have access to resources or expertise to validate findings in areas that are new to you. You don’t want to spend a significant amount of time uncovering something that is already general knowledge and therefore not very useful to your stakeholders.

Building your components using the analytics infrastructure model ensures that you have reusable assets in each of the major parts of the model. Sometimes you will spend many hours, days, or weeks developing an analysis, only to find that there are no interesting insights. This is common in data science work. By using the analytics infrastructure model, you can maintain some parts of your work to build other solutions in the future.

The Analytics Infrastructure Model In Depth

So what are the “reusable and repeatable components” touted in the analytics infrastructure model? This section digs into the details of what needs to happen in each part of the model. Let’s start by digging into the lower-left data component of the model, looking at the data that is commonly available in an IT environment. Data pipelines are big business and well covered in the “for fee” and free literature.

Building analytics models usually involves getting and modeling some data from the infrastructure, which includes spending a lot of time on research, data munging, data wrangling, data cleansing, ETL (Extract, Transform, Load), and other tasks. The true power of what you build is realized when you deploy your model into an environment and turn it on. As the analytics infrastructure model indicates, this involves acquiring useful data and transporting it into an accessible place. What are some examples of the data that you may need to acquire? Expanding on the data and transport sections of the model in Figure 2-9, you will find many familiar terms related to the combination of networking and data.

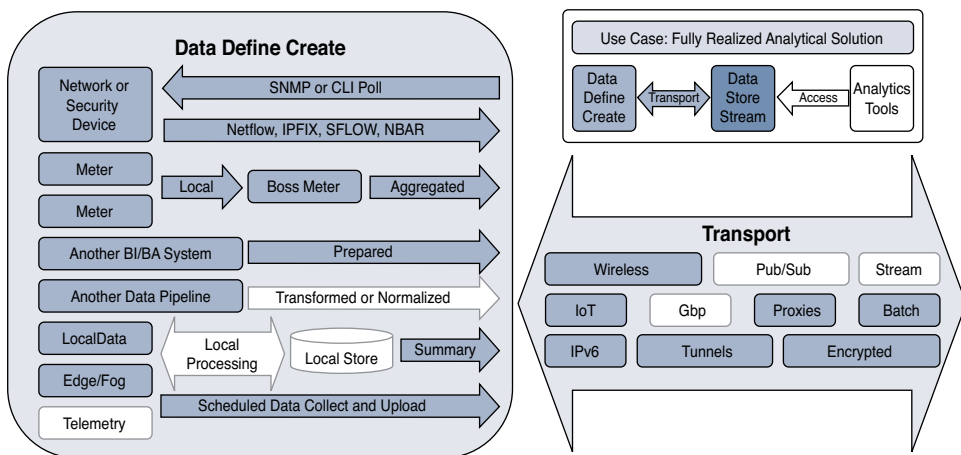


**Figure 2-9** *Analytics Infrastructure Model Data and Transport Examples*

Implementing a model involves setting up a full pipeline of new data (or reusing a part of a previous pipeline) to run through your newly modeled use cases, and this involves “turning on” the right data and transporting it to where you need it to be. Sometimes this is kept local (as in the case of many Internet of Things [IoT] solutions), and sometimes data needs to be transported. This is all part of setting up the full data pipeline. If you need to examine data in flight for some real-time analysis, you may need to have full data streaming capabilities built from the data source to the place where the analysis happens.

Do not let the number of words in Figure 2-9 scare you; not all of these things are used. This diagram simply shares some possibilities and is in no way a complete set of everything that could be at each layer.

To illustrate how this model works, let's return to the earlier example of the router problem. If latency and sometimes router crashes are associated with a memory leak in some software versions of a network router, you can use a telemetry data source to access memory statistics in a router. Telemetry data, covered in Chapter 4, is a push model whereby network devices send periodic or triggered updates to a specified location in the analytics solution overlay. Telemetry is like a hospital heart monitor that gets constant updates from probes on a patient. Getting router memory-related telemetry data to the analytics layer involves using the components identified in white in Figure 2-10—for just a single stream. By setting this up for use, you create a reusable data pipeline with telemetry-supplied data. A new instance of this full pipeline must be set up for each device in the network that you want to analyze for this problem. The hard part—the “feature engineering” of building a pipeline—needs to happen only once. You can easily replicate and reuse that pipeline, as you now have your memory “heart rate monitor” set up for all devices that support telemetry. The left side of Figure 2-10 shows many ways data can originate, including methods and local data manipulations, and the arrow on the right side of the figure shows potential transport methods. There are many types of data sources and access methods.



**Figure 2-10** *Analytics Infrastructure Model Telemetry Data Example*

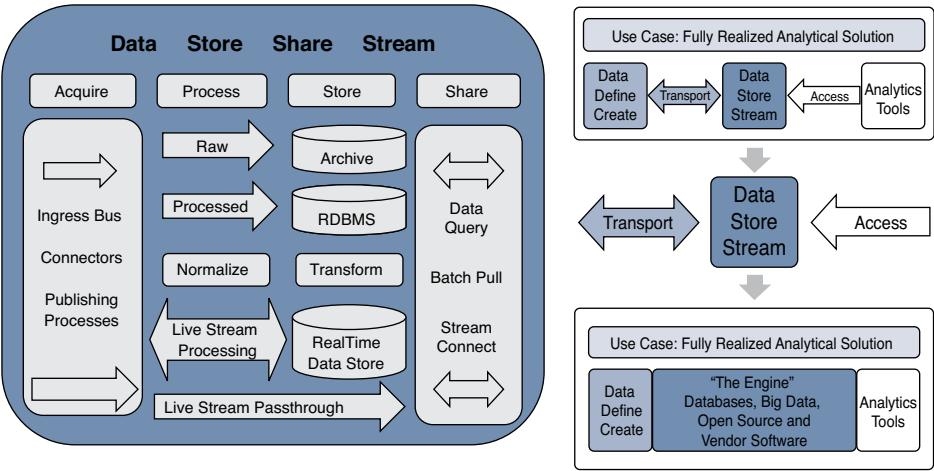
In this example, you are taking in telemetry data at the data layer, and you may also do some local processing of the data and store it in a localized database. In order to send the memory data upstream, you may standardize it to a megabyte or gigabyte number, standardize it to a “z” value, or perform some other transformation. This design work must happen once for each source. Does this data transformation and standardization stuff sound tedious to you? Consider that in 1999, NASA lost a \$125 million Mars orbiter due to a mismatch of metric to English units in the software. Standardization, transformation, and data design are important.

Now, assuming that you have the telemetry data you want, how do you send it to a storage location? You need to choose transport options. For this example, say that you choose to send a steady stream to a Kafka publisher/subscriber location by using Google Protocol Buffers (GPB) encoding. There are lots of capabilities, and lots of options, but after a one-time design, learning, and setup process, you can document it and use it over and over again. What happens when you need to check another router for this same memory leak? You call up the specification that you designed here and retrofit it for the new requirement.

While data platforms and data movement are not covered in detail in this book, it is important that you have a basic understanding of what is happening inside the engine, all around the “the data platform.”

The Analytics Engine

Unless you have a dedicated team to do this, much of this data storage work and setup may fall in your lap during model building. You can find a wealth of instruction for building your own data environments by doing a simple Internet search. Figure 2-11 shows many of the activities related to this layer. Note how the transport and data access relate to the configuration of this centralized engine. You need a destination for your prepared data, and you need to know the central location configuration so you can send it there. On the access side, the central data location will have access methods and security, which you must know or design in order to consume data from this layer.



**Figure 2-11** *The Analytics Infrastructure Model Data Engine*

Once you have defined the data parameters, and you understand where to send the data, you can move the data into the engine for storage, analysis, and streaming. From each individual source perspective, the choice comes down to push or pull mechanisms, as



per the component capabilities available to you in your data-producing entities. This may include pull methods using polling protocols such as Simple Network Management Protocol (SNMP) or push methods such as the telemetry used in this example.

This centralized data-engineering environment is where the Hadoop, Spark, or commercial big data platform lives. Such platforms are often set up with receivers for each individual type of data. The pipeline definition for each of these types of data includes the type and configuration of this receiver at the central data environment. Very common within analytics engines today is something called a publisher/subscriber environment, or “pub/sub” bus. Apache Kafka is a very common bus used in these engines today.

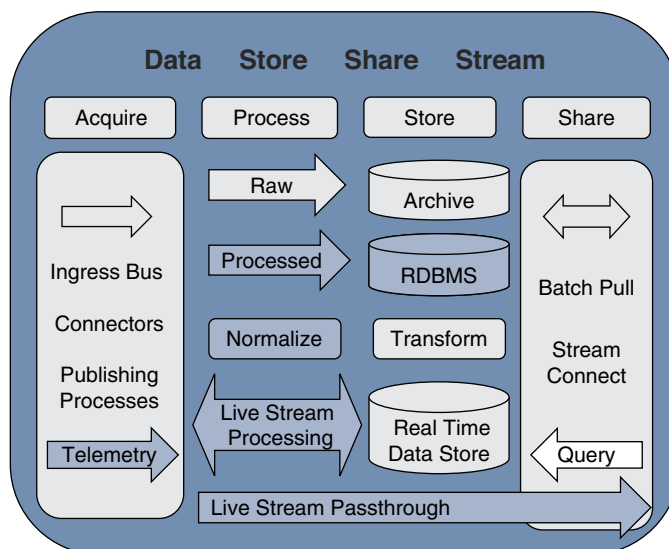
A good analogy for the pub/sub bus is broadcast TV channels with a DVR. Data feeds (through analytics infrastructure model transports) are sent to specific channels from data producers, and subscribers (data consumers) can choose to listen to these data feeds and subscribe (using some analytics infrastructure model access method, such as a Kafka consumer) to receive them. In this telemetry example, the telemetry receiver takes interesting data and copies or publishes it to this bus environment. Any package requiring data for doing analytics subscribes to a stream and has it copied to its location for analysis in the case of streaming data. This separation of the data producers and consumers makes for very flexible application development. It also means that your single data feed could be simultaneously used by multiple consumers.

What else happens here at the central environment? There are receivers for just about any data type. You can both stream into the centralized data environment and out of the centralized environment in real time. While this is happening, processing functions decode the stream, extract interesting data, and put the data into relational databases or raw storage. It is also common to copy items from the data into some type of “object” storage environment for future processing. During the transform process, you may standardize, summarize, normalize, and store data. You transform data to something that is usable and standardized to fit into some existing analytics use case. This centralized environment, often called the “data warehouse” or “data lake,” is accessed through a variety of methods, such as Structured Query Language (SQL), application programming interface (API) calls, Kafka consumers, or even simple file access, just to name a few.

Before the data is stored at the central location, you may need to adjust these data, including doing the following:

- Data cleansing to make sure the data matches known types that your storage expects
- Data reconciliation, including filling missing data, cleaning up formats, removing duplicates, or bounding values to known ranges
- Deriving or generating any new values that you want included in the records
- Splitting or combining data into meaningful values for the domain
- Standardizing the data ingress or splitting a stream to keep standardized and raw data

Now let's return to the memory example: These telemetry data streams (subject: memory leak) from the network infrastructure must now be made available to the analytics tools and data scientists for analysis or application of the models. This availability must happen through the analytics engine part of the analytics infrastructure model. Figure 2-12 shows what types of activities are involved if there is a query or request for this data stream from analytics tools or packages. This query is requesting that a live feed of the stream be passed through the publisher/subscriber bus architecture and a normalized feed of the same stream be copied to a database for batch analysis. This is all set up in the software at the central data location.



