Data Matching Concepts And Techniques For Record Linkage Entity Resolution And Duplicate Detection Data Centric Systems And Applications 2012 Edition By Christen Peter Published By Springer 2012

Thank you for reading **data matching concepts and techniques for record linkage entity resolution and duplicate detection data centric systems and applications 2012 edition by christen peter published by springer 2012**. As you may know, people have look hundreds times for their chosen readings like this data matching concepts and techniques for record linkage entity resolution and duplicate detection data centric systems and applications 2012 edition by christen peter published by springer 2012, but end up in harmful downloads.
Rather than enjoying a good book with a cup of tea in the afternoon, instead they cope with some malicious virus inside their laptop.

data matching concepts and techniques for record linkage entity resolution and duplicate detection data centric systems and applications 2012 edition by christen peter published by springer 2012 is available in our digital library an online access to it is set as public so you can download it instantly.
Our book servers saves in multiple locations, allowing you to get the most less latency time to download any of our books like this one.
Merely said, the data matching concepts and techniques for record linkage entity resolution and duplicate detection data centric systems and applications 2012 edition by christen peter published by springer 2012 is universally compatible with any devices to read

The Data Warehouse Toolkit - Ralph Kimball 2011-08-08
This old edition was published in 2002. The current and final edition of this book is The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, 3rd Edition which was published in 2013 under ISBN: 9781118530801. The authors begin with fundamental design recommendations and gradually progress step-by-step through increasingly complex scenarios. Clear-cut guidelines for designing dimensional models are illustrated using real-world data warehouse case studies drawn from a variety of business application areas and industries, including: Retail sales and e-commerce Inventory management Procurement Order management Customer relationship management (CRM) Human resources management Accounting Financial services Telecommunications and utilities Education Transportation Health care and insurance By the end of the book, you will have mastered the full range of powerful techniques for designing dimensional databases that are easy to understand and provide fast query response. You will also learn how to create an architected framework that integrates the distributed data warehouse using standardized dimensions and facts.
Mining of Massive Datasets - Jure Leskovec 2014-11-13
Now in its second edition, this book focuses on practical algorithms for mining data from even the largest datasets.
Data Quality - Carlo Batini 2006-09-27
Poor data quality can seriously hinder or damage the efficiency and effectiveness of organizations and businesses. The growing awareness of such repercussions has led to major public initiatives like the "Data Quality Act" in the USA and the "European 2003/98" directive of the European Parliament. Batini and Scannapieco present a comprehensive and systematic introduction to the wide set of issues related to data quality. They start with a detailed description of different data quality dimensions, like accuracy, completeness, and consistency, and their importance in different types of data, like federated data, web data, or time-dependent data, and in different data categories classified according to frequency of change, like stable, long-term, and frequently changing data. The book's extensive description of techniques and methodologies from core data quality research as well as from related fields like data mining, probability theory, statistical data analysis, and machine learning gives an excellent overview of the current state of the art. The presentation is completed by a short description and critical comparison of tools and practical methodologies, which will help readers to resolve their own quality problems. This book is an ideal combination of the soundness of theoretical foundations and the applicability of practical approaches. It is ideally suited for everyone – researchers, students, or professionals – interested in a comprehensive overview of data quality issues. In addition, it will serve as the basis for an introductory course or for self-study on this topic.
Health Data in the Information Age - Institute of Medicine 1994-01-01
Regional health care databases are being established around the country with the goal of providing timely and useful information to policymakers, physicians, and patients. But their emergence is raising important and sometimes controversial questions about the collection, quality, and appropriate use of health care data. Based on experience with databases now in operation and in development, Health Data in the Information Age provides a clear set of guidelines and principles for exploiting the potential benefits of aggregated health dataâ€"without jeopardizing confidentiality. A panel of experts identifies characteristics of emerging health database organizations (HDOs). The committee explores how HDOs can maintain the quality of their data, what policies and practices they should adopt, how they can prepare for linkages with computer-based patient records, and how diverse groups from researchers to health care administrators might use aggregated data. Health Data in the Information Age offers frank analysis and guidelines that will be invaluable to anyone interested in the operation of health care databases.
Linking Sensitive Data - Peter Christen 2021-10-19
This book provides modern technical answers to the legal requirements of pseudonymisation as recommended by privacy legislation. It covers topics such as modern regulatory frameworks for sharing and linking sensitive information, concepts and algorithms for privacy-preserving record linkage and their computational aspects, practical considerations such as dealing with dirty and missing data, as well as privacy, risk, and performance assessment measures. Existing techniques for privacy-preserving record linkage are evaluated empirically and real-world application examples that scale to population sizes are described. The book also includes pointers to freely available software tools, benchmark data sets, and tools to generate synthetic data that can be used to test and evaluate linkage techniques. This book consists of fourteen chapters grouped into four parts, and two appendices. The first part introduces the reader to the topic of linking sensitive data, the second part covers methods and techniques to link such data, the third part discusses aspects of practical importance, and the fourth part provides an outlook of future challenges and open research problems relevant to linking sensitive databases. The appendices provide pointers and describe freely available, open-source software systems that allow the linkage of sensitive data, and provide further details about the evaluations presented. A companion Web site at https://dmm.anu.edu.au/lsdbook2020 provides additional material and Python programs used in the book. This book is mainly written for applied scientists, researchers, and advanced practitioners in governments,

