# The Data Science Handbook

The Data Science HandbookJohn Wiley & Sons

The statistics profession is at a unique point in history. The need for valid statistical tools is greater than ever; data sets are massive, often measuring hundreds of thousands of measurements for a single subject. The field is ready to move towards clear objective benchmarks under which tools can be evaluated. Targeted learning allows (1) the full generalization and utilization of cross-validation as an estimator selection tool so that the subjective choices made by humans are now made by the machine, and (2) targeting the fitting of the probability distribution of the data toward the target parameter representing the scientific question of interest. This book is aimed at both statisticians and applied researchers interested in causal inference and general effect estimation for observational and experimental data. Part I is an accessible introduction to super learning and the targeted maximum likelihood estimator, including related concepts necessary to understand and apply these methods. Parts II-IX handle complex data structures and topics applied researchers will immediately recognize from their own research, including time-to-event outcomes, direct and indirect effects, positivity violations, case-control studies, censored data, longitudinal data, and genomic studies.

Principles and Methods for Data Science, Volume 43 in the Handbook of Statistics series, highlights new advances in the field, with this updated volume presenting interesting and timely topics, including Competing risks, aims and methods, Data analysis and mining of microbial community dynamics, Support Vector Machines, a robust prediction method with applications in bioinformatics, Bayesian Model Selection for Data with High Dimension, High dimensional statistical inference: theoretical development to data analytics, Big data challenges in genomics, Analysis of microarray gene expression data using information theory and stochastic algorithm, Hybrid Models, Markov Chain Monte Carlo Methods: Theory and Practice, and more. Provides the authority and expertise of leading contributors from an international board of authors Presents the latest release in the Handbook of Statistics series Updated release includes the latest information on Principles and Methods for Data Science

A comprehensive overview of data science covering the analytics, programming, and business skills necessary to master the discipline Finding a good data scientist has been likened to hunting for a unicorn: the required combination of technical skills is simply very hard to find in one person. In addition, good data science is not just rote application of trainable skill sets; it requires the ability to think flexibly about all these areas and understand the connections between them. This book provides a crash course in data science, combining all the necessary skills into a unified discipline. Unlike many analytics books, computer science and software engineering are given extensive coverage since they play such a central role in the daily work of a data scientist. The author also describes classic machine learning algorithms, from their mathematical foundations to real-world applications. Visualization tools are reviewed, and their central importance in data science is highlighted. Classical statistics is addressed to help readers think critically about the interpretation of data and its common pitfalls. The clear communication of technical results, which is perhaps the most undertrained of data science skills, is given its own chapter, and all topics are explained in the context of solving real-world data problems. The book also features: • Extensive sample code and tutorials using Python™ along with its technical libraries • Core technologies of “Big Data,” including their strengths and limitations and how they can be used to solve real-world problems • Coverage of the practical realities of the tools, keeping theory to a minimum; however, when theory is presented,

it is done in an intuitive way to encourage critical thinking and creativity • A wide variety of case studies from industry • Practical advice on the realities of being a data scientist today, including the overall workflow, where time is spent, the types of datasets worked on, and the skill sets needed The Data Science Handbook is an ideal resource for data analysis methodology and big data software tools. The book is appropriate for people who want to practice data science, but lack the required skill sets. This includes software professionals who need to better understand analytics and statisticians who need to understand software. Modern data science is a unified discipline, and it is presented as such. This book is also an appropriate reference for researchers and entry-level graduate students who need to learn real-world analytics and expand their skill set. FIELD CADY is the data scientist at the Allen Institute for Artificial Intelligence, where he develops tools that use machine learning to mine scientific literature. He has also worked at Google and several Big Data startups. He has a BS in physics and math from Stanford University, and an MS in computer science from Carnegie Mellon.

Data science is expanding across industries at a rapid pace, and the companies first to adopt best practices will gain a significant advantage. To reap the benefits, decision makers need to have a confident understanding of data science and its application in their organization. It is easy for novices to the subject to feel paralyzed by intimidating buzzwords, but what many don’t realize is that data science is in fact quite multidisciplinary—useful in the hands of business analysts, communications strategists, designers, and more. With the second edition of The Decision Maker’s Handbook to Data Science, you will learn how to think like a veteran data scientist and approach solutions to business problems in an entirely new way. Author Stylianos Kampakis provides you with the expertise and tools required to develop a solid data strategy that is continuously effective. Ethics and legal issues surrounding data collection and algorithmic bias are some common pitfalls that Kampakis helps you avoid, while guiding you on the path to build a thriving data science culture at your organization. This updated and revised second edition, includes plenty of case studies, tools for project assessment, and expanded content for hiring and managing data scientists Data science is a language that everyone at a modern company should understand across departments. Friction in communication arises most often when management does not connect with what a data scientist is doing or how impactful data collection and storage can be for their organization. The Decision Maker’s Handbook to Data Science bridges this gap and readies you for both the present and future of your workplace in this engaging, comprehensive guide. What You Will Learn Understand how data science can be used within your business. Recognize the differences between AI, machine learning, and statistics. Become skilled at thinking like a data scientist, without being one. Discover how to hire and manage data scientists. Comprehend how to build the right environment in order to make your organization data-driven. Who This Book Is For Startup founders, product managers, higher level managers, and any other non-technical decision makers who are thinking to implement data science in their organization and hire data scientists. A secondary audience includes people looking for a soft introduction into the subject of data science.