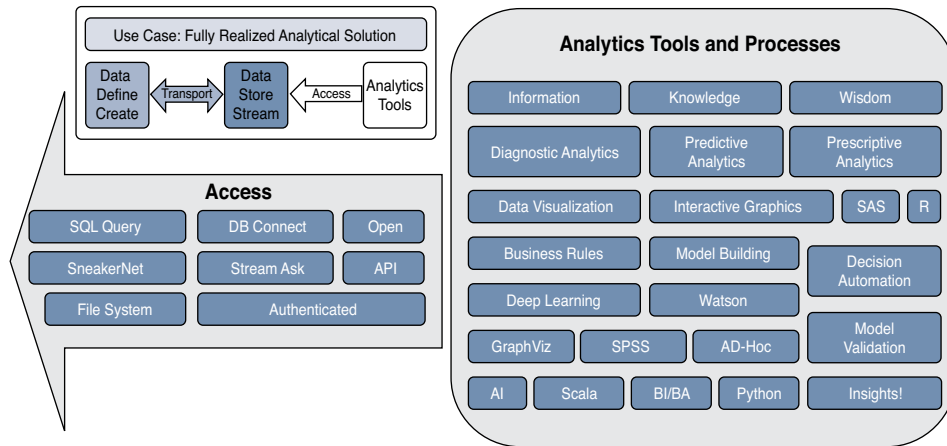
**Figure 2-12** *Analytics Infrastructure Model Streaming Data Example*

## Data Science

Data science is the sexy part of analytics. Data science includes the data mining, statistics, visualization, and modeling activities performed on readily available data. People often forget about the requirements to get the proper data to solve the individual use cases. The focus for most analysts is to start with the business problem first and then determine which type of data is required to solve or provide insights from the particular use case. Do not underestimate the time and effort required to set up the data for these use cases. Research shows that analysts spend 80% or more of their time on acquiring, cleaning, normalizing, transforming, or otherwise manipulating the data. I've spent upward of 90% on some problems.

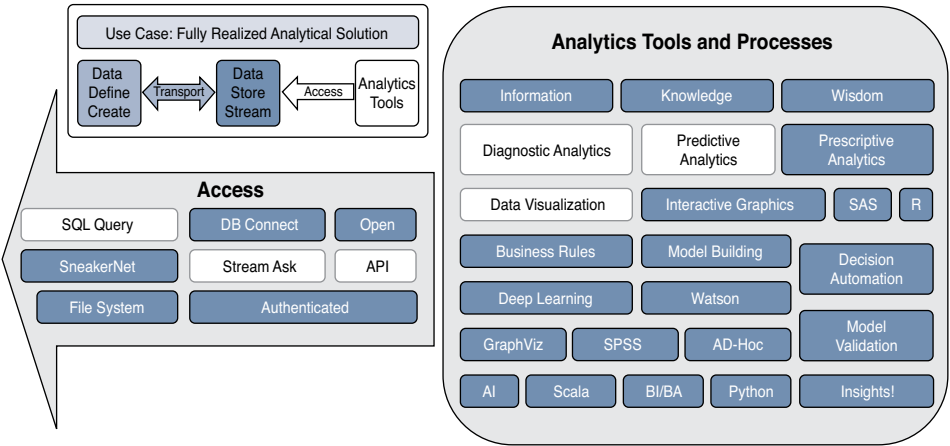
Analysts must spend so much time because analytics algorithms require specific representations or encodings of the data. In some cases, encoding is required because the raw stream appears to be gibberish. You can commonly do the transformations, standardizations, and normalizations of data in the data pipeline, depending on the use case. First you need to figure out the required data manipulations through your model building phases; you will ultimately add them inline to the model deployment phases, as shown in the previous diagrams, such that your data arrives at the data science tools ready to use in the models.

The analytics infrastructure model is valuable from the data science tools perspective because you can assume that the data is ready, and you can focus clearly on the data access and the tools you need to work on that data. Now you do the data science part. As shown in Figure 2-13, the data science part of the model highlights tools, processes, and capabilities that are required to build and deploy models.



**Figure 2-13** *Analytics Infrastructure Model Analytics Tools and Processes*

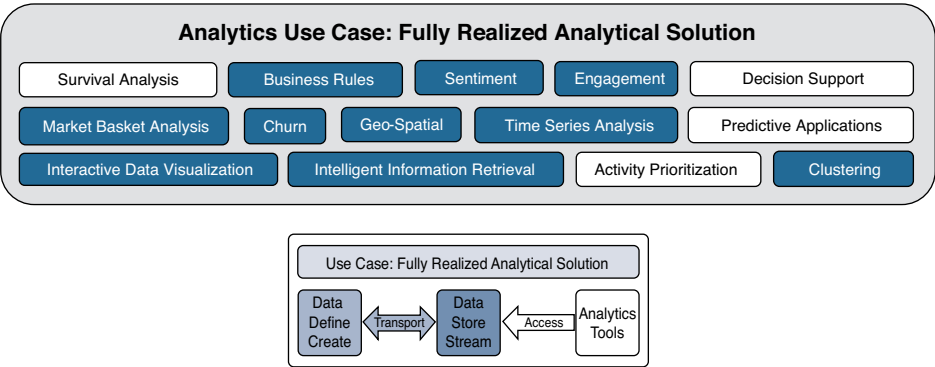
Going back to the streaming telemetry memory leak example, what should you do here? As highlighted in Figure 2-14, you use a SQL query to an API to set up the storage of the summary data. You also request full stream access to provide data visualization. Data visualization then easily shows both your technical and nontechnical stakeholders the obvious untamed growth of memory on certain platforms, which ultimately provides some “diagnostic analytics.” Insight: This platform, as you have it deployed, leaks memory with the current network conditions. You clearly show this with a data visualization, and now that you have diagnosed it, you can even build a predictive model for catching it before it becomes a problem in your network.



**Figure 2-14** *Analytics Infrastructure Model Streaming Analytics Example*

### Analytics Use Cases

The final section of the analytics infrastructure model is the use cases built on all this work that you performed: the “analytics solution.” Figure 2-15 shows some examples of generalized use cases that are supported with this example. You can build a predictive application for your memory case and use survival analysis techniques to determine which routers will hit this memory leak in the future. You can also use your analytics for decision support to management in order to prioritize activities required to correct the memory issue. Survival analysis here is an example of how to use common industry intuition to develop use cases for your own space. Survival analysis is about recognizing that something will not survive, such as a part in an industrial machine. You can use the very same techniques to recognize that a router will not survive a memory leak.

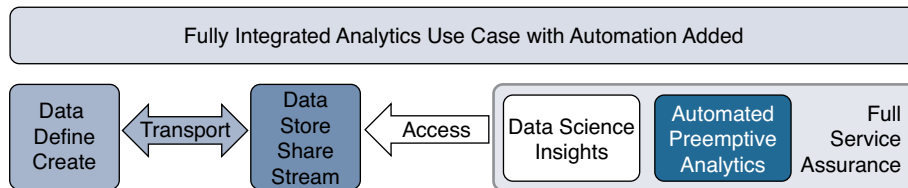


**Figure 2-15** *Analytics Infrastructure Model Analytics Use Cases Example*

As you go through the analytics use cases in later chapters, it is up to you and your context bias to determine how far to take each of the use cases. Often simple descriptive

analytics or a picture of what is in the environment is enough to provide a solution. Working toward wisdom from the data for predictive, prescriptive, and preemptive analytics solutions is well worth the effort in many cases. The determination of whether it is worth the effort is highly dependent on the capabilities of the systems, people, process, and tools available in your organization (including you).

Figure 2-16 shows where fully automated service assurance is added to the analytics infrastructure model. When you combine the analytics solution with fully automated remediation, you build a full-service assurance layer. Cisco builds full-service assurance layers into many architectures today, in solutions such as Digital Network Architecture (DNA), Application Centric Infrastructure (ACI), Crosswork Network Automation, and more that are coming in the near future. Automation is beyond the scope of this book, but rest assured that your analytics solutions are a valuable source for the automated systems to realize full-service assurance.



**Figure 2-16** *Analytics Infrastructure Model with Service Assurance Attachment*

## Summary

Now you understand that there is a method to the analytics madness. You also now know that there are multiple approaches you can take to data science problems. You understand that building a model on captive data in your own machine is an entirely different process from deploying a model in a production environment. You also understand different approaches to the process and that you and your stakeholders may each show preferences for different ones. Whether you are starting with the data exploration or the problem statement, you can find useful and interesting insights.

You may also have had your first introduction to the overlays and underlays concepts, which are important concepts as you go deeper into the data that is available to you from your network in the next chapter. Getting data to and from other overlay applications, as well as to and from other layers of the network is an important part of building complete solutions.

You now have a generalized analytics infrastructure model that helps you understand how the parts of analytics solutions come together to form a use case. Further, you understand that using the analytics infrastructure model allows you to build many different levels of analytics and provides repeatable, reusable components. You can choose how mature you wish your solution to be, based on factors from your own environment. The next few chapters take a deep dive into understanding the networking data from that environment.

*This page intentionally left blank*

# Index

## Symbols

---

& (ampersand), 306  
\  
(backslash), 288  
~ (tilde), 291–292, 370  
2×2 charts, 9–10  
5-tuple, 65

## A

---

access, data. *See* data access  
ACF (autocorrelation function), 262  
ACI (Application Centric Infrastructure),  
20, 33, 430–431  
active-active load balancing, 186  
activity prioritization, 170–173  
AdaBoost, 252  
Address Resolution Protocol (ARP), 61  
addresses  
    IP (Internet Protocol)  
        *packet counts*, 395–397  
        *packet format*, 390–391  
    MAC, 61, 398  
algorithms, 3–4, 217–218, 439  
    apriori, 242–243, 381–382  
    artificial intelligence, 267  
    assumptions of, 218–219

classification  
    *choosing algorithms for*, 248–249  
    *decision trees*, 249–250  
    *gradient boosting methods*,  
        251–252  
    *neural networks*, 252–258  
    *random forest*, 250–251  
    SVMs (*support vector machines*),  
        258–259  
    *time series analysis*, 259–262  
confusion matrix, 267–268  
contingency tables, 267–268  
cumulative gains and lift, 269–270  
data-encoding methods, 232–233  
dimensionality reduction, 233–234  
feature selection, 230–232  
regression analysis, 246–247  
simulation, 271  
statistical analysis  
    ANOVA (*analysis of variance*), 227  
    *Bayes' theorem*, 228–230  
    *box plots*, 221–222  
    *correlation*, 224–225  
    *longitudinal data*, 225–226  
    *normal distributions*, 222–223  
    *outliers*, 223  
    *probability*, 228  
    *standard deviation*, 222–223