industry, and universities who are concerned with developing, implementing, and deploying systems and tools to share sensitive information in administrative, commercial, or medical databases. The Book describes how linkage methods work and how to evaluate their performance. It covers all the major concepts and methods and also discusses practical matters such as computational efficiency, which are critical if the methods are to be used in practice - and it does all this in a highly accessible way!David J. Hand, Imperial College, London

Computers at Risk - National Research Council 1990-02-01

Computers at Risk presents a comprehensive agenda for developing nationwide policies and practices for computer security. Specific recommendations are provided for industry and for government agencies engaged in computer security activities. The volume also outlines problems and opportunities in computer security research, recommends ways to improve the research infrastructure, and suggests topics for investigators. The book explores the diversity of the field, the need to engineer countermeasures based on speculation of what experts think computer attackers may do next, why the technology community has failed to respond to the need for enhanced security systems, how innovators could be encouraged to bring more options to the marketplace, and balancing the importance of security against the right of privacy.

Database Technologies: Concepts, Methodologies, Tools, and Applications - Erickson, John 2009-02-28

"This reference expands the field of database technologies through four-volumes of in-depth, advanced research articles from nearly 300 of the world's leading professionals"--Provided by publisher.

Methodological Developments in Data Linkage - Katie Harron 2015-09-22

A comprehensive compilation of new developments in data linkage methodology The increasing availability of large administrative databases has led to a dramatic rise in the use of data linkage, yet the standard texts on linkage are still those which describe the seminal work from the 1950-60s, with some updates. Linkage and analysis of data across sources remains problematic due to lack of discriminatory and accurate identifiers, missing data and regulatory issues. Recent developments in data linkage methodology have concentrated on bias and analysis of linked data, novel approaches to organising relationships between databases and privacy-preserving linkage. Methodological Developments in Data Linkage brings together a collection of contributions from members of the international data linkage community, covering cutting edge methodology in this field. It presents opportunities and challenges provided by linkage of large and often complex datasets, including analysis problems, legal and security aspects, models for data access and the development of novel research areas. New methods for handling uncertainty in analysis of linked data, solutions for anonymised linkage and alternative models for data collection are also discussed. Key Features: Presents cutting edge methods for a topic of increasing importance to a wide range of research areas, with applications to data linkage systems internationally Covers the essential issues associated with data linkage today Includes examples based on real data linkage systems, highlighting the opportunities, successes and challenges that the increasing availability of linkage data provides Novel approach incorporates technical aspects of both linkage, management and analysis of linked data This book will be of core interest to academics, government employees, data holders, data managers, analysts and statisticians who use administrative data. It will also appeal to researchers in a variety of areas, including epidemiology, biostatistics, social statistics, informatics, policy and public health.

