BEST BOOKS ON BIG DATA & DATA SCIENCE

In this article, we'll recommend books and other resources specifically about big data and data science.

(This article is part of our Tech Books & Talks Guide. Use the right-hand menu to navigate.)

Big Data: A Revolution That Will Transform How We Live, Work, and Think

Authors: Viktor Mayer-Schonberger and Kenneth Cukier

Big Data: A Revolution provides a broad overview of big data and the

impact that it's making on modern society. While by no means a technical book, it does provide a good high-level introduction to what big data is and how it's affecting practices in areas as diverse as:

- Fraud detection
- International law enforcement
- Linguistics
- Automated language translation

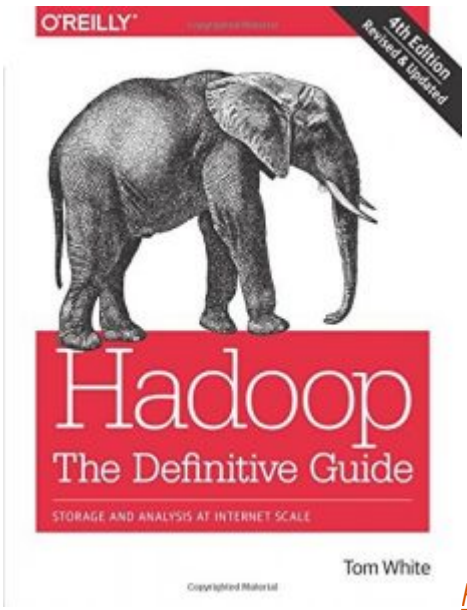
Well-suited for business managers and analysts, or maybe even C-level executives, *Big Data: A Revolution* provides insight and guidance on how industries should move forward in the wake of today's information revolution.

One of the book's central premises is the notion of "*why*, not *what*". For example, the book states:

> *"The era of big data challenges the way we live and interact with the world. Most strikingly, society will need to shed some of its obsession for causality in exchange for simple correlations: not knowing why but only what."*

Hadoop: The Definitive Guide, 4th Edition

Author: Tom White



Hadoop: The Definitive Guide is a big data book that's targeted at technical audiences. The book was originally published in 2009 and is currently sold as a 4th edition update.

Praised by developers and data engineers the world-over, *Hadoop: The Definitive Guide* provides how-to's on building and maintaining distributed, parallel processing data systems with Apache Hadoop (HDFS, MapReduce, and YARN). The 4th Edition update even goes into details on Hadoop 2 deployment, including technical details that you should know about YARN, HBase, Parquet, Flume, Crunch, Pig, Hive, and Spark.

The book also presents interesting case studies from the healthcare and genomic sciences industries.

Data Smart: Using Data Science to Transform Information Into Insight

Author: John Foreman

Business professionals love Data Smart, it's as simple as that! Written especially for data science newbies, Data Smart provides a really easy way for readers to grasp the concepts and techniques that underlie data science. Furthermore, the book provides step-by-step tutorials on how to execute these techniques in the ubiquitous Microsoft Excel. Some data science methods covered in Data Smart include:

- Cluster analysis (including k-means and k-medians methods)
- Linear programming for document classification
- Various forms of linear regression analysis
- Time series forecasting

Although this book won't teach you everything you need to know in order to start deploying large-scale analytics projects, it will help you learn the basic ABCs of data science and some of the methods comprising it.

Introduction to Machine Learning with Python: A Guide for Data Scientists

Authors: Andreas C. Müller and Sarah Guido

Introduction to Machine Learning with Python provides practical guidelines for building machine learning models with a variety of Python data science libraries. The book covers fundamental machine learning concepts and provides coding examples on various applications. The modeling guidelines include:

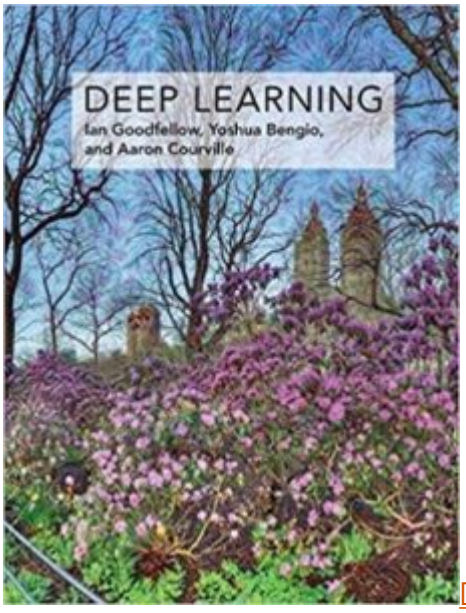
- Data preparation
- Model tuning and evaluation
- AI algorithms including supervised and unsupervised learning
- Statistical optimization techniques

Dedicated chapters on text and image data processing and visualization help users learn practical machine learning with Python end-to-end.

(See why Python is perfect for big data.)