- supervised learning, 246
- terminology, 219–221
- text and document analysis, 256–262
  - information retrieval*, 263–264
  - NLP (natural language processing)*, 262–263
  - sentiment analysis*, 266–267
  - topic modeling*, 265–266
- unsupervised learning
  - association rules*, 240–243
  - clustering*, 234–239
  - collaborative filtering*, 244–246
  - defined*, 234
  - sequential pattern mining*, 243–244
- alpha, 261
- Amazon, recommender system for, 191–194
- ambiguity bias, 115–116
- ampersand (&), 306
- analysis of variance. *See* ANOVA (analysis of variance)
- analytics algorithms. *See* algorithms
- analytics experts, 25
- analytics infrastructure model, 22–25, 275–276
  - data and transport, 26–28
  - data engine, 28–30
  - data science, 30–32
  - data streaming example, 30
  - illustrated, 437
  - publisher/subscriber environment, 29
  - roles, 24–25
  - service assurance, 33
  - traditional thinking versus, 22–24
  - use cases
    - algorithms*, 3–4
    - defined*, 18–19
    - development*, 2–3
    - examples of*, 32–33
- analytics maturity, 7–8
- analytics models, building, 2, 14–15, 19–20. *See also* use cases
- analytics infrastructure model, 22–25, 275–276, 437
  - data and transport*, 26–28
  - data engine*, 28–30
  - data science*, 30–32
  - data streaming example*, 30
  - publisher/subscriber environment*, 29
  - roles*, 24–25
  - service assurance*, 33
  - traditional thinking versus*, 22–24
- deployment, 2, 14–15, 17–18
- EDA (exploratory data analysis)
  - defined*, 15–16
  - use cases versus solutions*, 18–19
  - walkthrough*, 17–18
- feature engineering, 219
- feature selection, 219
- interpretation, 220
- overfitting, 219
- overlay, 20–22
- problem-centric approach
  - defined*, 15–16
  - use cases versus solutions*, 18–19
  - walkthrough*, 17–18
- underlay, 20–22
- validation, 219
- analytics process, 437
- analytics scales, 436
- analytics solutions, defined, 150
- anchoring effect, 107–109
- AND operator, 306
- ANNs (artificial neural networks), 254–255
- anomaly detection, 153–155
  - clustering, 239
  - statistical, 318–320
- ANOVA (analysis of variance), 227, 305–310
  - data filtering, 305–306
  - describe function, 308
  - drop command, 309
  - groupby command, 307



- homogeneity of variance, 313–318
- Levene's test, 313
- outliers, dropping, 307–310
- pairwise, 317
- Apache Kafka, 28–29
- API (application programming interface)
  - calls, 29
- App iQ platform, 430
- AppDynamics, 6, 428–430
- Application Centric Infrastructure (ACI), 20, 33, 430–431
- application programming interface (API)
  - calls, 29
- application-specific integrated circuits (ASICs), 67
- apply method, 295–296, 346
- approaches. *See* methodology and approach
- apriori algorithms, 242–243, 381–382
- architecture
  - architecture and advisory services, 426–427
  - big data, 4–5
  - microservices, 5–6
- Ariely, Dan, 108
- ARIMA (autoregressive integrated moving average), 101–102, 262
- ARP (Address Resolution Protocol), 61
- artificial general intelligence, 267
- artificial intelligence, 11, 267
- artificial neural networks (ANNs), 254–255
- ASICs (application-specific integrated circuits), 67
- assets
  - data plane analytics use case, 422–423
  - tracking, 173–175
- association rules, 240–243
- associative thinking, 131–132
- authority bias, 113–114
- autocorrelation function (ACF), 262
- automation, 11, 33, 431–432
- autonomous applications, use cases for, 200–201

- autoregressive integrated moving average (ARIMA), 101–102, 262
- autoregressive process, 262
- availability bias, 111
- availability cascade, 112, 141
- averages
  - ARIMA (autoregressive integrated moving average), 262
  - moving averages, 262
- Azure Cloud Network Watcher, 68

## B

---

- BA (business analytics) dashboards, 13, 42
- back-propagation, 254
- backslash (\), 288
- bagging, 250–251
- bar charts, platform crashes example, 289–290
- base-rate neglect, 117
- Bayes' theorem, 228–230
- Bayesian methods, 230
- BCI (Business Critical Insights), 335, 425
- behavior analytics, 175–178
- benchmarking use cases, 155–157
- BGP (Border Gateway Protocol), 41, 61
- BI (business intelligence) dashboards, 13, 42
- bias, 2–3, 439
  - ambiguity, 115–116
  - anchoring effect, 107–109
  - authority, 113–114
  - availability, 111
  - availability cascade, 112
  - base-rate neglect, 117
  - clustering, 112
  - concept of, 104–105
  - confirmation, 114–115
  - context, 116–117
  - correlation, 112
  - “curse of knowledge”, 119
  - Dunning-Kruger effect, 120–121

- empathy gap, 123
- endowment effect, 121
- expectation, 114–115
- experimenter's, 116
- focalism, 107
- framing effect, 109–110, 151
- frequency illusion, 117
- group, 120
- group attribution error, 118
- halo effect, 123–124
- hindsight, 9, 123–124
- HIPPO (highest paid persons' opinion)
  - impact, 113–114
- IKEA effect, 121–122
- illusion of truth effect, 112–113
- impact of, 105–106
- imprinting, 107
- innovation and, 128
- “law of small numbers”, 117–118
- mirroring, 110–111
- narrative fallacy, 107–108
- not-invented-here syndrome, 122
- outcome, 124
- priming effect, 109, 151
- pro-innovation, 121
- recency, 111
- solutions and, 106–107
- status-quo, 122
- sunk cost fallacy, 122
- survivorship, 118–119
- table of, 124–126
- thrashing, 122
- tunnel vision, 107
- WYSIATI (What You See Is All There Is), 118
- zero price effect, 123

**Bias, Randy, 204**

**big data, 4–5**

**Border Gateway Protocol (BGP), 41, 61**

**box plots, 221–222**

- platform crashes example, 297–299
- software crashes example, 300–305

- Box-Jenkins method, 262
- breaking anchors, 140
- Breusch-Pagan tests, 220
- budget analysis, 169
- bug analysis use cases, 178–179
- business analytics (BA) dashboards,
  - 13, 42
- Business Critical Insights (BCI), 335, 425
- business domain experts, 25
- business intelligence (BI) dashboards,
  - 13, 42
- business model
  - analysis, 200–201
  - optimization, 201–202

## C

---

- capacity planning, 180–181
- CARESS technique, 137
- cat /etc/\*release command, 61
- categorical data, 77–78
- causation, correlation versus, 112
- CDP (Cisco Discovery Protocol), 60, 93
- charts
  - cumulative gains, 269–270
  - lift, 269–270
  - platform crashes use case, 289–290
- churn use cases, 202–204
- Cisco analytics solutions, 6,
  - 425–426, 442
- analytics platforms and partnerships, 433
- AppDynamics, 428–430
- architecture and advisory services,
  - 426–427
- BCI (Business Critical Insights), 335, 425
- CMS (Cisco Managed Services), 425
- Crosswork automation, 431–432
- DNA (Digital Network Architecture), 428
- IoT (Internet of Things) analytics, 432
- open source platform, 433–434
- Stealthwatch, 427
- Tetration, 430–431

- Cisco Application Centric Infrastructure (ACI), 20
- Cisco Discovery Protocol (CDP), 60
- Cisco Identity Service Engine (ISE), 427
- Cisco IMC (Integrated Management Controller), 40–41
- Cisco iWAN+Viptela, 20
- Cisco TrustSec, 427
- Cisco Unified Computing System (UCS), 62
- citizen data scientists, 11
- classification, 157–158
  - algorithms
    - choosing*, 248–249
    - decision trees*, 249–250
    - gradient boosting methods*, 251–252
    - neural networks*, 252–258
    - random forest*, 250–251
  - SVMs (support vector machines), 258–259
  - time series analysis, 259–262
- cleansing data, 29, 86
- CLI (command-line interface) scraping, 59, 92
- cloud software, 5–6
- Cloudera, 433
- clustering, 234–239
  - K-means, 344–349, 373–375
  - machine learning-guided troubleshooting, 350–353
  - SME port clustering, 407–413
    - cluster scatterplot*, 410–411
    - host patterns*, 411–413
    - K-means clustering*, 408–410
    - port profiles*, 407–408
  - use cases, 158–160
- clustering bias, 112
- CMS (Cisco Managed Services), 425
- CNNs (convolutional neural networks), 254–255
- cognitive bias. *See* bias
- Cognitive Reflection Test (CRT), 98
- cognitive trickery, 143
- cohorts, 160
- collaborative filtering, 244–246
- collinearity, 225
- columns
  - dropping, 287
  - grouping, 307
- columns command, 286
- Colvin, Geoff, 103
- command-line interface (CLI) scraping, 59, 92
- commands. *See also* functions
  - `cat /etc/*release`, 61
  - columns, 286
  - drop, 309
  - groupby, 307, 346, 380, 398
  - head, 396, 404
  - join, 291
  - tcpdump, 68
- comma-separated values (CSV) files, 82
- communication, control plane, 38
- Competing on Analytics* (Davenport and Harris), 148
- compliance to benchmark, 155
- computer thrashing, 140
- condition-based maintenance, 189
- confirmation bias, 114–115
- confusion matrix, 267–268
- container on box, 74–75
- context
  - context bias, 116–117
  - context-sensitive stop words, 329
  - external data for, 89
- contingency tables, 267–268
- continuous numbers, 78–79
- control plane, 441
  - activities in, 41
  - communication, 38
  - data examples, 46–47, 67–68
  - defined, 37
  - syslog telemetry use case, 355
    - data encoding*, 371–373
    - data preparation*, 356–357, 369–371

- high-volume producers, identifying*, 362–366
- K-means clustering*, 373–375
- log analysis with pandas*, 357–360
- machine learning-based evaluation*, 366–367
- noise reduction*, 360–362
- OSPF (Open Shortest Path First) routing*, 357
- syslog severities*, 359–360
- task list*, 386–387
- transaction analysis*, 379–386
- word cloud visualization*, 367–369, 375–379
- convolutional neural networks (CNNs), 254–255
- correlation
  - correlation bias, 112
  - explained, 224–225
  - use cases, 160–162
- cosine distance, 236
- count-encoded matrix, 336–338
- CountVectorizer method, 338
- covariance, 167
- Covey, Stephen, 10
- crashes, device. *See* device crash use cases
- crashes, network. *See* network infrastructure analytics use case
- CRISP-DM (cross-industry standard process for data mining), 18
- critical path, 172, 211
- CRM (customer relationship management) systems, 25, 187
- cross-industry standard process for data mining (CRISP-DM), 18
- Crosswork Network Automation, 33, 431–432
- crowdsourcing, 133–134
- CRT (Cognitive Reflection Test), 98
- CSV (comma-separated value) files, 82
- cumulative gains, 269–270
- curse of dimensionality, 159
- “curse of knowledge”, 119
- custom labels, 93

- customer relationship management (CRM) systems, 25, 187
- customer segmentation, 160