The Semantic Web. Latest Advances and New Domains - Harald Sack 2016-05-22

The 47 revised full papers presented together with three invited talks were carefully reviewed and selected from 204 submissions. This program was completed by a demonstration and poster session, in which researchers had the chance to present their latest results and advances in the form of live demos. In addition, the PhD Symposium program included 10 contributions, selected out of 21 submissions. The core tracks of the research conference were complemented with new tracks focusing on linked data; machine learning; mobile web, sensors and semantic streams; natural language processing and information retrieval; reasoning; semantic data management, big data, and scalability; services, APIs, processes and cloud computing; smart cities, urban and geospatial data; trust and privacy; and vocabularies, schemas, and ontologies.

Entity Resolution and Information Quality - John R. Talburt 2011-01-14

Entity Resolution and Information Quality presents topics and definitions, and clarifies confusing terminologies regarding entity resolution and information quality. It takes a very wide view of IQ, including its six-domain framework and the skills formed by the International Association for Information and Data Quality {IAIDQ). The book includes chapters that cover the principles of entity resolution and the principles of Information Quality, in addition to their concepts and terminology. It also discusses the Fellegi-Sunter theory of record linkage, the Stanford Entity Resolution Framework, and the Algebraic Model for Entity Resolution, which are the major theoretical models that support Entity Resolution. In relation to this, the book briefly discusses entity-based data integration (EBDI) and its model, which serve as an extension of the Algebraic Model for Entity Resolution. There is also an explanation of how the three commercial ER systems operate and a description of the non-commercial open-source system known as OYSTER. The book concludes by discussing trends in entity resolution research and practice. Students taking IT courses and IT professionals will find this book invaluable. First authoritative reference explaining entity resolution and how to use it effectively Provides practical system design advice to help you get a competitive advantage Includes a companion site with synthetic customer data for applicatory exercises, and access to a Java-based Entity Resolution program.

Data Matching - Peter Christen 2014-08-09

Data matching (also known as record or data linkage, entity resolution, object identification, or field matching) is the task of identifying, matching and merging records that correspond to the same entities from several databases or even within one database. Based on research in various domains including applied statistics, health informatics, data mining, machine learning, artificial intelligence, database management, and digital libraries, significant advances have been achieved over the last decade in all aspects of the data matching process, especially on how to improve the accuracy of data matching, and its scalability to large databases. Peter Christen’s book is divided into three parts: Part I, “Overview”, introduces the subject by presenting several sample applications and their special challenges, as well as a general overview of a generic data matching process. Part II, “Steps of the Data Matching Process”, then details its main steps like pre-processing, indexing, field and record comparison, classification, and quality evaluation. Lastly, part III, “Further Topics”, deals with specific aspects like privacy, real-time matching, or matching unstructured data. Finally, it briefly describes the main features of many research and open source systems available today. By providing the reader with a broad range of data matching concepts and techniques and touching on all aspects of the data matching process, this book helps researchers as well as students specializing in data quality or data matching aspects to familiarize themselves with recent research advances and to identify open research challenges in the area of data matching. To this end, each chapter of the book includes a final section that provides pointers to further background and research material. Practitioners will better understand the current state of the art in data matching as well as the internal workings and limitations of current systems. Especially, they will learn that it is often not feasible to simply implement an existing off-the-shelf data matching system without substantial adaption and customization. Such practical considerations are discussed for each of the major steps in the data matching process.

Handbook on Impact Evaluation - Shahidur R. Khandker 2009-10-13

Public programs are designed to reach certain goals and beneficiaries. Methods to understand whether such programs actually work, as well as the level and nature of impacts on intended beneficiaries, are main themes of this book.

Sharing Clinical Trial Data - Institute of Medicine 2015-04-20

Data sharing can accelerate new discoveries by avoiding duplicative trials, stimulating new ideas for research, and enabling the maximal scientific knowledge and benefits to be gained from the efforts of clinical trial participants and investigators. At the same time, sharing clinical trial data presents risks, burdens, and challenges. These include the need to protect the privacy and honor the consent of clinical trial participants; safeguard the legitimate economic interests of sponsors; and guard against invalid secondary analyses, which could undermine trust in clinical trials or otherwise harm public health. Sharing Clinical Trial Data presents activities and strategies for the responsible sharing of clinical trial data. With the goal of increasing scientific knowledge to lead to better therapies for patients, this book identifies guiding principles and makes recommendations to maximize the benefits and minimize risks. This report

offers guidance on the types of clinical trial data available at different points in the process, the points in the process at which each type of data should be shared, methods for sharing data, what groups should have access to data, and future knowledge and infrastructure needs. Responsible sharing of clinical trial data will allow other investigators to replicate published findings and carry out additional analyses, strengthen the evidence base for regulatory and clinical decisions, and increase the scientific knowledge gained from investments by the funders of clinical trials. The recommendations of Sharing Clinical Trial Data will be useful both now and well into the future as improved sharing of data leads to a stronger evidence base for treatment. This book will be of interest to stakeholders across the spectrum of research--from funders, to researchers, to journals, to physicians, and ultimately, to patients.