Deep Learning

Authors: Aaron Courville, Ian Goodfellow, and Yoshua Bengio



Deep Learning is one of the most comprehensive books written on the subject by three of the most prominent experts in the deep learning domain. The book provides extensive mathematical details and sufficiently covers all of the relevant subject domains in statistical machine learning, including:

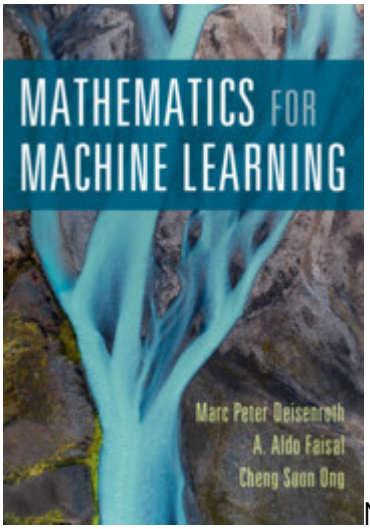
- Probability and information theory
- Optimization
- Numerical computation

The book provides general machine learning and mathematical modeling guidelines applicable to natural language processing, computer vision and autonomous driving, finance, healthcare data and bioinformatics, and entertainment, among others.

Who should use this book? It's most suitable for undergraduate and graduate students pursuing a career in machine learning. Professionals with background in statistics and computer science can also take advantage of the book Deep Learning in order to truly understand how machine learning works and delivers value to their business.

Mathematics for Machine Learning

Authors: Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong.



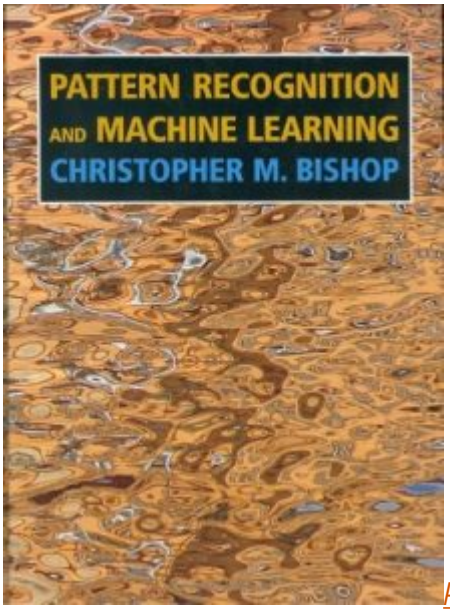
Mathematics is the most integral component of machine learning, which goes well beyond deep learning. Mathematics for Machine Learning takes a step back and covers a variety of basic machine learning concepts with great detail.

This book is a great starting point for professionals pursuing a shift in career toward machine learning and especially deep learning but lack the knowledge of the underlying mathematics. This book covers a variety of topics from calculus, linear algebra, probability, and optimization. It then looks into the central machine learning algorithms including Support Vector Machines and Principal Component Analysis.

This latter part of the book, however, is not exhaustive on machine learning techniques and algorithms, as a variety of central ML methods are not covered—but in this growing field, who can cover everything?

Pattern Recognition and Machine Learning

Author: Christopher Bishop



Pattern Recognition and Machine Learning is another great book about data science. In contrast to *Data Smart*, however, *Pattern Recognition* was written to satisfy the interests and technical capacities of already advanced information scientists and statisticians.

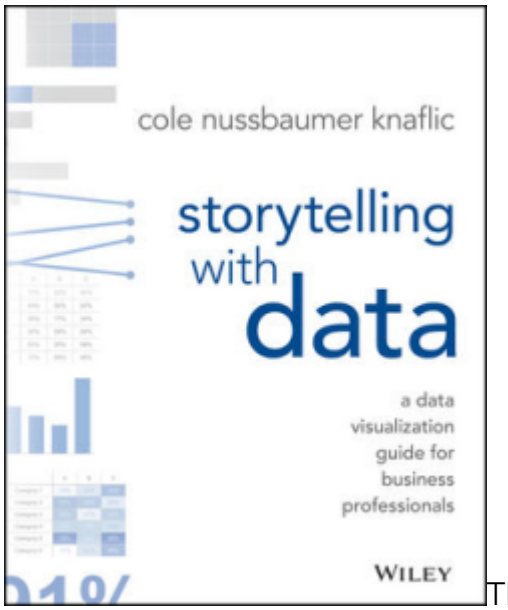
The book introduces inferential approximation algorithms that are useful in generating fast answers from questions asked of big data sets. Although the book requires no prerequisite knowledge of

pattern recognition or machine learning, it does specify that readers should be skillful in calculus and the basics of probability and linear algebra.

Engineers and statisticians have praised the book for its readability and comprehensiveness, although critics have voiced frustration with its non-intuitive math-heavy approach.

Storytelling with Data: A Data Visualization Guide for Business Professionals

Author: Cole Nussbaumer Knaflic



The final part of the data science journey is to present insights learned from the lens of artificial intelligence. Storytelling with Data allows data scientists, machine learning experts and IT professionals to present their ML findings as simple and intuitive data visualizations

Specifically, the book teaches the importance of context and simplicity of data visualization, determining appropriate techniques for various statistical techniques and audiences, eliminating clutter and unnecessary complex information as well as design concepts to present a story or prove a point.

Complementary to the book is another hands-on practice guide by the same author: Storytelling with Data: Let's Practice!.

Related reading

- BMC Machine Learning & Big Data Blog
- Top Machine Learning Algorithms & How To Get Started
- Bias & Variance in Machine Learning: Concepts & Tutorials
- Data Architecture Explained: Components, Standards & Changing Architectures
- Anomaly Detection with Machine Learning: An Introduction
- Data Science Certifications: An Introduction