## D

---

- data. *See also* data access
  - domain experts, 25
  - encoding, 232–233
    - network infrastructure analytics use case*, 328–336
    - syslog telemetry use case*, 371–373
  - engine, 28–30
  - gravity, 76
  - loading
    - data plane analytics use case*, 390–394
    - network infrastructure analytics use case*, 325–328
    - statistics use cases*, 286–288
  - mining, 150
  - munging, 85
  - network, 35–37
    - business and applications data relative to*, 42–44
    - control plane*, 37, 38, 41, 46–47
    - data plane*, 37, 41, 47–49
    - management plane*, 37, 40–41, 44–46
    - network virtualization*, 49–51
    - OpenStack nodes*, 39–40
    - planes, combining across virtual and physical environments*, 51–52
    - sample network*, 38
  - normalization, 85
  - preparation, 29, 86
    - encoding methods*, 85
    - KPIs (key performance indicators)*, 86–87
    - made-up data*, 84–85
    - missing data*, 86
    - standardized data*, 85
    - syslog telemetry use case*, 355, 369–371, 379

- reconciliation, 29
  - regularization, 85
  - scaling, 298
  - standardizing, 85
  - storage, 6
  - streaming, 30
  - structure, 82
    - JSON (JavaScript Object Notation)*, 82–83
    - semi-structured data*, 84
    - structured data*, 82
    - unstructured data*, 83–84
  - transformation, 310
  - transport, 89–90
    - CLI (command-line interface) scraping*, 92
    - HLD (high-level design)*, 90
    - IPFIX (IP Flow Information Export)*, 95
    - LLD (low-level design)*, 90
    - NetFlow*, 94
    - other data*, 93
    - sFlow*, 95
    - SNMP (Simple Network Management Protocol)*, 90–92
    - SNMP (Simple Network Management Protocol) traps*, 93
    - Syslog*, 93–94
    - telemetry*, 94
  - types, 76–77
    - continuous numbers*, 78–79
    - discrete numbers*, 79
    - higher-order numbers*, 81–82
    - interval scales*, 80
    - nominal data*, 77–78
    - ordinal data*, 79–80
    - ratios*, 80–81
  - warehouses, 29
- data access. *See also* data structure; transport of data; types**
- container on box, 74–75
  - control plane data, 67–68
  - data plane traffic capture, 68–69
  - ERSPAN (Encapsulated Remote Switched Port Analyzer)*, 69
  - inline security appliances*, 69
  - port mirroring*, 69
  - RSPAN (Remote SPAN)*, 69
  - SPAN (Switched Port Analyzer)*, 69
  - virtual switch operations*, 69–70
  - DPI (deep packet inspection), 56
  - external data for context, 89
  - IoT (Internet of Things) model, 75–76
  - methods of, 55–57
  - observation effect, 88
  - packet data, 70–74
    - HTTP (Hypertext Transfer Protocol)*, 71–72
    - IPsec (Internet Protocol Security)*, 73–74
    - IPv4*, 70–71
    - SSL (Secure Sockets Layer)*, 74
    - TCP (Transmission Control Protocol)*, 71–72
    - VXLAN (Virtual Extensible LAN)*, 74
  - panel data, 88
  - pull data availability
    - CLI (command-line interface) scraping*, 59, 92
    - NETCONF (Network Configuration Protocol)*, 60
    - SNMP (Simple Network Management Protocol)*, 57–59
    - unconventional data sources*, 60–61
    - YANG (Yet Another Next Generation)*, 60
  - push data availability
    - IPFIX (IP Flow Information Export)*, 64–67
    - NetFlow*, 65–66
    - sFlow*, 67, 95
    - SNMP (Simple Network Management Protocol) traps*, 61–62, 93
    - Syslog*, 62–63, 93–94
    - telemetry*, 63–64
  - timestamps, 87–88

- data lake, 29
- data pipeline engineering, 90
- data plane. *See also* data plane analytics use case
  - activities in, 41
  - data examples, 47–49
  - defined, 37
  - traffic capture, 68–69
    - ERSPAN (Encapsulated Remote Switched Port Analyzer)*, 69
    - inline security appliances*, 69
    - port mirroring*, 69
    - RSPAN (Remote SPAN)*, 69
    - SPAN (Switched Port Analyzer)*, 69
    - virtual switch operations*, 69–70
- data plane analytics use case, 389, 442
  - assets, 422–423
  - data loading and exploration, 390–394
    - IP package format*, 390–391
    - packet file loading*, 390
    - parsed fields*, 392–393
    - Python packages, importing*, 390
    - TCP package format*, 391
  - full port profiles, 413–419
  - investigation task list, 423–424
  - SME analysis
    - dataframe and visualization library loading*, 394
    - host analysis*, 399–404
    - IP address packet counts*, 395–397
    - IP packet protocols*, 398
    - MAC addresses*, 398
    - output*, 404–406
    - time series counts*, 395
    - timestamps and time index*, 394–395
    - topology mapping information*, 398
  - SME port clustering, 407–413
    - cluster scatterplot*, 410–411
    - host patterns*, 411–413
    - K-means clustering*, 408–410
    - port profiles*, 407–408
  - source port profiles, 419–422
- data science, 25, 30–32, 278–280
- data structure, 82
- databases, 6
- dataframes
  - combining, 292–293
  - defined, 286–287
  - dropping columns from, 287
  - filtering, 287, 290–292, 300, 330, 370
  - grouping, 293–296, 299–300, 307
  - loading, 394
  - outlier analysis, 318–320
  - PCA (principal component analysis), 339–340, 372–373
  - sorting without, 326–327
  - `value_counts` function, 288–290
  - views, 329–330, 347
- data-producing sensors, 210–211
- Davenport, Thomas, 148
- de Bono, Edward, 132
- decision trees
  - example of, 249–250
  - random forest, 250–251
- deep packet inspection (DPI), 56
- defocusing, 140
- deliberate practice, 100, 102
- delivery models, use cases for, 210–212
- delta, 262
- dependence, 261
- deployment of models, 2, 14–15, 17–18
- describe function, 308
- descriptive analytics, 8–9
- descriptive analytics use cases, 167–168
- designing solutions. *See* solution design
- destination IP address packet counts, 396–397
- deviation, standard, 222–223
- device crash use cases, 285
  - anomaly detection, 318–320
  - ANOVA (analysis of variance), 305–310
    - data filtering*, 305–306
    - describe function*, 308

- drop command*, 309
  - groupby command*, 307
  - homogeneity of variance*, 313–318
  - outliers, dropping*, 307–310
  - pairwise*, 317
  - data loading and exploration, 286–288
  - data transformation, 310
  - normality, tests for, 311–313
  - platform crashes, 288–299
    - apply method*, 295–296
    - box plot*, 297–298
    - crash counts by product ID*, 294–295
    - crash counts/rate comparison plot*, 298–299
    - crash rates by product ID*, 296–298
    - crashes by platform*, 292–294
    - data scaling*, 298
    - dataframe filtering*, 290–292
    - groupby object*, 293–296
    - horizontal bar chart*, 289–290
    - lambda function*, 296
    - overall crash rates*, 292
    - router reset reasons*, 290
    - simple bar chart*, 289
    - value\_counts function*, 288–289
  - software crashes, 299–305
    - box plots*, 300–305
    - dataframe filtering*, 300
    - dataframe grouping*, 299–300
  - diagnostic targeting, 209
  - “dial-in” telemetry configuration, 64
  - “dial-out” telemetry configuration, 64
  - dictionaries, tokenization and, 328
  - diffs function, 352
  - Digital Network Architecture (DNA), 33, 428
  - dimensionality
    - curse of, 159
    - reduction, 233–234, 337–340
  - discrete numbers, 79
  - distance methods, 236
  - divisive clustering, 236
  - DNA (Digital Network Architecture), 33, 428
  - DNA mapping, 324–325
  - DNAC (DNA Center), 428
  - doc2bow, 331–332
  - document analysis, 256–262
    - information retrieval, 263–264
    - NLP (natural language processing), 262–263
    - sentiment analysis, 266–267
    - topic modeling, 265–266
  - DPI (deep packet inspection), 56
  - drop command, 309
  - dropouts, 204–206
  - dropping columns, 287
  - Duhigg, Charles, 99
  - dummy variables, 232
  - Dunning-Kruger effect, 120–121
- ## E
- 
- EDA (exploratory data analysis)
    - defined, 15–16
    - use cases versus solutions, 18–19
    - walkthrough, 17–18
  - edit distance, 236
  - EDT (event-driven telemetry), 64
  - EIGRP (Enhanced Interior Gateway Routing Protocol), 61, 398
  - ElasticNet regression, 247
  - electronic health records, 210
  - empathy gap, 123
  - Encapsulated Remote Switched Port Analyzer (ERSPAN), 69
  - encoding methods, 85, 232–233
    - network infrastructure analytics use case, 328–336
    - syslog telemetry use case, 371–373
  - Encrypted Traffic Analytics (ETA), 427
  - endowment effect, 121
  - engagement models, 206–207



engine, analytics infrastructure model, 28–30  
 Enhanced Interior Gateway Routing Protocol (EIGRP), 61, 398  
 entropy, 250  
 environment setup, 282–284, 325–328  
 episode mining, 244  
 errors, group attribution, 118. *See also* bias  
 ERSPAN (Encapsulated Remote Switched Port Analyzer), 69  
 ETA (Encrypted Traffic Analytics), 427  
 ETL (Extract, Transform, Load), 26  
 ETSI (European Telecommunications Standards Institute), 75  
 Euclidean distance, 236  
 European Telecommunications Standards Institute (ETSI), 75  
 event log analysis use cases, 181–183  
 event-driven telemetry (EDT), 64  
 expectation bias, 114–115  
 experimentation, 141–142  
 experimenter's bias, 116  
 expert systems deployment, 214  
 exploratory data analysis. *See* EDA (exploratory data analysis)  
 exponential smoothing techniques, 261  
 external data for context, 89  
 Extract, Transform, Load (ETL), 26

## F

---

F statistic, 220  
 failure analysis use cases, 183–185  
 fast path, 211  
 features  
   defined, 42–43  
   feature engineering, 219  
   selection, 219, 230–232  
 Few, Stephen, 163  
 fields, data plane analytics use case, 392–393  
 files, CSV (comma-separated value), 82. *See also* logs  
 fillna, 342–343

filtering  
   ANOVA and, 305–306  
   collaborative, 244–246  
   dataframes, 287, 290–292, 300, 330, 370  
   platform crashes example, 290–292  
   software crashes example, 300  
 fingerprinting, 324–325  
 “Five whys” technique, 137–138  
 Flexible NetFlow, 65  
 Flight 1549, 99–100  
 focalism, 107  
 fog computing, 76  
 foresight, 9  
 FP growth algorithms, 242  
 framing effect, 109–110, 151  
 Franks, Bill, 147  
 fraud detection use cases, 207–209  
 Frederick, Shane, 98  
 FreeSpan, 244  
 frequency illusion, 117  
 F-tests, 227, 314  
 full host profiles, 401–403  
 full port profiles, 413–419  
 functions  
   apply, 295–296, 346  
   apriori, 242–243, 381–382  
   CountVectorizer, 338  
   describe, 308  
   diffs, 352  
   host\_profile, 403  
   join, 370  
   lambda, 296  
   max, 347  
   reset\_index, 414  
   split, 368  
   value\_counts, 288–289, 396, 400, 403