"This book describes the process of analyzing data. The authors have extensive experience both managing data analysts and conducting their own data analyses, and this book is a distillation of their experience in a format that is applicable to both practitioners and managers in data science."--Leanpub.com.

Written by experts that include originators of some key ideas, chapters in the Handbook of Multiple Testing cover multiple comparison problems big and small, with guidance toward error rate control and insights on how principles developed earlier can be applied to current and emerging problems. Some highlights of the coverages are as follows. Error rate control is useful for controlling the incorrect decision rate. Chapter 1 introduces Tukey's original multiple comparison error rates and point to how they have been applied and adapted to modern

multiple comparison problems as discussed in the later chapters. Principles endure. While the closed testing principle is more familiar, Chapter 4 shows the partitioning principle can derive confidence sets for multiple tests, which may become important as the profession goes beyond making decisions based on p-values. Multiple comparisons of treatment efficacy often involve multiple doses and endpoints. Chapter 12 on multiple endpoints explains how different choices of endpoint types lead to different multiplicity adjustment strategies, while Chapter 11 on the MCP-Mod approach is particularly useful for dose-finding. To assess efficacy in clinical trials with multiple doses and multiple endpoints, the reader can see the traditional approach in Chapter 2, the Graphical approach in Chapter 5, and the multivariate approach in Chapter 3. Personalized/precision medicine based on targeted therapies, already a reality, naturally leads to analysis of efficacy in subgroups. Chapter 13 draws attention to subtle logical issues in inferences on subgroups and their mixtures, with a principled solution that resolves these issues. This chapter has implication toward meeting the ICHE9R1 Estimands requirement. Besides the mere multiple testing methodology itself, the handbook also covers related topics like the statistical task of model selection in Chapter 7 or the estimation of the proportion of true null hypotheses (or, in other words, the signal prevalence) in Chapter 8. It also contains decision-theoretic considerations regarding the admissibility of multiple tests in Chapter 6. The issue of selected inference is addressed in Chapter 9. Comparison of responses can involve millions of voxels in medical imaging or SNPs in genome-wide association studies (GWAS). Chapter 14 and Chapter 15 provide state of the art methods for large scale simultaneous inference in these settings.

This book provides an introduction to the mathematical and algorithmic foundations of data science, including machine learning, high-dimensional geometry, and analysis of large networks. Topics include the counterintuitive nature of data in high dimensions, important linear algebraic techniques such as singular value decomposition, the theory of random walks and Markov chains, the fundamentals of and important algorithms for machine learning, algorithms and analysis for clustering, probabilistic models for large networks, representation learning including topic modelling and non-negative matrix factorization, wavelets and compressed sensing. Important probabilistic techniques are developed including the law of large numbers, tail inequalities, analysis of random projections, generalization guarantees in machine learning, and moment methods for analysis of phase transitions in large random graphs. Additionally, important structural and complexity measures are discussed such as matrix norms and VC-dimension. This book is suitable for both undergraduate and graduate courses in the design and analysis of algorithms for data.

Build machine and deep learning systems with the newly released TensorFlow 2 and Keras for the lab, production, and mobile devices Key Features Introduces and then uses TensorFlow 2 and Keras right from the start Teaches key machine and deep learning techniques Understand the fundamentals of deep learning and machine learning through clear explanations and extensive code samples Book Description Deep Learning with TensorFlow 2 and Keras, Second Edition teaches neural networks and deep learning techniques alongside TensorFlow (TF) and Keras. You'll learn how to write deep learning applications in the most powerful, popular, and scalable machine learning stack available. TensorFlow is the machine learning library of choice for professional applications, while Keras offers a simple and powerful Python API for accessing TensorFlow. TensorFlow 2 provides full Keras integration, making advanced machine learning easier and more convenient than ever before. This book also introduces neural networks with TensorFlow, runs

through the main applications (regression, ConvNets (CNNs), GANs, RNNs, NLP), covers two working example apps, and then dives into TF in production, TF mobile, and using TensorFlow with AutoML. What you will learn Build machine learning and deep learning systems with TensorFlow 2 and the Keras API Use Regression analysis, the most popular approach to machine learning Understand ConvNets (convolutional neural networks) and how they are essential for deep learning systems such as image classifiers Use GANs (generative adversarial networks) to create new data that fits with existing patterns Discover RNNs (recurrent neural networks) that can process sequences of input intelligently, using one part of a sequence to correctly interpret another Apply deep learning to natural human language and interpret natural language texts to produce an appropriate response Train your models on the cloud and put TF to work in real environments Explore how Google tools can automate simple ML workflows without the need for complex modeling Who this book is for This book is for Python developers and data scientists who want to build machine learning and deep learning systems with TensorFlow. Whether or not you have done machine learning before, this book gives you the theory and practice required to use Keras, TensorFlow 2, and AutoML to build machine learning systems.