Handbook of Big Data Technologies - Albert Y. Zomaya 2017-02-25

This handbook offers comprehensive coverage of recent advancements in Big Data technologies and related paradigms. Chapters are authored by international leading experts in the field, and have been reviewed and revised for maximum reader value. The volume consists of twenty-five chapters organized into four main parts. Part one covers the fundamental concepts of Big Data technologies including data curation mechanisms, data models, storage models, programming models and programming platforms. It also dives into the details of implementing Big SQL query engines and big stream processing systems. Part Two focuses on the semantic aspects of Big Data management including data integration and exploratory ad hoc analysis in addition to structured querying and pattern matching techniques. Part Three presents a comprehensive overview of large scale graph processing. It covers the most recent research in large scale graph processing platforms, introducing several scalable graph querying and mining mechanisms in domains such as social networks. Part Four details novel applications that have been made possible by the rapid emergence of Big Data technologies such as Internet-of-Things (IOT), Cognitive Computing and SCADA Systems. All parts of the book discuss open research problems, including potential opportunities, that have arisen from the rapid progress of Big Data technologies and the associated increasing requirements of application domains. Designed for researchers, IT professionals and graduate students, this book is a timely contribution to the growing Big Data field. Big Data has been recognized as one of leading emerging technologies that will have a major contribution and impact on the various fields of science and varies aspect of the human society over the coming decades. Therefore, the content in this book will be an essential tool to help readers understand the development and future of the field.

Hadoop Application Architectures - Mark Grover 2015-06-30

Get expert guidance on architecting end-to-end data management solutions with Apache Hadoop. While many sources explain how to use various components in the Hadoop ecosystem, this practical book takes you through architectural considerations necessary to tie those components together into a complete tailored application, based on your particular use case. To reinforce those lessons, the book's second section provides detailed examples of architectures used in some of the most commonly found Hadoop applications. Whether you're designing a new Hadoop application, or planning to integrate Hadoop into your existing data infrastructure, Hadoop Application Architectures will skillfully guide you through the process. This book covers: Factors to consider when using Hadoop to store and model data Best practices for moving data in and out of the system Data processing frameworks, including MapReduce, Spark, and Hive Common Hadoop processing patterns, such as removing duplicate records and using windowing analytics Giraph, GraphX, and other tools for large graph processing on Hadoop Using workflow orchestration and scheduling tools such as Apache Oozie Near-real-time stream processing with Apache Storm, Apache Spark Streaming, and Apache Flume Architecture examples for clickstream analysis, fraud detection, and data warehousing

The Behavioral and Social Sciences - National Research Council 1988-02-01

This volume explores the scientific frontiers and leading edges of research across the fields of anthropology, economics, political science, psychology, sociology, history, business, education, geography, law, and psychiatry, as well as the newer, more specialized areas of artificial intelligence, child development, cognitive science, communications, demography, linguistics, and management and decision science. It includes recommendations concerning new resources, facilities, and programs that may be needed over the next several years to ensure rapid progress and provide a high level of returns to basic research.