## G

---

gains, cumulative, 269–270  
 gamma, 261  
 Gartner analytics, 8  
 gender bias, 97–98



generalized sequential pattern (GSP), 244  
 Gensim package, 264, 283, 328, 331–332  
 Gladwell, Malcolm, 99  
 Global Positioning System (GPS), 210–211  
 Goertzel, Ben, 267  
 GPS (Global Positioning System), 210–211  
 gradient boosting methods, 251–252  
 gravity, data, 76  
 group attribution error, 118  
 group bias, 120  
 group-based strong learners, 250  
 groupby command, 307, 346, 380, 398  
 groupby object, 293–296  
 grouping  
   columns, 307  
   dataframes, 293–296, 299–300  
 GSP (generalized sequential pattern), 244

## H

---

Hadoop, 28–29  
 halo effect, 123–124  
 hands-on experience, mental models and, 100  
 hard data, 150  
 Harris, Jeanne, 148  
 head command, 396, 404  
*Head Game* (Mudd), 110  
 healthcare use cases, 209–210  
 Hewlett-Packard iLO (Integrated Lights Out), 40–41  
 hierarchical agglomerative clustering, 236–237  
 higher-order numbers, 81–82  
 highest paid persons' opinion (HIPPO) impact, 113–114  
 high-level design (HLD), 90  
 high-volume producers, identifying, 362–366  
 hindsight bias, 9, 123–124  
 HIPPO (highest paid persons' opinion) impact, 113–114

HLD (high-level design), 90  
 homogeneity of variance, 313–318  
 homoscedasticity, 313–318  
 Hortonworks, 433  
 host analysis, 399–404  
   data plane analytics use case, 411–413  
   full host profile analysis, 401–403  
   per-host analysis function, 399  
   per-host conversion analysis, 400–401  
   per-host port analysis, 403  
 host\_profile function, 403  
*How Not to Be Wrong* (Ellenberg), 118–119  
 HTTP (Hypertext Transfer Protocol), 71–72  
 human bias, 97–98  
 Hypertext Transfer Protocol (HTTP), 71–72  
 Hyper-V, 70

## I

---

IBM, Cisco's partnership with, 433  
 IBN (intent-based networking), 11, 428  
 ICMP (Internet Control Message Protocol), 398  
 ID3 algorithm, 250  
 Identity Service Engine (ISE), 427  
 IETF (Internet Engineering Task Force), 66–67, 95  
 IGMP (Internet Group Management Protocol), 398  
 IGP (interior gateway protocols), 357  
 IIA (International Institute for Analytics), 147  
 IKEA effect, 121–122  
 illusion of truth effect, 112–113  
 iLO (Integrated Lights Out), 40–41  
 image recognition use cases, 170  
 IMC (Integrated Management Controller), 40–41  
 importing Python packages, 390  
 imprinting, 107  
 industry terminology, 7

**inference, statistical, 228**

**influence, 227**

**information retrieval**

algorithms, 263–264

use cases, 185–186

**Information Technology Infrastructure Library (ITIL), 161**

**infrastructure analytics use case, 323–324**

data encoding, 328–336

data loading, 325–328

data visualization, 340–344

dimensionality reduction, 337–340

DNA mapping and fingerprinting, 324–325

environment setup, 325–328

K-means clustering, 344–349

machine learning-guided troubleshooting, 350–353

search challenges and solutions, 331–336

**in-group bias, 120**

**inline security appliances, 69**

**innovative thinking techniques, 127–128, 439**

associative thinking, 131–132

bias and, 128

breaking anchors, 140

cognitive trickery, 143

crowdsourcing, 133–134

defocusing, 140

experimentation, 141–142

inverse thinking, 139–140, 204–206

lean thinking, 142

metaphoric thinking, 130–131

mindfulness, 128

networking, 133–135

observation, 138–139

perspectives, 130–131

questioning

*CARESS technique, 137*

*example of, 135–137*

*“Five whys”, 137–138*

quick innovation wins, 143–144

six hats thinking approach, 132–133

unpriming, 140

*The Innovator's DNA* (Dyer et al), 128

**insight, 9**

**installing Jupyter Notebook, 282–283**

**Integrated Lights Out (iLO), 40–41**

**Integrated Management Controller (IMC), 40–41**

**Intelligent Wide Area Networks (iWAN), 20, 428**

**intent-based networking (IBN), 11, 428**

**interior gateway protocols (IGPs), 357**

**International Institute for Analytics (IIA), 147**

**Internet clickstream analysis, 169**

**Internet Control Message Protocol (ICMP), 398**

**Internet Engineering Task Force (IETF), 66–67, 95**

**Internet Group Management Protocol (IGMP), 398**

**Internet of Things (IoT), 75–76**

analytics, 432

growth of, 214

*Internet of Things—From Hype to Reality* (Rayes and Salam), 75

**Internet Protocol (IP)**

IP address packet counts, 395–397

packet format, 390–391

packet protocols, 398

**Internet Protocol Security (IPsec), 73–74**

**interval scales, 80**

**intrusion detection use cases, 207–209**

**intuition**

explained, 103–104

System 1/System 2, 102–103

**inventory management, 169**

**inverse problem, 206**

**inverse thinking, 139–140, 204–206**

**IoT. See Internet of Things (IoT)**

**IP (Internet Protocol)**

IPFIX (IP Flow Information Export),  
64–67, 95

packet counts, 395–397

packet data, 70–71

packet format, 390–391

packet protocols, 398

IPFIX (IP Flow Information Export),  
64–67, 95

IPsec (Internet Protocol Security),  
73–74

ISE (Identity Service Engine), 427

isin keyword, 366

IT analytics use cases, 170

activity prioritization, 170–173

asset tracking, 173–175

behavior analytics, 175–178

bug and software defect analysis,  
178–179

capacity planning, 180–181

event log analysis, 181–183

failure analysis, 183–185

information retrieval, 185–186

optimization, 186–188

prediction of trends, 190–194

predictive maintenance, 188–189

scheduling, 194–195

service assurance, 195–197

transaction analysis, 197–199

ITIL (Information Technology  
Infrastructure Library), 161

iWAN (Intelligent Wide Area Networks),  
20, 428

## J

---

Jaccard distance, 236

Jasper, 432

JavaScript Object Notation (JSON),  
82–83

join command, 291

join function, 370

JSON (JavaScript Object Notation), 82–83

Jupyter Notebook, installing, 282–283

## K

---

Kafka (Apache), 28–29

Kahneman, Daniel, 102–103

kcluster values, 347. *See also* K-means  
clustering

Kendall's tau, 225, 236

Kenetic, 430–433

key performance indicators (KPIs),  
86–87

keys, 82–83

key/value pairs, 82–83

keywords, isin, 366

Kinetic, 430–433

K-means clustering

data plane analytics use case, 408–410

network infrastructure analytics use case,  
344–349

syslog telemetry use case, 373–375

knowledge

curse of, 119

management of, 8

known attack vectors, 214

KPIs (key performance indicators), 86–87

Kurzweil, Ray, 267

## L

---

labels, 151

ladder of powers methods, 310

lag, 262

lambda function, 296

language

selection, 6

translation, 11

lasso regression, 247

latent Dirichlet allocation (LDA), 265,  
334–335

latent semantic indexing (LSI), 265–266,  
334–335

law of parsimony, 120, 152

“law of small numbers”, 117–118

- LDA (latent Dirichlet allocation), 265, 334–335
- The Lean Startup* (Ries), 142
- lean thinking, 142
- learning reinforcement, 212–213
- left skewed distribution, 310
- lemmatization, 263
- Levene's test, 313
- leverage, 227
- lift charts, 269–270
- lift-and-gain analysis, 194
- LightGBM, 252
- linear regression, 246–247
- Link Layer Discovery Protocol (LLDP), 61
- Linux servers, pull data availability, 61
- LLD (low-level design), 90
- LLDP (Link Layer Discovery Protocol), 61, 93
- load balancing, active-active, 186
- loading data
  - data plane analytics use case, 390–394
    - dataframes*, 394
    - IP package format*, 390–391
    - packet file loading*, 390
    - parsed fields*, 392–393
    - Python packages, importing*, 390
    - TCP package format*, 391
  - network infrastructure analytics use case, 325–328
  - statistics use cases, 286–288
- logical AND, 306
- logistic regression, 101–102, 247
- logistics use cases, 210–212
- logs
  - event log analysis, 181–183
  - syslog telemetry use case, 355
    - data encoding*, 371–373
    - data preparation*, 356–357, 369–371
    - high-volume producers, identifying*, 362–366
    - K-means clustering*, 373–375

- log analysis with pandas*, 357–360
- machine learning-based evaluation*, 366–367
- noise reduction*, 360–362
- OSPF (Open Shortest Path First) routing*, 357
- syslog severities*, 359–360
- task list*, 386–387
- transaction analysis*, 379–386
- word cloud visualization*, 367–369, 375–379

- Long Short Term Memory (LSTM)
  - networks, 254–258

- longitudinal data, 225–226

- low-level design (LLD), 90

- LSI (latent semantic indexing), 265–266, 334–335

- LSTM (Long Short Term Memory)
  - networks, 254–258

## M

---

- M2M initiatives, 75

- MAC addresses, 61, 398

- machine learning

- classification algorithms

- choosing*, 248–249

- decision trees*, 249–250

- gradient boosting methods*, 251–252

- neural networks*, 252–258

- random forest*, 250–251

- defined, 150

- machine learning-based log evaluation, 366–367

- supervised, 151, 246

- troubleshooting with, 350–353

- unsupervised

- association rules*, 240–243

- clustering*, 234–239

- collaborative filtering*, 244–246

- defined*, 151, 234

- sequential pattern mining*, 243–244

- use cases, 153
  - anomalies and outliers*, 153–155
  - benchmarking*, 155–157
  - classification*, 157–158
  - clustering*, 158–160
  - correlation*, 160–162
  - data visualization*, 163–165
  - descriptive analytics*, 167–168
  - NLP (natural language processing)*, 165–166
  - time series analysis*, 168–169
  - voice, video, and image recognition*, 170
- making your own data, 84–85
- Management Information Bases (MIBs), 57
- management plane
  - activities in, 40–41
  - data examples, 44–46
  - defined, 37
- Manhattan distance, 236
- manipulating data
  - encoding methods, 85
  - KPIs (key performance indicators), 86–87
  - made-up data, 84–85
  - missing data, 86
  - standardized data, 85
- manufacturer's suggested retail price (MSRP), 108
- mapping, DNA, 324–325
- market basket analysis, 199
- Markov Chain Monte Carlo (MCMC) systems, 271
- matplotlib package, 283
- maturity levels, 7–8
- max method, 347
- MBIs (Management Information Bases), 57
- MCMC (Markov Chain Monte Carlo) systems, 271
- MDT (model-driven telemetry), 64
- mean squared error (MSE), 227
- memory, muscle, 102
- mental models
  - bias
    - ambiguity*, 115–116
    - anchoring effect*, 107–109
    - authority*, 113–114
    - availability*, 111, 112
    - base-rate neglect*, 117
    - clustering*, 112
    - concept of*, 104–105
    - confirmation*, 114–115
    - context*, 116–117
    - correlation*, 112
    - “curse of knowledge”*, 119
    - Dunning-Kruger effect*, 120–121
    - empathy gap*, 123
    - endowment effect*, 121
    - expectation*, 114–115
    - experimenter's*, 116
    - focalism*, 107
    - framing effect*, 109–110, 151
    - frequency illusion*, 117
    - group*, 120
    - group attribution error*, 118
    - halo effect*, 123–124
    - hindsight*, 9, 123–124
    - HIPPO (highest paid persons' opinion) impact*, 113–114
    - IKEA effect*, 121–122
    - illusion of truth effect*, 112–113
    - impact of*, 105–106
    - imprinting*, 107
    - “law of small numbers”*, 117–118
    - mirroring*, 110–111
    - narrative fallacy*, 107–108
    - not-invented-here syndrome*, 122
    - outcome*, 124
    - priming effect*, 109, 151
    - pro-innovation*, 121
    - recency*, 111
    - solutions and*, 106–107
    - status-quo*, 122
    - sunk cost fallacy*, 122