Data Science gets thrown around in the press like it'smagic. Major retailers are predicting everything from when theircustomers are pregnant to when they want a new pair of ChuckTaylors. It's a brave new world where seemingly meaningless datacan be transformed into valuable insight to drive smart businessdecisions. But how does one exactly do data science? Do you have to hireone of these priests of the dark arts, the "data scientist," toextract this gold from your data? Nope. Data science is little more than using straight-forward steps toprocess raw data into actionable insight. And in DataSmart, author and data scientist John Foreman will show you howthat's done within the familiar environment of aspreadsheet. Why a spreadsheet? It's comfortable! You get to look at the dataevery step of the way, building confidence as you learn the tricksof the trade. Plus, spreadsheets are a vendor-neutral place tolearn data science without the hype. But don't let the Excel sheets fool you. This is a book forthose serious about learning the analytic techniques, the math andthe magic, behind big data. Each chapter will cover a different technique in aspreadsheet so you can follow along: Mathematical optimization, including non-linear programming andgenetic algorithms Clustering via k-means, spherical k-means, and graphmodularity Data mining in graphs, such as outlier detection Supervised AI through logistic regression, ensemble models, andbag-of-words models Forecasting, seasonal adjustments, and prediction intervalsthrough monte carlo simulation Moving from spreadsheets into the R programming language You get your hands dirty as you work alongside John through eachtechnique. But never fear, the topics are readily applicable andthe author laces humor throughout. You'll even learnwhat a dead squirrel has to do with optimization modeling, whichyou no doubt are dying to know.

Analyzing data sets has continued to be an invaluable application for numerous industries. By combining different algorithms, technologies, and systems used to extract information from data and solve complex problems, various sectors have reached new heights and have changed our world for the better. The Handbook of Research on Engineering, Business, and Healthcare Applications of Data Science and Analytics is a collection of innovative research on the methods and applications of data analytics. While highlighting topics including artificial intelligence, data security, and information systems, this book is ideally designed for researchers, data analysts, data scientists, healthcare administrators, executives, managers, engineers, IT consultants, academicians, and students interested in the potential of data application technologies.

If you're tired of licensing third-party software for data analysis, Python Data Science will help you do it for yourself! Recently, more and more companies are learning that they need to make DATA-DRIVEN decisions. And with big data and data science on the rise, we now have more data than we know what to do with. In fact, without a doubt, you have already experienced data science in one way or another. Obviously, you are interacting with data science products every time you search for information on the web by using search engines such as Google, or asking for directions with your mobile phone. Data science is the science and technology focused on collecting raw data and processing it in an effective manner. It is the combination of concepts and methods that make it possible to give meaning and understandability to huge volumes of data. Data science has been the force behind resolving some of our most common daily tasks for several years. In nearly all of our daily work, we directly or indirectly work on storing and exchanging data. With the rapid development of technology, the need to store data effectively is also increasing. That's why it needs to be handled properly. Basically, data science unearths the hidden insights of raw-data and uses them for productive output. Python is often used in data science today because it is a mature programming language that has excellent properties for newbie programmers. Some of the most remarkable of these properties are its easy to read code, suppression of non-mandatory delimiters, dynamic typing, and dynamic memory usage. Python is an interpreted language, and it can be executed in the Python console without any need to compile to machine language. "Python Data Science" teaches a complete course of data science, including key topics like data integration, data mining, python etc. We will explore NumPy for numerical data, Pandas for data analysis, IPython, Scikit-learn and Tensorflow for machine learning and business. Each of the chapters in this book is devoted to one of the most interesting aspects of data analysis and processing. The following are some of the major topics covered in Python Data Science: Understanding Data Science Getting Started with Python for Data Scientists Descriptive statistics Data Analysis and Libraries NumPy Arrays and Vectorized Computation Data Analysis with Pandas Data Visualization Data Mining Classifying with Scikit-learn

Estimators Giving Computers the Ability to Learn from Data Training Machine Learning Algorithms The Python ecosystem for data science discussed within Python Data Science includes SciPy, NumPy, Matplotlib, Pandas, and Scikit-learn, which provides all of the data science algorithms. Data processing and analysis is one of the hottest areas of IT, where developers who can handle projects of any level, from social networks to trained systems, are constantly required. We hope this book will be the starting point for your journey into the fascinating world of Data Science. To get started on your Python adventure, just scroll back up and click the 'Buy' button.

"This book introduces you to R, RStudio, and the tidyverse, a collection of R packages designed to work together to make data science fast, fluent, and fun. Suitable for readers with no previous programming experience"--

