



HIGH-TECH

Big Data and Machine Learning



by Brett S. Martin



HIGH-TECH

Big Data and Machine Learning

by Brett S. Martin

NORWOODHOUSE  PRESS

Cover: Facilities with large numbers of powerful computers make big data and machine learning possible.

Norwood House Press
P.O. Box 316598
Chicago, Illinois 60631

For information regarding Norwood House Press, please visit our website at:
www.norwoodhousepress.com or call 866-565-2900.

PHOTO CREDITS: Cover: © Scanrail1/Shutterstock Images; © Akhenaton Images/Shutterstock Images, 29; © Alexander Kirch/Shutterstock Images, 33; © Andrew Matthews/Press Association/PA Wire URN:30099549/AP Images, 12; © Bunditinay/Shutterstock Images, 28; © Chris Radburn/Press Association/PA Wire URN:28954785/AP Images, 37; © dotstock/Shutterstock Images, 16; © Eric Risberg/AP Images, 36; © Everett Collection/Shutterstock Images, 9; © Gorodenkoff/Shutterstock Images, 25; © Henny Ray Abrams/AP Images, 42; © Jon Simon/Feature Photo Service for IBM/AP Images, 38; © Kaspars Grinvalds/Shutterstock Images, 14; © Kite_rin/Shutterstock Images, 10; © Mark Agnor/Shutterstock Images, 23; © mavo/Shutterstock Images, 5; © Michael Dwyer/AP Images, 32; © Pe3k/Shutterstock Images, 19; © Pierre-Olivier/Shutterstock Images, 40; © rawpixel.com/Shutterstock Images, 21; © Wdnet Creation/Shutterstock Images, 18; © Worawee Meepian/Shutterstock Images, 26

Content Consultant: Jake Williams, Assistant Professor, Information Science, Drexel University

Hardcover ISBN: 978-1-59953-938-6
Paperback ISBN: 978-1-68404-217-3

© 2019 by Norwood House Press.

All rights reserved.

No part of this book may be reproduced without written permission from the publisher.

Library of Congress Cataloging-in-Publication Data

Names: Martin, Brett S., author.
Title: Big data and machine learning / by Brett Martin.
Description: Chicago, Illinois : Norwood House Press, [2018] | Series: Tech bytes | Includes bibliographical references and index.
Identifiers: LCCN 2018003241 (print) | LCCN 2018011594 (ebook) | ISBN 9781684042227 (ebook) | ISBN 9781599539386 (hardcover : alk. paper) | ISBN 9781684042173 (pbk. : alk. paper)
Subjects: LCSH: Big data--Juvenile literature. | Machine learning--Juvenile literature.
Classification: LCC QA76.9.B45 (ebook) | LCC QA76.9.B45 M37 2018 (print) | DDC 005.7--dc23
LC record available at <https://lcn.loc.gov/2018003241>

312N—072018

Manufactured in the United States of America in North Mankato, Minnesota.

CONTENTS

Chapter 1: An Explosion of Big Data	4
Chapter 2: Solving Big Data Challenges	15
Chapter 3: Putting Big Data and Machine Learning to Work	24
Chapter 4: A Future of Possibilities	33
Glossary	44
For More Information	45
Index	46
About the Author	48


Note: Words that are **bolded** in the text are defined in the glossary.

An Explosion of Big Data

A young woman is walking downtown. As she passes her favorite department store, her smartphone beeps. The store's computer system knows she is close by. It tracks her shopping habits. The computer system is aware that she recently bought a new coat. It knows that many people who buy a coat also buy a scarf. So, the store sends a coupon for a scarf to her phone. But she is not interested in shopping. Instead, she walks down a block to the coffee shop.

Her phone beeps again. The coffee shop's mobile app knows she is in the store. She pulls her coffee shop loyalty card up on the app. She has earned a free drink. Otherwise, she would have paid with the app.

Drink in hand, she gets comfortable in a chair. She sets up her laptop. Then she logs on to a social media website. She sends messages to her friends. The site shows her people it thinks she might know. It asks if she wants to connect

A young woman with voluminous, curly, reddish-brown hair is smiling warmly. She is wearing a dark blue t-shirt and holding a white ceramic coffee cup with her right hand. Her left hand is resting on a wooden surface, possibly a table. The background is softly blurred, suggesting an indoor setting like a cafe or office.

Personalized coupons
are just one way stores
use shoppers' data.

with them. Some are coworkers. Others are friends of friends. While sipping her coffee, she browses the site. She makes a new connection then likes a local bakery's page. Almost immediately, a coupon pops up. If she buys a cake from the bakery today, she'll get 15 percent off.