- survivorship*, 118–119
- table of*, 124–126
- thrashing*, 122
- tunnel vision*, 107
- WYSIATI (*What You See Is All There Is*), 118
- zero price effect*, 123
- changing how you think, 98–99
- concept of, 97–98, 99–102
- CRT (Cognitive Reflection Test), 98
- human bias, 97–98
- intuition, 103–104
- System 1/System 2, 102–103
- metaphoric thinking**, 130–131
- meters, smart**, 189
- methodology and approach**, 13–14
  - analytics infrastructure model, 22–25.
    - See also* use cases
    - data and transport*, 26–28
    - data engine*, 28–30
    - data science*, 30–32
    - data streaming example*, 30
    - publisher/subscriber environment*, 29
    - roles*, 24–25
    - service assurance*, 33
    - traditional thinking versus*, 22–24
- BI/BA dashboards, 13
- CRISP-DM (cross-industry standard process for data mining), 18
- EDA (exploratory data analysis)
  - defined*, 15–16
  - use cases versus solutions*, 18–19
  - walkthrough*, 17–18
- overlay/underlay, 20–22
- problem-centric approach
  - defined*, 15–16
  - use cases versus solutions*, 18–19
  - walkthrough*, 17–18
- SEMMA (Sample Explore, Modify, Model, and Assess), 18
- microservices architectures**, 5–6
- Migration Analytics**, 425

- mindfulness**, 128–129
- mindset**. *See* mental models
- mirror-image bias**, 110–111
- mirroring**, 69, 110–111
- missing data**, 86
- mlexend package**, 283
- model-driven telemetry (MDT)**, 64
- models**. *See* analytics models, building; mental models
- Monte Carlo simulation**, 202, 271
- moving averages**, 262
- MSE (mean squared error)**, 227
- MSRP (manufacturer's suggested retail price)**, 108
- Mudd, Philip**, 110
- multicollinearity**, 225
- muscle memory**, 102–103

## N

---

- narrative fallacy**, 107–108
- natural language processing (NLP)**, 165–166, 262–263
- negative correlation**, 224
- NETCONF (Network Configuration Protocol)**, 60
- Netflix recommender system**, 191–194
- NetFlow**
  - architecture of, 65
  - capabilities of, 65–66
  - data transport, 94
  - versions of, 65
- Network Configuration Protocol (NETCONF)**, 60
- network functions virtualization (NFV)**, 5–6, 51–52, 365
- network infrastructure analytics use case**, 323–324, 441
  - data encoding, 328–336
  - data loading, 325–328
  - data visualization, 340–344
  - dimensionality reduction, 337–340

- DNA mapping and fingerprinting, 324–325
- environment setup, 325–328
- K-means clustering, 344–349
- machine learning-guided troubleshooting, 350–353
- search challenges and solutions, 331–336
- Network Time Protocol (NTP)**, 87–88
- Network Watcher**, 68
- networking, social**, 133–135
- networking data**, 35–37
  - business and applications data relative to, 42–44
  - control plane
    - activities in*, 41
    - data examples*, 46–47
    - defined*, 37
  - control plane communication, 38
  - data access
    - container on box*, 74–75
    - control plane data*, 67–68
    - data plane traffic capture*, 68–70
    - DPI (deep packet inspection)*, 56
    - external data for context*, 89
    - IoT (Internet of Things) model*, 75–76
    - methods of*, 55–57
    - observation effect*, 88
    - packet data*, 70–74
    - panel data*, 88
    - pull data availability*, 57–61
    - push data availability*, 61–67
    - timestamps*, 87–88
  - data manipulation
    - KPIs (key performance indicators)*, 86–87
    - made-up data*, 84–85
    - missing data*, 86
    - standardized data*, 85
  - data plane
    - activities in*, 41
    - data examples*, 47–49
    - defined*, 37
- data structure
  - JSON (JavaScript Object Notation)*, 82–83
  - semi-structured data*, 84
  - structured data*, 82
  - unstructured data*, 83–84
- data transport, 89–90
  - CLI (command-line interface) scraping*, 92
  - HLD (high-level design)*, 90
  - IPFIX (IP Flow Information Export)*, 95
  - LLD (low-level design)*, 90
  - NetFlow*, 94
  - other data*, 93
  - sFlow*, 95
  - SNMP (Simple Network Management Protocol)*, 90–92
  - SNMP (Simple Network Management Protocol) traps*, 93
  - Syslog*, 93–94
  - telemetry*, 94
- data types, 76–77
  - continuous numbers*, 78–79
  - discrete numbers*, 79
  - higher-order numbers*, 81–82
  - interval scales*, 80
  - nominal data*, 77–78
  - ordinal data*, 79–80
  - ratios*, 80–81
- encoding methods, 85
- management plane
  - activities in*, 40–41
  - data examples*, 44–46
  - defined*, 37
- network virtualization, 49–51
- OpenStack nodes, 39–40
- planes, combining across virtual and physical environments, 51–52
- sample network, 38
- networks, computer. *See also* IBN (intent-based networking)**
  - DNA (Digital Network Architecture), 428

IBN (intent-based networking), 11, 428  
 NFV (network functions virtualization), 51–52  
 overlay/underlay, 20–22  
 planes of operation, 36–37  
     *business and applications data relative to*, 42–44  
     *combining across virtual and physical environments*, 51–52  
     *control plane*, 37, 41, 46–47  
     *control plane communication*, 38  
     *data plane*, 37, 41, 47–49  
     *illustrated*, 438  
     *management plane*, 37, 40–41, 44–46  
     *network virtualization*, 49–51  
     *NFV (network functions virtualization)*, 51–52  
     *OpenStack nodes*, 39–40  
     *sample network*, 38  
     *virtualized environment*, 438  
 SD-WANs (software-defined wide area networks), 20  
 virtualization, 49–51  
 networks, neural. *See* neural networks  
 neural networks, 11, 252–258  
 next-best-action analysis, 193  
 next-best-offer analysis, 193  
 NFV (network functions virtualization), 5–6, 51–52, 365  
 Ng, Andrew, 267  
 N-grams, 263  
 NLP (natural language processing), 165–166, 262–263  
 NLTK, 263  
 nltk package, 283, 328  
 noise reduction, syslog telemetry use case, 360–362  
 nominal data, 77–78  
 normal distributions, 222–223  
 normality, tests for, 311–313  
 not-invented-here syndrome, 122  
 novelty detection, 153–155  
 np (numpy package), 313

NTOP, 68  
 NTP (Network Time Protocol), 87–88

## numbers

continuous, 78–79  
 discrete, 79  
 higher-order, 81–82  
 interval scales, 80  
 nominal data, 77–78  
 ordinal data, 79–80  
 ratios, 80–81

numpy package, 283, 313

## O

---

objects, groupby, 293–296  
 observation, 138–139  
 observation effect, 88  
 Occam's razor, 120  
 one-hot encoding, 232–233, 336  
 oneM2M, 75  
 Open Shortest Path First (OSPF), 41, 61, 357  
 open source software, 5–6, 11, 433–434  
 OpenNLP, 263  
 OpenStack, 5–6, 39–41  
 operation, planes of. *See* planes of operation  
 operations research, 214  
 operators, logical AND, 306  
 optimization, business model, 201–202  
 optimization use cases, 186–188  
 orchestration, 11  
 ordinal data, 79–80  
 ordinal numbers, 232  
 orthodoxies, 139–140  
 OSPF (Open Shortest Path First), 41, 61, 357  
 outcome bias, 124  
 out-group bias, 120  
 outlier analysis, 153–155, 307–310, 318–320



*Outliers* (Gladwell), 99  
 overfitting, 219  
 overlay, analytics as, 20–22

## P

PACF (partial autocorrelation function), 262

### packages

fillna, 342–343  
 Gensim, 264, 283, 328, 331–332  
 importing, 390  
 matplotlib, 283  
 mlexend, 283  
 nltk, 283, 328  
 numpy, 283, 313  
 pandas, 283, 346, 357–360  
 pylab, 283  
 scipy, 283  
 sklearn, 283  
 statsmodels, 283  
 table of, 283–284  
 wordcloud, 283

### packets

file loading, 390  
 HTTP (Hypertext Transfer Protocol), 71–72  
 IP (Internet Protocol), 390–391  
   *packet counts*, 395–397  
   *packet protocols*, 398  
 IPsec (Internet Protocol Security), 73–74  
 IPv4, 70–74  
 port assignments, 393–394  
 SSL (Secure Sockets Layer), 74  
 TCP (Transmission Control Protocol), 71–72, 391  
 VXLAN (Virtual Extensible LAN), 74