Summary Dask is a native parallel analytics tool designed to integrate seamlessly with the libraries you're already using, including Pandas, NumPy, and Scikit-Learn. With Dask you can crunch and work with huge datasets, using the tools you already have. And Data Science with Python and Dask is your guide to using Dask for your data projects without changing the way you work! Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. You'll find registration instructions inside the print book. About the Technology An efficient data pipeline means everything for the success of a data science project. Dask is a flexible library for parallel computing in Python that makes it easy to build intuitive workflows for ingesting and analyzing large, distributed datasets. Dask provides dynamic task scheduling and parallel collections that extend the functionality of NumPy, Pandas, and Scikit-learn, enabling users to scale their code from a single laptop to a cluster of hundreds of machines with ease. About the Book Data Science with Python and Dask teaches you to build scalable projects that can handle massive datasets. After meeting the Dask framework, you'll analyze data in the NYC Parking Ticket database and use DataFrames to streamline your process. Then, you'll create machine learning models using Dask-ML, build interactive visualizations, and build clusters using AWS and Docker. What's inside Working with large, structured and unstructured datasets Visualization with Seaborn and Datashader Implementing your own algorithms Building distributed apps with Dask Distributed Packaging and deploying Dask apps About the Reader For data scientists and developers with experience using Python and the PyData stack. About the Author Jesse Daniel is an experienced Python developer. He taught Python for Data Science at the University of Denver and leads a team of data scientists at a Denver-based media technology company. Table of Contents PART 1 - The Building Blocks of scalable computing Why scalable computing matters Introducing Dask PART 2 - Working with Structured Data using Dask DataFrames Introducing Dask DataFrames Loading data into DataFrames Cleaning and transforming DataFrames Summarizing and analyzing DataFrames Visualizing DataFrames with Seaborn Visualizing location data with Datashader PART 3 - Extending and deploying Dask Working with Bags and

Arrays Machine learning with Dask-ML Scaling and deploying Dask

A concise introduction to the emerging field of data science, explaining its evolution, relation to machine learning, current uses, data infrastructure issues, and ethical challenges. The goal of data science is to improve decision making through the analysis of data. Today data science determines the ads we see online, the books and movies that are recommended to us online, which emails are filtered into our spam folders, and even how much we pay for health insurance. This volume in the MIT Press Essential Knowledge series offers a concise introduction to the emerging field of data science, explaining its evolution, current uses, data infrastructure issues, and ethical challenges. It has never been easier for organizations to gather, store, and process data. Use of data science is driven by the rise of big data and social media, the development of high-performance computing, and the emergence of such powerful methods for data analysis and modeling as deep learning. Data science encompasses a set of principles, problem definitions, algorithms, and processes for extracting non-obvious and useful patterns from large datasets. It is closely related to the fields of data mining and machine learning, but broader in scope. This book offers a brief history of the field, introduces fundamental data concepts, and describes the stages in a data science project. It considers data infrastructure and the challenges posed by integrating data from multiple sources, introduces the basics of machine learning, and discusses how to link machine learning expertise with real-world problems. The book also reviews ethical and legal issues, developments in data regulation, and computational approaches to preserving privacy. Finally, it considers the future impact of data science and offers principles for success in data science projects.

Learn the techniques and math you need to start making sense of your data About This Book Enhance your knowledge of coding with data science theory for practical insight into data science and analysis More than just a math class, learn how to perform real-world data science tasks with R and Python Create actionable insights and transform raw data into tangible value Who This Book Is For You should be fairly well acquainted with basic algebra and should feel comfortable reading snippets of R/Python as well as pseudo code. You should have the urge to learn and apply the techniques put forth in this book on either your own data sets or those provided to you. If you have the basic math skills but want to apply them in data science or you have good programming skills but lack math, then this book is for you. What You Will Learn Get to know the five most important steps of data science Use your data intelligently and learn how to handle it with care Bridge the gap between mathematics and programming Learn about probability, calculus, and how to use statistical models to control and clean your data and drive actionable results Build and evaluate baseline machine learning models Explore the most effective metrics to determine the success of your machine learning models Create data visualizations that communicate actionable insights Read and apply machine learning concepts to your problems

and make actual predictions In Detail Need to turn your skills at programming into effective data science skills? Principles of Data Science is created to help you join the dots between mathematics, programming, and business analysis. With this book, you'll feel confident about asking—and answering—complex and sophisticated questions of your data to move from abstract and raw statistics to actionable ideas. With a unique approach that bridges the gap between mathematics and computer science, this books takes you through the entire data science pipeline. Beginning with cleaning and preparing data, and effective data mining strategies and techniques, you'll move on to build a comprehensive picture of how every piece of the data science puzzle fits together. Learn the fundamentals of computational mathematics and statistics, as well as some pseudocode being used today by data scientists and analysts. You'll get to grips with machine learning, discover the statistical models that help you take control and navigate even the densest datasets, and find out how to create powerful visualizations that communicate what your data means. Style and approach This is an easy-to-understand and accessible tutorial. It is a step-by-step guide with use cases, examples, and illustrations to get you well-versed with the concepts of data science. Along with explaining the fundamentals, the book will also introduce you to slightly advanced concepts later on and will help you implement these techniques in the real world.