Python for Data Analysis - Wes McKinney 2017-09-25

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

Entity Information Life Cycle for Big Data - John R. Talburt 2015-04-20

Entity Information Life Cycle for Big Data walks you through the ins and outs of managing entity information so you can successfully achieve master data management (MDM) in the era of big data. This book explains big data's impact on MDM and the critical role of entity information management system (EIMS) in successful MDM. Expert authors Dr. John R. Talburt and Dr. Yinle Zhou provide a thorough background in the principles of managing the entity information life cycle and provide practical tips and techniques for implementing an EIMS, strategies for exploiting distributed processing to handle big data for EIMS, and examples from real applications. Additional material on the theory of EIIM and methods for assessing and evaluating EIMS performance also make this book appropriate for use as a textbook in courses on entity and identity management, data management, customer relationship management (CRM), and related topics. Explains the business value and impact of entity information management system (EIMS) and directly addresses the problem of EIMS design and operation, a critical issue organizations face when implementing MDM systems Offers practical guidance to help you design and build an EIM system that will successfully handle big data Details how to measure and evaluate entity integrity in MDM systems and explains the principles and processes that comprise EIM Provides an understanding of features and functions an EIM system should have that will assist in evaluating commercial EIM systems Includes chapter review questions, exercises, tips, and free downloads of demonstrations that use the OYSTER open source EIM system Executable code (Java .jar files), control scripts, and synthetic input data illustrate various aspects of CSRUD life cycle such as identity capture, identity update, and assertions

Model Rules of Professional Conduct - American Bar Association. House of Delegates 2007

The Model Rules of Professional Conduct provides an up-to-date resource for information on legal ethics. Federal, state and local courts in all jurisdictions look to the Rules for guidance in solving lawyer malpractice cases, disciplinary actions, disqualification issues, sanctions questions and much more. In this volume, black-letter Rules of Professional Conduct are followed by numbered Comments that explain each Rule's purpose and provide suggestions for its practical application. The Rules will help you identify proper conduct in a variety of given situations, review those instances where discretionary action is possible, and define the nature of the relationship between you and your clients, colleagues and the courts.

Data Profiling - Ziawasch Abedjan 2018-11-08

Data profiling refers to the activity of collecting data about data, i.e., metadata. Most IT professionals and researchers who work with data have engaged in data profiling, at least informally, to understand and explore an unfamiliar dataset or to determine whether a new dataset is appropriate for a particular task at hand. Data profiling results are also important in a variety of other situations, including query optimization, data integration, and data cleaning. Simple metadata are statistics, such as the number of rows and columns, schema and datatype information, the number of distinct values, statistical value distributions, and the number of null or empty values in each column. More complex types of metadata are statements about multiple columns and their correlation, such as candidate keys, functional dependencies, and other types of dependencies. This book provides a classification of the various types of profilable metadata,

discusses popular data profiling tasks, and surveys state-of-the-art profiling algorithms. While most of the book focuses on tasks and algorithms for relational data profiling, we also briefly discuss systems and techniques for profiling non-relational data such as graphs and text. We conclude with a discussion of data profiling challenges and directions for future work in this area.

Fixing Access Annoyances - Phil Mitchell 2006-02-21

Provides a collection of tips on fixing annoyances found in Microsoft Access, covering such topics as performance, security, database design, queries, forms, page layout, macros, and expressions.

Registries for Evaluating Patient Outcomes - Agency for Healthcare Research and Quality/AHRQ 2014-04-01

This User's Guide is intended to support the design, implementation, analysis, interpretation, and quality evaluation of registries created to increase understanding of patient outcomes. For the purposes of this guide, a patient registry is an organized system that uses observational study methods to collect uniform data (clinical and other) to evaluate specified outcomes for a population defined by a particular disease, condition, or exposure, and that serves one or more predetermined scientific, clinical, or policy purposes. A registry database is a file (or files) derived from the registry. Although registries can serve many purposes, this guide focuses on registries created for one or more of the following purposes: to describe the natural history of disease, to determine clinical effectiveness or cost-effectiveness of health care products and services, to measure or monitor safety and harm, and/or to measure quality of care. Registries are classified according to how their populations are defined. For example, product registries include patients who have been exposed to biopharmaceutical products or medical devices. Health services registries consist of patients who have had a common procedure, clinical encounter, or hospitalization. Disease or condition registries are defined by patients having the same diagnosis, such as cystic fibrosis or heart failure. The User's Guide was created by researchers affiliated with AHRQ's Effective Health Care Program, particularly those who participated in AHRQ's DEcIDE (Developing Evidence to Inform Decisions About Effectiveness) program. Chapters were subject to multiple internal and external independent reviews.