pairwise ANOVA (analysis of variance), 317

### pandas package, 283

apply, 346  
 fillna, 342–343  
 log analysis with, 357–360

panel data, 88, 225–226

parsimony, law of, 120, 152

partial autocorrelation function (PACF), 262

partnerships, Cisco, 433

part-of-speech tagging, 263

pattern mining, 243–244

pattern recognition, 190

PCA (principal component analysis), 233–234

  network infrastructure analytics use case, 339–340

  syslog telemetry use case, 372–373

Pearson's correlation coefficient, 225, 236

perceptrons, 252

perspectives, gaining new, 130–131

phi, 262

physical environments, combining planes across, 51–52

pivoting, 142

planes of operation, 36–37

  business and applications data relative to, 42–44

  combining across virtual and physical environments, 51–52

  control plane

*activities in*, 41

*communication*, 38

*data examples*, 46–47

*defined*, 37

  data plane. *See also* data plane analytics use case

*activities in*, 41

*data examples*, 47–49

*defined*, 37

  illustrated, 438

  management plane

*activities in*, 40–41

*data examples*, 44–46

*defined*, 37

  network virtualization, 49–51

  NFV (network functions virtualization), 51–52

  OpenStack nodes, 39–40

- sample network, 38
- virtualized environments, 438
- planning, capacity, 180–181**
- platform crashes, statistics use case for, 288–299**
  - apply method, 295–296
  - box plot, 297–298
  - crash counts by product ID, 294–295
  - crash counts/rate comparison plot, 298–299
  - crash rates by product ID, 296–298
  - crashes by platform, 292
  - data scaling, 298
  - dataframe filtering, 290–292
  - groupby object, 293–296
  - horizontal bar chart, 289–290
  - lambda function, 296
  - overall crash rates, 292
  - router reset reasons, 290
  - simple bar chart, 289
  - value\_counts function, 288–289
- Platform for Network Data Analytics (PNDA), 433**
- platforms, Cisco analytics solutions, 433**
- plots**
  - box, 221–222
  - cluster scatterplot, 410–411
  - defined, 220
  - platform crashes example, 297–299
  - Q-Q (quartile-quantile), 220, 311–312
  - software crashes example, 300–305
- PNDA (Platform for Network Data Analytics), 433**
- polynomial regression, 247**
- population variance, 167**
- ports**
  - assignments, 393–394
  - mirroring, 69
  - per-host port analysis, 403
  - profiles, 407–408
    - full*, 413–419
    - source*, 419–422
  - SME port clustering, 407–413
    - cluster scatterplot*, 410–411
    - host patterns*, 411–413
    - K-means clustering*, 408–410
    - port profiles*, 407–408
- positive correlation, 224**
- post-algorithmic era, 147–148**
- post-hoc testing, 317**
- preconceived notions, 107–108**
- Predictably Irrational* (Ariely), 108**
- prediction of trends, use cases for, 190–191**
- Predictive Analytics* (Siegel), 148**
- predictive maintenance use cases, 188–189**
- predictive maturity, 8**
- preemptive analytics, 9**
- preemptive maturity, 8**
- PrefixScan, 244**
- prescriptive analytics, 9**
- priming effect, 109, 151**
- principal component analysis (PCA), 233–234**
  - network infrastructure analytics use case, 339–340
  - syslog telemetry use case, 372–373
- proactive maturity, 8**
- probability, 228**
- problem-centric approach**
  - defined, 15–16
  - use cases versus solutions, 18–19
  - walkthrough, 17–18
- process, analytics, 437**
- profiles, port, 407–408**
  - full, 413–419
  - source, 419–422
- pro-innovation bias, 121**
- psychology use cases, 209–210**
- publisher/subscriber environment, 29**
- pub/sub bus, 29**
- pull data availability**
  - CLI (command-line interface) scraping, 59, 92

NETCONF (Network Configuration Protocol), 60  
 SNMP (Simple Network Management Protocol), 57–59  
 unconventional data sources, 60–61  
 YANG (Yet Another Next Generation), 60  
**pull methods**, 28–29  
**push data availability**  
   IPFIX (IP Flow Information Export), 64–67, 95  
   NetFlow, 65–66, 94  
   sFlow, 67, 95  
   SNMP (Simple Network Management Protocol) traps, 61–62, 93  
   Syslog, 62–63, 93–94  
   telemetry, 63–64, 94  
**push methods**, 28–29  
**p-values**, 227, 314–317  
**pylab package**, 283  
**pyplot**, 395  
**Python packages**. *See* **packages**

## Q

---

**Q-Q (quartile-quantile) plots**, 220, 311–312  
**qualitative data**, 77–78  
**queries (SQL)**, 82  
**questioning**  
   CARESS technique, 137  
   example of, 135–137  
   “Five whys”, 137–138

## R

---

**race bias**, 97–98  
**radio frequency identification (RFID)**, 210–211  
**random forest**, 250–251  
**ratios**, 80–81  
**RCA (root cause analysis)**, 184  
**RcmdrPLugin.temis**, 263  
**reactive maturity**, 7–8

**recency bias**, 111  
**recommender systems**, 191–194  
**reconciling data**, 29  
**recurrent neural networks (RNNs)**, 254–256  
**regression analysis**, 101–102, 246–247  
**reinforcement learning**, 173, 212–213  
**relational database management system (RDBMS)**, 82  
**Remote SPAN (RSPAN)**, 69  
**reset\_index function**, 414  
**retention use cases**, 202–204  
**retrieval of information**  
   algorithms, 263–264  
   use cases, 185–186  
**reward functions**, 186  
**RFIS (radio frequency identification)**, 210–211  
**ridge regression**, 247  
**right skewed distribution**, 310  
**RNNs (recurrent neural networks)**, 254–256  
**roles**  
   analytics experts, 25  
   analytics infrastructure model, 24–25  
   business domain experts, 25  
   data domain experts, 25  
   data scientists, 25  
**root cause analysis (RCA)**, 184  
**RSBMS (relational database management system)**, 82  
**RSPAN (Remote SPAN)**, 69  
**R-squared**, 227  
**Rube Goldberg machines**, 151–152  
**rules, association**, 240–243

## S

---

**Sample Explore, Modify, Model, and Assess (SEMMA)**, 18  
**Sankey diagrams**, 199  
**SAS, Cisco's partnership with**, 433  
**scaling data**, 298

- scatterplots, 410–411
- scheduling use cases, 194–195
- scipy package, 283
- scraping, CLI (command-line interface), 59
- SDA (Secure Defined Access), 428
- SDN (software-defined networking), 61, 365
- SD-WANs (software-defined wide area networks), 20
- searches, network infrastructure analytics use case, 331–336
- seasonality, 261
- Secure Defined Access (SDA), 428
- Secure Sockets Layer (SSL), 74
- security signatures, 214
- segmentation, customer, 160
- self-leveling wireless networks, 186
- SELs (system event logs), 62
- semi-structured data, 84
- SEMMA (Sample Explore, Modify, Model, and Assess), 18
- sentiment analysis, 266–267
- sequential pattern mining, 243–244
- sequential patterns, 197
- service assurance
  - analytics infrastructure model with, 33
  - defined, 11–12
  - Service Assurance Analytics, 425
  - use cases for, 195–197
- service-level agreements (SLAs), 11–12, 196
- The Seven Habits of Highly Successful People* (Covey), 10
- severities, syslog, 359–360
- sFlow, 67, 95
- Shapiro-Wilk test, 311
- Siegel, Eric, 148
- signatures, security, 214
- Simple Network Management Protocol. *See* SNMP (Simple Network Management Protocol)
- simulations, 271
- Sinek, Simon, 148
- singular value decomposition (SVD), 265
- six hats thinking approach, 132–133
- sklearn package, 283
- SLAs (service-level agreements), 11–12, 196
- slicing data, 286
- small numbers, mental models and, 117–118
- smart meters, 189
- smart society, 213–214
- Smarter, Faster, Better* (Duhigg), 99
- SME analysis
  - dataframe and visualization library loading, 394
  - host analysis, 399–404
  - IP address packet counts, 395–397
  - IP packet protocols, 398
  - MAC addresses, 398
  - output, 404–406
  - time series counts, 395
  - timestamps and time index, 394–395
  - topology mapping information, 398
- SME port clustering, 407–413
  - cluster scatterplot, 410–411
  - host patterns, 411–413
  - K-means clustering, 408–410
  - port profiles, 407–408
- SMEs (subject matter experts), 1–2
- SNMP (Simple Network Management Protocol), 28–29
  - data transport, 90–92
  - pull data availability, 57–59
  - traps, 61–62, 93
- social filtering solution, 191
- soft data, 150
- software
  - crashes use case, 299–305
    - box plots*, 300–305
    - dataframe filtering*, 300
    - dataframe grouping*, 299–300
  - defect analysis use cases, 178–179
  - open source, 5–6, 11
- software-defined networking (SDN), 61, 365

- software-defined wide area networks (SD-WANs), 20
- solution design, 150, 274
  - breadth of focus, 274
  - operationalizing as use cases, 281
  - time expenditure, 274–275
  - workflows, 282
- sorting dataframes, 326–327
- source IP address packet counts, 396
- source port profiles, 419–422
- SPADE, 244
- SPAN (Switched Port Analyzer), 69
- Spanning Tree Protocol (STP), 41
- Spark, 28–29
- SPC (statistical process control), 189
- Spearman's rank, 225, 236
- split function, 368
- SQL (Structured Query Language), 29, 82
- SSE (sum of squares error), 227
- SSL (Secure Sockets Layer), 74
- standard deviation, 167, 222–223
- standardizing data, 85
- Stanford CoreNLP, 263
- Starbucks, 110
- Start with Why* (Sinek), 148
- stationarity, 261
- statistical analysis, 440. *See also* statistics use cases
  - ANOVA (analysis of variance), 227
  - Bayes' theorem, 228–230
  - box plots, 221–222
  - correlation, 224–225
  - defined, 220
  - longitudinal data, 225–226
  - normal distributions, 222–223
  - probability, 228
  - standard deviation, 222–223
  - statistical inference, 228
- statistical process control (SPC), 189
- statistics use cases, 153, 285
  - anomalies and outliers, 153–155
  - anomaly detection, 318–320
  - ANOVA (analysis of variance), 305–310
  - data filtering*, 305–306
  - describe function*, 308
  - drop command*, 309
  - groupby command*, 307
  - homogeneity of variance*, 313–318
  - outliers, dropping*, 307–310
  - pairwise*, 317
- benchmarking, 155–157
- classification, 157–158
- clustering, 158–160
- correlation, 160–162
- data loading and exploration, 286–288
- data transformation, 310
- data visualization, 163–165
- descriptive analytics, 167–168
- NLP (natural language processing), 165–166
- normality, tests for, 311–313
- platform crashes example, 288–299
  - apply method*, 295–296
  - box plot*, 297–298
  - crash counts by product ID*, 294–295
  - crash counts/rate comparison plot*, 298–299
  - crash rates by product ID*, 296–298
  - crashes by platform*, 292–294
  - data scaling*, 298
  - dataframe filtering*, 290–292
  - groupby object*, 293–296
  - horizontal bar chart*, 289–290
  - lambda function*, 296
  - overall crash rates*, 292
  - router reset reasons*, 290
  - simple bar chart*, 289
  - value\_counts function*, 288–289
- software crashes example, 299–305
  - box plots*, 300–305
  - dataframe filtering*, 300
  - dataframe grouping*, 299–300
- time series analysis, 168–169
- voice, video, and image recognition, 170

- statsmodels package, 283
- status-quo bias, 122
- Stealthwatch, 6, 65, 427
- Steltzner, Adam, 202
- stemming, 263
- stepwise regression, 247
- stop words, 263, 329
- STP (Spanning Tree Protocol), 41
- strategic thinking, 9
- streaming data, 30
- structure. *See* data structure
- Structured Query Language (SQL), 29, 82
- subject matter experts (SMEs), 1–2
- Sullenberger, Chesley “Sully”, 99–100
- Sully*, 99–100
- sum of squares error (SSE), 227
- sums-of-squares distance measures, 167
- sunk cost fallacy, 122
- supervised machine learning, 151, 246
- support vector machines (SVMs), 258–259
- survivorship bias, 118–119
- SVD (singular value decomposition), 265
- SVMs (support vector machines), 258–259
- swim lanes configuration, 161
- Switched Port Analyzer (SPAN), 69
- switches, virtual, 69–70
- syslog, 62–63, 93–94
- syslog telemetry use case, 355, 441
  - data encoding, 371–373
  - data preparation, 356–357, 369–371
  - high-volume producers, identifying, 362–366
  - K-means clustering, 373–375
  - log analysis with pandas, 357–360
  - machine learning-based evaluation, 366–367
  - noise reduction, 360–362
  - OSPF (Open Shortest Path First) routing, 357
  - syslog severities, 359–360
  - task list, 386–387