"This book is a great way to both start learning data science through the promising Julia language and to become an efficient data scientist."- Professor Charles Bouveyron, INRIA Chair in Data Science, Université Côte d'Azur, Nice, France Julia, an open-source programming language, was created to be as easy to use as languages such as R and Python while also as fast as C and Fortran. An accessible, intuitive, and highly efficient base language with speed that exceeds R and Python, makes Julia a formidable language for data science. Using well known data science methods that will motivate the reader, Data Science with Julia will get readers up to speed on key features of the Julia language and illustrate its facilities for data science and machine learning work. Features: Covers the core components of Julia as well as packages relevant to the input, manipulation and representation of data. Discusses several important topics in data science including supervised and unsupervised learning. Reviews data visualization using the Gadfly package, which was designed to emulate the very popular ggplot2 package in R. Readers will learn how to make many common plots and how to visualize model results. Presents how to optimize Julia code for performance. Will be an ideal source for people who already know R and want to learn how to use Julia (though no previous knowledge of R or any other programming language is required). The advantages of Julia for data science cannot be understated. Besides speed and ease of use, there are already over 1,900 packages available and Julia can interface (either directly or through packages) with libraries written in R, Python, Matlab, C, C++ or Fortran. The book is for senior undergraduates, beginning graduate students, or practicing data scientists who want to learn how to use Julia for data science. "This book is a great way to

both start learning data science through the promising Julia language and to become an efficient data scientist." Professor Charles Bouveyron INRIA Chair in Data Science Université Côte d'Azur, Nice, France

Data Science and Big Data Analytics is about harnessing the power of data for new insights. The book covers the breadth of activities and methods and tools that Data Scientists use. The content focuses on concepts, principles and practical applications that are applicable to any industry and technology environment, and the learning is supported and explained with examples that you can replicate using open-source software. This book will help you: Become a contributor on a data science team Deploy a structured lifecycle approach to data analytics problems Apply appropriate analytic techniques and tools to analyzing big data Learn how to tell a compelling story with data to drive business action Prepare for EMC Proven Professional Data Science Certification Corresponding data sets are available from the book's page at Wiley which you can find on the Wiley site by searching for the ISBN 9781118876138. Get started discovering, analyzing, visualizing, and presenting data in a meaningful way today!

Tap into the power of data science with this comprehensive resource for non-technical professionals Data Science: The Executive Summary – A Technical Book for Non-Technical Professionals is a comprehensive resource for people in non-engineer roles who want to fully understand data science and analytics concepts. Accomplished data scientist and author Field Cady describes both the "business side" of data science, including what problems it solves and how it fits into an organization, and the technical side, including analytical techniques and key technologies. Data Science: The Executive Summary covers topics like: Assessing whether your organization needs data scientists, and what to look for when hiring them When Big Data is the best approach to use for a project, and when it actually ties analysts' hands Cutting edge Artificial Intelligence, as well as classical approaches that work better for many problems How many techniques rely on dubious mathematical idealizations, and when you can work around them Perfect for executives who make critical decisions based on data science and analytics, as well as mangers who hire and assess the work of data scientists, Data Science: The Executive Summary also belongs on the bookshelves of salespeople and marketers who need to explain what a data analytics product does. Finally, data scientists themselves will improve their technical work with insights into the goals and constraints of the business situation.

Praise for the Second Edition: "The authors present an intuitive and easy-to-read book. ... accompanied by many examples, proposed exercises, good references, and comprehensive appendices that initiate the reader unfamiliar with MATLAB." —Adolfo Alvarez Pinto, International Statistical Review "Practitioners of EDA who use MATLAB will want a copy of this book. ... The authors have done a great service by bringing together so many EDA routines, but their main accomplishment in this dynamic text is providing the understanding and tools to do EDA. —David A Huckaby, MAA

Reviews Exploratory Data Analysis (EDA) is an important part of the data analysis process. The methods presented in this text are ones that should be in the toolkit of every data scientist. As computational sophistication has increased and data sets have grown in size and complexity, EDA has become an even more important process for visualizing and summarizing data before making assumptions to generate hypotheses and models. Exploratory Data Analysis with MATLAB, Third Edition presents EDA methods from a computational perspective and uses numerous examples and applications to show how the methods are used in practice. The authors use MATLAB code, pseudo-code, and algorithm descriptions to illustrate the concepts. The MATLAB code for examples, data sets, and the EDA Toolbox are available for download on the book's website. New to the Third Edition Random projections and estimating local intrinsic dimensionality Deep learning autoencoders and stochastic neighbor embedding Minimum spanning tree and additional cluster validity indices Kernel density estimation Plots for visualizing data distributions, such as beanplots and violin plots A chapter on visualizing categorical data

JavaScript is the native language of the Internet. Originally created to make web pages more dynamic, it is now used for software projects of all kinds, including scientific visualization and data services. However, most data scientists have little or no experience with JavaScript, and most introductions to the language are written for people who want to build shopping carts rather than share maps of coral reefs. This book will introduce you to JavaScript's power and idiosyncrasies and guide you through the key features of the language and its tools and libraries. The book places equal focus on client- and server-side programming, and shows readers how to create interactive web content, build and test data services, and visualize data in the browser. Topics include: The core features of modern JavaScript Creating templated web pages Making those pages interactive using React Data visualization using Vega-Lite Using Data-Forge to wrangle tabular data Building a data service with Express Unit testing with Mocha All of the material is covered by the Creative Commons Attribution-Noncommercial 4.0 International license (CC-BY-NC-4.0) and is included in the book's companion website at http://js4ds.org . Maya Gans is a freelance data scientist and front-end developer by way of quantitative biology. Toby Hodges is a bioinformatician turned community coordinator who works at the European Molecular Biology Laboratory. Greg Wilson co-founded Software Carpentry, and is now part of the education team at RStudio