From the department store to the social media network, the woman's entire experience is closely linked with big data and machine learning. Big data is a body of information that is large and varied. Big data is generated very quickly. It can be challenging to **analyze** and interpret. But with the right tools, big data can be extremely valuable. Machine learning is when computer systems are trained to learn on their own. Humans program the computers to learn. But they do not tell



Location-Based Advertising

Advertising was very different before the Internet. Companies often placed expensive ads in newspapers. The entire newspaper readership would see the same ad. Today, a personalized online ad can be sent to an audience of one. This is partially because of big data. It allows companies to advertise by specific location. Data from a person's phone tells his or her location. The data alerts companies when the customer is close to the store. The company's computer system automatically sends an ad to the customer. The goal is to draw the person inside to spend money.

them what to do. They learn by looking at data.

When big data and machine learning work together, the result can seem like magic. **Geolocation** data from the woman's phone told the store she was nearby. When she went online, the social network identified people she might know. This was based on her profile.

The site looked at where she went to college and worked. It looked at who she sent messages to. It used this data to recommend new connections.

A Short History of Big Data

People have kept and analyzed records, or data, since prehistoric times. Ancient

humans made notches in sticks and bones. They did this to keep track of food and other supplies. Around 2400 BCE, the abacus was invented. The abacus is a device with beads that slide along strings or rods. It helps people do math. The origin is still in dispute. Some historians say it was invented in China. Others credit the Babylonians or Egyptians. Ancient Greeks invented another early calculating device between 100 and 200 CE. Called the Antikythera mechanism, it had gears to track the solar system.

In 1663, statistician John Graunt used data to warn where the bubonic plague would strike. He noticed trends in where people were dying. He identified the locations and people most at risk of infection.

In the 1880s, a young engineer at the US Census Bureau invented a machine. Herman Hollerith's device was called a tabulating machine. The machine punched holes in paper cards, then counted the results. The technology analyzed census data in three months. The project would have taken ten years without the device. Hollerith went on to start the technology company IBM.

In the 1960s, government researchers wanted to share information more easily. In 1962, researcher Joseph Carl Robnett Licklider at the Massachusetts Institute of Technology (MIT) made a proposal. He wanted to connect computers together to share data.

Meanwhile, companies started developing technologies that became



The Rise of the Internet

It is hard to imagine life today without the Internet. Many people use it daily. But it did not start out as a way for people to access information. In the 1960s, computers were large and immovable. Sharing data between computers required sending computer tapes through the mail. In 1965, a computer scientist at MIT connected two computers. One was in Massachusetts, and the other was in California. They connected over a phone line. This network was called Advanced Research Projects Agency Network, or ARPANET. This network became the Internet.

supercomputers. Supercomputers are fast, high-performance systems. They can process data very quickly. Work that would take a human thousands of years to complete takes a supercomputer only a few seconds. The Advanced Scientific Computer was one of the first supercomputers. Texas Instruments designed it between 1966 and 1973.

By 1965, connecting computers had made it possible to collect large amounts of data. That year, the US government proposed the first large **data center**. It wanted the data center to store 742 million tax returns. It was also supposed to store 175 million sets of fingerprints. The government planned to store the data on **magnetic computer tape**. These tapes