- transaction analysis, 379–386
  - apriori function*, 381–382
  - data preparation*, 379
  - dictionary-encoded message lookup*, 380–381
  - groupby method*, 380
  - log message groups*, 382–386
  - tokenization*, 381
- word cloud visualization, 367–369, 375–379

System 1/System 2 intuition, 102–103

system event logs (SELs), 62

## T

---

tables, contingency, 267–268

tags, data transport, 93

*Talent Is Overrated* (Colvin), 103

*Taming the Big Data Tidal Wave* (Franks), 147

task lists

- data plane analytics use case, 423–424
- syslog telemetry use case, 386–387

TCP (Transmission Control Protocol)

- packet data, 71–72
- packet format, 391

tcpdump, 68

telemetry, 441

- analytics infrastructure model, 27–28
- architecture of, 63
- capabilities of, 64
- data transport, 94
- EDT (event-driven telemetry), 64
- MDT (model-driven telemetry), 64
- syslog telemetry use case, 355
  - data encoding*, 371–373
  - data preparation*, 356–357, 369–371
  - high-volume producers, identifying*, 362–366
  - K-means clustering*, 373–375
  - log analysis with pandas*, 357–360

- machine learning-based evaluation*, 366–367
- noise reduction*, 360–362
- OSPF (Open Shortest Path First) routing*, 357
- syslog severities*, 359–360
- task list*, 386–387
- transaction analysis*, 379–386
- word cloud visualization*, 367–369, 375–379
- term document matrix**, 336
- term frequency-inverse document frequency (TF-IDF)**, 232
- terminology**, 7
- tests**, 219, 220
  - F-tests, 227
  - Levene's, 313
  - normality, 311–313
  - post-hoc testing, 317
  - Shapiro-Wilk, 311
- Tetration**, 6, 430–431
- text analysis**, 256–262
  - information retrieval, 263–264
  - NLP (natural language processing), 262–263
  - nominal data, 77–78
  - ordinal data, 79–80
  - sentiment analysis, 266–267
  - topic modeling, 265–266
- TF-IDF (term frequency-inverse document frequency)**, 232
- thinking**
  - innovative, 127–128, 439
    - associative thinking*, 131–132
    - bias and*, 128
    - breaking anchors*, 140
    - cognitive trickery*, 143
    - crowdsourcing*, 133–134
    - defocusing*, 140
    - experimentation*, 141–142
    - inverse*, 204–206
    - inverse thinking*, 139–140
    - lean thinking*, 142
    - metaphoric thinking*, 130–131
    - mindfulness*, 128–129
    - networking*, 133–135
    - observation*, 138–139
    - perspectives*, 130–131
    - questioning*, 135–138
    - quick innovation wins*, 143–144
    - six hats thinking approach*, 132–133
    - unpriming*, 140
  - strategic, 9
- Thinking Fast and Slow* (Kahneman), 102
- thinking hats approach**, 132–133
- thrashing**, 122
- tilde (~)**, 291–292, 370
- time index**
  - creating from timestamp, 357–358
  - data plane analytics use case, 394–395
- time series analysis**, 168–169, 259–262
- time series counts**, 395
- time to failure**, 183–184
- TimeGrouper**, 395
- timestamps**, 87–88
  - creating time index from, 357–358
  - data plane analytics use case, 394–395
- tm**, 263
- tokenization**, 263, 328
  - syslog telemetry use case, 371
  - tokenization, 381
- topic modeling**, 265–266
- traffic capture, data plane**, 68–69
  - ERSPAN (Encapsulated Remote Switched Port Analyzer), 69
  - inline security appliances, 69
  - port mirroring, 69
  - RSPAN (Remote SPAN), 69
  - SPAN (Switched Port Analyzer), 69
  - virtual switch operations, 69–70
- training data**, 219
- transaction analysis**
  - explained, 193, 197–199

- syslog telemetry use case, 379–386
  - apriori function*, 381–382
  - data preparation*, 379
  - dictionary-encoded message lookup*, 380–381
  - groupby method*, 380
  - log message groups*, 382–386
  - tokenization*, 381
- transformation, data, 310
- translation, language, 11
- Transmission Control Protocol (TCP), 391
- transport of data, 89–90
  - analytics infrastructure model, 26–28
  - CLI (command-line interface) scraping, 92
  - HLD (high-level design), 90
  - IPFIX (IP Flow Information Export), 95
  - LLD (low-level design), 90
  - NetFlow, 94
  - other data, 93
  - sFlow, 95
  - SNMP (Simple Network Management Protocol), 90–92, 93
  - Syslog, 93–94
  - telemetry, 94
- traps (SNMP), 61–62
- trees, decision
  - example of, 249–250
  - random forest, 250–251
- trends, prediction of, 11–12, 190–191
- troubleshooting, machine learning-guided, 350–353
- truncation, 263
- TrustSec, 427
- Tufte, Edward, 163
- Tukey post-hoc test, 317
- tunnel vision, 107
- types, 76–77
  - continuous numbers, 78–79
  - discrete numbers, 79
  - higher-order numbers, 81–82
  - interval scales, 80

- nominal data, 77–78
- ordinal data, 79–80
- ratios, 80–81

## U

---

- UCS (Unified Computing System), 62
- unconventional data sources, 60–61
- underlay, 20–22
- Unified Computing System (UCS), 62
- unpriming, 140
- unstructured data, 83–84
- unsupervised machine learning
  - association rules, 240–243
  - clustering, 234–239
  - collaborative filtering, 244–246
  - defined, 151, 234
  - sequential pattern mining, 243–244
- use cases, 439
  - algorithms, 3–4
  - autonomous applications, 200–201
  - benefits of, 147–149, 273–274
  - building
    - analytics infrastructure model*, 275–276
    - analytics solution design*, 274
    - code*, 280–281
    - data*, 276–278
    - data science*, 278–280
    - environment setup*, 282–284
    - time expenditure*, 440
    - workflows*, 282
  - business model analysis, 200–201
  - business model optimization, 201–202
  - churn and retention, 202–204
  - control plane analytics, 441
  - data plane analytics, 389, 442
    - assets*, 422–423
    - data loading and exploration*, 390–394
  - full port profiles*, 413–419
  - investigation task list*, 423–424
  - SME analysis*, 394–406



- SME port clustering*, 407–413
  - source port profiles*, 419–422
- defined, 18–19, 150
- development, 2–3
- dropouts and inverse thinking, 204–206
- engagement models, 206–207
- examples of, 32–33
- fraud and intrusion detection, 207–209
- healthcare and psychology, 209–210
- IT analytics, 170
  - activity prioritization*, 170–173
  - asset tracking*, 173–175
  - behavior analytics*, 175–178
  - bug and software defect analysis*, 178–179
  - capacity planning*, 180–181
  - event log analysis*, 181–183
  - failure analysis*, 183–185
  - information retrieval*, 185–186
  - optimization*, 186–188
  - prediction of trends*, 190–191
  - predictive maintenance*, 188–189
  - recommender systems*, 191–194
  - scheduling*, 194–195
  - service assurance*, 195–197
  - transaction analysis*, 197–199
- logistics and delivery models, 210–212
- machine learning and statistics, 153
  - anomalies and outliers*, 153–155
  - benchmarking*, 155–157
  - classification*, 157–158
  - clustering*, 158–160
  - correlation*, 160–162
  - data visualization*, 163–165
  - descriptive analytics*, 167–168
  - NLP (natural language processing)*, 165–166
  - time series analysis*, 168–169
  - voice, video, and image recognition*, 170
- network infrastructure analytics, 323–324, 441
  - data encoding*, 328–331, 336–337
  - data loading*, 325–328
  - data visualization*, 340–344
  - dimensionality reduction*, 337–340
  - DNA mapping and fingerprinting*, 324–325
  - environment setup*, 325–328
  - K-means clustering*, 344–349
  - machine learning-guided troubleshooting*, 350–353
  - search challenges and solutions*, 331–336
- operationalizing solutions as, 281
- packages for, 283–284
- reinforcement learning, 212–213
- smart society, 213–214
- versus solutions, 18–19
- statistics, 153, 285, 440
  - anomalies and outliers*, 153–155
  - anomaly detection*, 318–320
  - ANOVA (analysis of variance)*, 305–310
  - benchmarking*, 155–157
  - classification*, 157–158
  - clustering*, 158–160
  - correlation*, 160–162
  - data loading and exploration*, 286–288
  - data transformation*, 310
  - data visualization*, 163–165
  - descriptive analytics*, 167–168
  - NLP (natural language processing)*, 165–166
  - normality, tests for*, 311–313
  - platform crashes example*, 288–299
  - software crashes example*, 299–305
  - time series analysis*, 168–169
  - voice, video, and image recognition*, 170
- summary table, 215
- syslog telemetry, 355
  - data encoding*, 371–373
  - data preparation*, 356–357, 369–371

*high-volume producers, identifying*, 362–366  
*K-means clustering*, 373–375  
*log analysis with pandas*, 357–360  
*machine learning-based evaluation*, 366–367  
*noise reduction*, 360–362  
*OSPF (Open Shortest Path First) routing*, 357  
*syslog severities*, 359–360  
*task list*, 386–387  
*transaction analysis*, 379–386  
*word cloud visualization*, 367–369, 375–379

## V

---

validation, 219  
 value\_counts function, 288–289, 396, 400, 403  
 values, key/value pairs, 82–83  
 variables, dummy, 232  
 variance, analysis of. *See* ANOVA (analysis of variance)  
 vectorized features, finding, 338  
 video recognition use cases, 170  
 views, dataframe, 329–330, 347  
 Viptela, 20  
 Virtual Extensible LAN (VXLAN), 74  
 virtual private networks (VPNs), 20  
 virtualization  
   network, 49–51  
   NFV (network functions virtualization), 51–52, 365

planes of operation, 51–52, 438  
 virtual switch operations, 69–70  
 VPNs (virtual private networks), 20  
 VXLAN (Virtual Extensible LAN), 74

voice recognition, 11, 170

VPNs (virtual private networks), 20

VXLAN (Virtual Extensible LAN), 74

## W

---

Wald, Abraham, 118–119

What You See Is All There Is (WYSIATI), 118

whys, “five whys” technique, 137–138

Windows Management Instrumentation (WMI), 61

Wireshark, 68

wisdom of the crowd, 250

WMI (Windows Management Instrumentation), 61

word clouds, 367–369, 375–379

wordcloud package, 283

workflows, designing, 282

WYSIATI (What You See Is All There Is), 118

## X-Y-Z

---

XGBoost, 252

YANG (Yet Another Next Generation), 60

Yau, Nathan, 163

Yet Another Next Generation (YANG), 60

zero price effect, 123