Written by renowned data science experts Foster Provost and Tom Fawcett, Data Science for Business introduces the fundamental principles of data science, and walks you through the "data-analytic thinking" necessary for extracting useful knowledge and business value from the data you collect. This guide also helps you understand the many data-mining techniques in use today. Based on an MBA course Provost has taught at New York University over the past ten years,

Data Science for Business provides examples of real-world business problems to illustrate these principles. You'll not only learn how to improve communication between business stakeholders and data scientists, but also how participate intelligently in your company's data science projects. You'll also discover how to think data-analytically, and fully appreciate how data science methods can support business decision-making. Understand how data science fits in your organization—and how you can use it for competitive advantage Treat data as a business asset that requires careful investment if you're to gain real value Approach business problems data-analytically, using the data-mining process to gather good data in the most appropriate way Learn general concepts for actually extracting knowledge from data Apply data science principles when interviewing data science job candidates

Handbook of Data Science Approaches for Biomedical Engineering covers the research issues and concepts of biomedical engineering progress and the ways they are aligning with the latest technologies in IoT and big data. In addition, the book includes various real-time/offline medical applications that directly or indirectly rely on medical and information technology. Case studies in the field of medical science, i.e., biomedical engineering, computer science, information security, and interdisciplinary tools, along with modern tools and the technologies used are also included to enhance understanding. Today, the role of Big Data and IoT proves that ninety percent of data currently available has been generated in the last couple of years, with rapid increases happening every day. The reason for this growth is increasing in communication through electronic devices, sensors, web logs, global positioning system (GPS) data, mobile data, IoT, etc. Provides in-depth information about Biomedical Engineering with Big Data and Internet of Things Includes technical approaches for solving real-time healthcare problems and practical solutions through case studies in Big Data and Internet of Things Discusses big data applications for healthcare management, such as predictive analytics and forecasting, big data integration for medical data, algorithms and techniques to speed up the analysis of big medical data, and more

The Data Science Handbook is a curated collection of 25 candid, honest and insightful interviews conducted with some of the world's top data scientists.In this book, you'll hear how the co-creator of the term 'data scientist' thinks about career and personal success. You'll hear from a young woman who created her own data scientist curriculum, subsequently landing her a role in the field. Readers of this book will be left with war stories, wisdom and

The contemporary world lives on the data produced at an unprecedented speed through social networks and the internet of things (IoT). Data has been called the new global currency, and its rise is transforming entire industries, providing a wealth of opportunities. Applied data science research is necessary to derive useful information from big data for the effective and efficient utilization to solve real-world problems. A broad analytical set allied with strong business logic is fundamental in today's

corporations. Organizations work to obtain competitive advantage by analyzing the data produced within and outside their organizational limits to support their decision-making processes. This book aims to provide an overview of the concepts, tools, and techniques behind the fields of data science and artificial intelligence (AI) applied to business and industries. The Handbook of Research on Applied Data Science and Artificial Intelligence in Business and Industry discusses all stages of data science to AI and their application to real problems across industries—from science and engineering to academia and commerce. This book brings together practice and science to build successful data solutions, showing how to uncover hidden patterns and leverage them to improve all aspects of business performance by making sense of data from both web and offline environments. Covering topics including applied AI, consumer behavior analytics, and machine learning, this text is essential for data scientists, IT specialists, managers, executives, software and computer engineers, researchers, practitioners, academicians, and students.
This thoroughly revised guide demonstrates how the flexibility of the command line can help you become a more efficient and productive data scientist. You'll learn how to combine small yet powerful command-line tools to quickly obtain, scrub, explore, and model your data. To get you started, author Jeroen Janssens provides a Docker image packed with over 80 tools--useful whether you work with Windows, macOS, or Linux. You'll quickly discover why the command line is an agile, scalable, and extensible technology. Even if you're comfortable processing data with Python or R, you'll learn how to greatly improve your data science workflow by leveraging the command line's power. This book is ideal for data scientists, analysts, and engineers; software and machine learning engineers; and system administrators. Obtain data from websites, APIs, databases, and spreadsheets Perform scrub operations on text, CSV, HTM, XML, and JSON files Explore data, compute descriptive statistics, and create visualizations Manage your data science workflow Create reusable command-line tools from one-liners and existing Python or R code Parallelize and distribute data-intensive pipelines Model data with dimensionality reduction, clustering, regression, and classification algorithms
Now that people are aware that data can make the difference in an election or a business model, data science as an occupation is gaining ground. But how can you get started working in a wide-ranging, interdisciplinary field that’s so clouded in hype? This insightful book, based on Columbia University’s Introduction to Data Science class, tells you what you need to know. In many of these chapter-long lectures, data scientists from companies such as Google, Microsoft, and eBay share new algorithms, methods, and models by presenting case studies and the code they use. If you’re familiar with linear algebra, probability, and statistics, and have programming experience, this book is an ideal introduction to data science. Topics include: Statistical inference, exploratory data analysis, and the data science process Algorithms Spam filters, Naive Bayes, and data wrangling Logistic regression Financial modeling Recommendation engines and causality Data visualization Social networks and data journalism Data engineering, MapReduce, Pregel, and Hadoop Doing Data Science is collaboration between course instructor Rachel Schutt, Senior VP of Data Science at News Corp, and data science consultant Cathy O’Neil, a senior data scientist at Johnson Research Labs, who attended and blogged about the course.