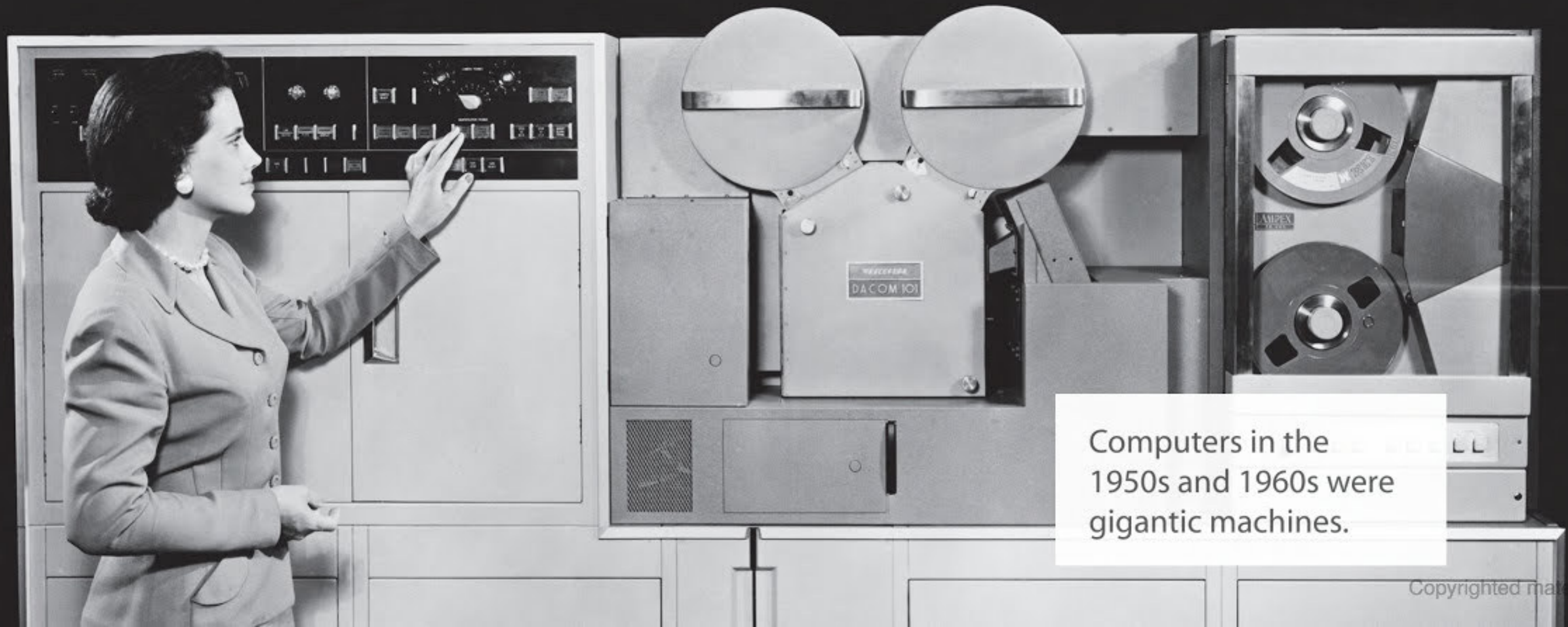


DID YOU KNOW?


In 1996, storing information digitally became cheaper than filing paper copies. This sparked a major shift in how information was kept.

would be stored in one location. However, the government dropped the project. The public worried about a potential invasion of people's privacy.

In 1989, the term *big data* first appeared in *Harper's Magazine*. This data was mostly customer data, such as people's addresses. Companies would use the data to send people marketing materials.



Computers in the 1950s and 1960s were gigantic machines.



Every time you use the Internet, you leave behind a trail of data.

This became known as junk mail. This was not big data as it is known today.

The World Wide Web is the most commonly used part of the Internet. The web became public in 1991. It connected people around the globe. This changed how people shopped and communicated. Internet users still create a lot of data.

We Produce and Consume Big Data Every Day

People across the planet produce data every single second. Everyday activities create data. Using a mobile phone or social media site creates data. Booking

a hotel or airline flight online produces data. Every time people log on to the Internet, they produce new data. So much information coming from so many sources has led to big data.

Data is coming from more places than ever before. Data sources are constantly growing. Trillions of devices called **sensors** now produce data. Vehicles, utility grids, and almost any new piece of machinery can have sensors. Big data is all of this information combined together.

Businesses use big data for many purposes. It makes marketing more effective. Companies can track patterns in what customers buy. This helps them design products or add features they know people want. Media companies use data to know what TV shows people like.

This lets them create new shows their audiences will enjoy.

Individual people also use big data. It can be used to recommend movies based on the movies a person has liked in the past. Banking applications allow people to see years of their own financial data. They can see how much they made and how much they spent in the last month,

DID YOU KNOW?



The amount companies spent on location-based advertising jumped from \$1.6 billion in 2013 to nearly \$15 billion in 2018.



year, or decade. They can use the data to make a budget for the future. Big data also lets people monitor their own health. The ways people use big data are always expanding.

Unlocking the Value of Big Data

For data to be useful, it must be analyzed. Programmers create computer software

Many people use the Internet and big data to track their money.

that finds useful patterns or connections. This process is known as analytics.

Machine learning is a common type of analytics. As machines see new data, they learn more. Faster computers can learn more quickly. Machine learning can apply complex math formulas to big data. It can find new information by analyzing big data. Machine learning can find patterns in data that human analysts fail to see.

Companies can use the information to understand what customers will buy. A common use of machine learning is

online recommendations. When people shop online, the online store will show related products based on their searches. This data is from other shoppers who performed similar searches. This is machine learning at work with big data.

Machine learning has helped people communicate with one another. Language

translation programs use machine learning to improve their translations. Translation programs train on data from thousands of translated texts. The data includes text in dozens of languages. Then the programs analyze patterns in the data. They use the patterns to decide how to translate a sentence from one language to another.



DID YOU KNOW?

More data was created in 2017 than in the previous 5,000 years of humanity. The amount of data produced doubles every two years.

Data at Your Fingertips

With so much data available, everyone can do tasks that used to require specialists. Travel agents used to book flights, hotels, or vacations for families. Apps now help people find the best deals. Personal devices tell people about their diet, sleep, and exercise. People can compare their