This engaging and clearly written textbook/reference provides a must-have introduction to the rapidly emerging interdisciplinary field of data science. It focuses on the principles fundamental to becoming a good data scientist and the key skills needed to build systems for collecting, analyzing, and interpreting data. The Data Science Design Manual is a source of practical insights that highlights what really matters in analyzing data, and provides an intuitive understanding of how these core concepts can be used. The book does not emphasize any particular programming language or suite of data-analysis tools, focusing instead on high-level discussion of important design principles. This easy-to-read text ideally serves the needs of undergraduate and early graduate students embarking on an “Introduction to Data Science” course. It reveals how this discipline sits at the intersection of statistics, computer science, and machine learning, with a distinct heft and character of its own. Practitioners in these and related fields will find this book perfect for self-study as well. Additional learning tools: Contains “War Stories,” offering perspectives on how data science applies in the real world Includes “Homework Problems,” providing a wide range of exercises and projects for self-study Provides a complete set of lecture slides and online video lectures at www.data-manual.com Provides “Take-Home Lessons,” emphasizing the big-picture concepts to learn from each chapter Recommends exciting “Kaggle Challenges” from the online platform Kaggle Highlights “False Starts,” revealing the subtle reasons why certain approaches fail Offers examples taken from the data science television show “The Quant Shop” (www.quant-shop.com)

Introduction to Data Science: Data Analysis and Prediction Algorithms with R introduces concepts and skills that can help you tackle real-world data analysis challenges. It covers concepts from probability, statistical inference, linear regression, and machine learning. It also helps you develop skills such as R programming, data wrangling, data visualization, predictive algorithm building, file organization with UNIX/Linux shell, version control with Git and GitHub, and reproducible document preparation. This book is a textbook for a first course in data science. No previous knowledge of R is necessary, although some experience with programming may be helpful. The book is divided into six parts: R, data visualization, statistics with R, data wrangling, machine learning, and productivity tools. Each part has several chapters meant to be presented as one lecture. The author uses motivating case studies that realistically mimic a data scientist’s experience. He starts by asking specific questions and answers these through data analysis so concepts are learned as a means to answering the questions. Examples of the case studies included are: US murder rates by state, self-reported student heights, trends in world health and economics, the impact of vaccines on infectious disease rates, the financial crisis of 2007-2008, election forecasting, building a baseball team, image processing of hand-written digits, and movie recommendation systems. The statistical concepts used to answer the case study questions are only briefly introduced, so complementing with a probability and statistics textbook is highly recommended for in-depth understanding of these concepts. If you read and understand the chapters and complete the exercises, you will be prepared to learn the more advanced concepts and skills needed to become an expert.

This beginning graduate textbook teaches data science and machine learning methods for modeling, prediction, and control of complex systems.

Get complete instructions for manipulating, processing, cleaning, and crunching datasets in Python. Updated for Python 3.6, the second edition of this hands-on guide is packed with practical case studies that show you how to solve a broad set of data analysis problems effectively. You'll learn the latest versions of pandas, NumPy, IPython, and Jupyter in the process. Written by Wes McKinney, the creator of the Python pandas project, this book is a practical, modern introduction to data science tools in Python. It's ideal for analysts new to Python and for Python programmers new to data science and scientific computing. Data files and related material are available on GitHub. Use the IPython shell and Jupyter notebook for exploratory computing Learn basic and advanced features in NumPy (Numerical Python) Get started with data analysis tools in the pandas library Use flexible tools to load, clean, transform, merge, and reshape data Create informative visualizations with matplotlib Apply the pandas groupby facility to slice, dice, and summarize datasets Analyze and manipulate regular and irregular time series data Learn how to solve real-world data analysis problems with thorough, detailed examples

Handbook of Big Data provides a state-of-the-art overview of the analysis of large-scale datasets. Featuring contributions from well-known experts in statistics and computer science, this handbook presents a carefully curated collection of techniques from both industry and academia. Thus, the text instills a working understanding of key statistical

"This edited book discusses data analytics and complex communication networks and recommends new methodologies, system architectures, and other solutions to prevail over the current limitations faced by the field"--

Knowing how to work with data to extract insights generates significant value. This book will help you to develop data analysis skills using a hands-on approach and real-world data. You'll get up to speed with pandas 1.x in no time and build some software engineering skills in the process, vastly expanding your data science toolbox.

The use of data in society has seen an exponential growth in recent years. Data science, the field of research concerned with understanding and analyzing data, aims to find ways to operationalize data so that it can be beneficially used in society, for example in health applications, urban governance or smart household devices. The legal questions that accompany the rise of new, data-driven technologies however are underexplored. This book is the first volume that seeks to map the legal implications of the emergence of data science. It discusses the possibilities and limitations imposed by the current legal framework, considers whether regulation is needed to respond to problems raised by data science, and which ethical problems occur in relation to the use of data. It also considers the emergence of Data Science and Law as a new legal discipline.

Data science libraries, frameworks, modules, and toolkits are great for doing data science, but they're also a good way to dive into the discipline without actually understanding data science. In this book, you'll learn how many of the most fundamental data science tools and algorithms work by implementing them from scratch. If you have an aptitude for mathematics and some programming skills, author Joel Grus will help you get comfortable with the math and statistics at the core of data science, and with hacking skills you need to get started as a data scientist. Today's messy glut of data holds answers to questions no one's even thought to ask. This book provides you with the know-how to dig those answers out. Get a crash course in Python Learn the

basics of linear algebra, statistics, and probability—and understand how and when they're used in data science Collect, explore, clean, munge, and manipulate data Dive into the fundamentals of machine learning Implement models such as k-nearest Neighbors, Naive Bayes, linear and logistic regression, decision trees, neural networks, and clustering Explore recommender systems, natural language processing, network analysis, MapReduce, and databases

The ultimate guide for anyone wondering how President Joe Biden will respond to the COVID-19 pandemic—all his plans, goals, and executive orders in response to the coronavirus crisis. Shortly after being inaugurated as the 46th President of the United States, Joe Biden and his administration released this 200 page guide detailing his plans to respond to the coronavirus pandemic. The National Strategy for the COVID-19 Response and Pandemic Preparedness breaks down seven crucial goals of President Joe Biden's administration with regards to the coronavirus pandemic: 1. Restore trust with the American people. 2. Mount a safe, effective, and comprehensive vaccination campaign. 3. Mitigate spread through expanding masking, testing, data, treatments, health care workforce, and clear public health standards. 4. Immediately expand emergency relief and exercise the Defense Production Act. 5. Safely reopen schools, businesses, and travel while protecting workers. 6. Protect those most at risk and advance equity, including across racial, ethnic and rural/urban lines. 7. Restore U.S. leadership globally and build better preparedness for future threats. Each of these goals are explained and detailed in the book, with evidence about the current circumstances and how we got here, as well as plans and concrete steps to achieve each goal. Also included is the full text of the many Executive Orders that will be issued by President Biden to achieve each of these goals. The National Strategy for the COVID-19 Response and Pandemic Preparedness is required reading for anyone interested in or concerned about the COVID-19 pandemic and its effects on American society.

If you're a developer looking to supplement your own data tools and services, this concise ebook covers the most useful sources of public data available today. You'll find useful information on APIs that offer broad coverage, tie their data to the outside world, and are either accessible online or feature downloadable bulk data. You'll also find code and helpful links. This guide organizes APIs by the subjects they cover—such as websites, people, or places—so you can quickly locate the best resources for augmenting the data you handle in your own service. Categories include: Website tools such as WHOIS, bit.ly, and Compete Services that use email addresses as search terms, including Github Finding information from just a name, with APIs such as WhitePages Services, such as Klout, for locating people with Facebook and Twitter accounts Search APIs, including BOSS and Wikipedia Geographical data sources, including SimpleGeo and U.S. Census Company information APIs, such as CrunchBase and ZoomInfo APIs that list IP addresses, such as MaxMind Services that list books, films, music, and products

FINALIST: Business Book Awards 2020 - HR & Management Category In a world of work where recruiters are constantly hearing that their role is at risk from AI, robotics and chatbots, it has never been more important to effectively attract and recruit the right people. Leveraging the power of social media and digital sourcing strategies is only part of the solution, and simply posting a job or sending a LinkedIn InMail is no longer enough. The Robot-Proof Recruiter shows you how to use the tools that reveal information

that can be used to grab a potential candidate's attention among the overwhelming volume of material online. Full of expert guidance and practical tips, this book explains what works, what doesn't, and how you can stand out and recruit effectively in a world of technology overload. The Robot-Proof Recruiter will enable you to become the recruiter that candidates trust and the one they want to talk to. It contains essential guidance on overcoming obstacles - including how to recruit without an existing online presence, how to work effectively with hiring managers to improve the candidate experience, and how to use technology to support the candidate's journey from initial outreach, to application, to employee, and through to alumnus. This is an indispensable book for all recruitment professionals and HR practitioners who want to recruit the right people for their organization.

For many researchers, Python is a first-class tool mainly because of its libraries for storing, manipulating, and gaining insight from data. Several resources exist for individual pieces of this data science stack, but only with the Python Data Science Handbook do you get them all—IPython, NumPy, Pandas, Matplotlib, Scikit-Learn, and other related tools. Working scientists and data crunchers familiar with reading and writing Python code will find this comprehensive desk reference ideal for tackling day-to-day issues: manipulating, transforming, and cleaning data; visualizing different types of data; and using data to build statistical or machine learning models. Quite simply, this is the must-have reference for scientific computing in Python. With this handbook, you'll learn how to use: IPython and Jupyter: provide computational environments for data scientists using Python NumPy: includes the ndarray for efficient storage and manipulation of dense data arrays in Python Pandas: features the DataFrame for efficient storage and manipulation of labeled/columnar data in Python Matplotlib: includes capabilities for a flexible range of data visualizations in Python Scikit-Learn: for efficient and clean Python implementations of the most important and established machine learning algorithms

Copyright: b2cf43b5d4f54a49905b8bf06e9fab1e