Data Quality and Record Linkage Techniques - Thomas N. Herzog 2007-05-23

This book offers a practical understanding of issues involved in improving data quality through editing, imputation, and record linkage. The first part of the book deals with methods and models, focusing on the Fellegi-Holt edit-imputation model, the Little-Rubin multiple-imputation scheme, and the Fellegi-Sunter record linkage model. The second part presents case studies in which these techniques are applied in a variety of areas, including mortgage guarantee insurance, medical, biomedical, highway safety, and social insurance as well as the construction of list frames and administrative lists. This book offers a mixture of practical advice, mathematical rigor, management insight and philosophy.

Designing Data-Intensive Applications - Martin Kleppmann 2017-03-16

Data is at the center of many challenges in system design today. Difficult issues need to be figured out, such as scalability, consistency, reliability, efficiency, and maintainability. In addition, we have an overwhelming variety of tools, including relational databases, NoSQL datastores, stream or batch processors, and message brokers. What are the right choices for your application? How do you make sense of all these buzzwords? In this practical and comprehensive guide, author Martin Kleppmann helps you navigate this diverse landscape by examining the pros and cons of various technologies for processing and storing data. Software keeps changing, but the fundamental principles remain the same. With this book, software engineers and architects will learn how to apply those ideas in practice, and how to make full use of data in modern applications. Peer under the hood of the systems you already use, and learn how to use and operate them more effectively Make informed decisions by identifying the strengths and weaknesses of different tools Navigate the trade-offs around consistency, scalability, fault tolerance, and complexity Understand the distributed systems research upon which modern databases are built Peek behind the scenes of major online services, and learn from their architectures

Semantic Systems. In the Era of Knowledge Graphs - Eva Blomqvist 2020-10-26

This open access book constitutes the refereed proceedings of the 16th International Conference on Semantic Systems, SEMANTiCS 2020, held in Amsterdam, The Netherlands, in September 2020. The conference was held virtually due to the COVID-19 pandemic.

Secondary Analysis of Electronic Health Records - MIT Critical Data 2016-09-09

This book trains the next generation of scientists representing different disciplines to leverage the data generated during routine patient care. It formulates a more complete lexicon of evidence-based recommendations and support shared, ethical decision making by doctors with their patients. Diagnostic and therapeutic technologies continue to evolve rapidly, and both individual practitioners and clinical teams face increasingly complex ethical decisions. Unfortunately, the current state of medical knowledge does not provide the guidance to make the majority of clinical decisions on the basis of evidence. The present research infrastructure is inefficient and frequently produces unreliable results that cannot be replicated. Even randomized controlled trials (RCTs), the traditional gold standards of the research reliability hierarchy, are not without limitations. They can be costly, labor intensive, and slow, and can return results that are seldom generalizable to every patient population. Furthermore, many pertinent but unresolved clinical and medical systems issues do not seem to have attracted the interest of the research enterprise, which has come to focus instead on cellular and molecular investigations and single-agent (e.g., a drug or device) effects. For clinicians, the end result is a bit of a "data desert" when it comes to making decisions. The new research infrastructure proposed in this book will help the medical profession to make ethically sound and well informed decisions for their patients.

R for Data Science - Hadley Wickham 2016-12-12

Learn how to use R to turn raw data into insight, knowledge, and understanding. 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, R for Data Science is designed to get you doing data science as quickly as possible. Authors Hadley Wickham and Garrett Grolemund guide you through the steps of importing, wrangling, exploring, and modeling your data and communicating the results. You'll get a complete, big-picture understanding of the data science cycle, along with basic tools you need to manage the details. Each section of the book is paired with exercises to help you practice what you've learned along the way. You'll learn how to: Wrangle—transform your datasets into a form convenient for analysis Program—learn powerful R tools for solving data problems with greater clarity and ease Explore—examine your data, generate hypotheses, and quickly test them Model—provide a low-dimensional summary that captures true "signals" in your dataset Communicate—learn R Markdown for integrating prose, code, and results

Strengthening Forensic Science in the United States - National Research Council 2009-07-29

Scores of talented and dedicated people serve the forensic science community, performing vitally important work. However, they are often constrained by lack of adequate resources, sound policies, and national support. It is clear that change and advancements, both systematic and scientific, are needed in a number of forensic science disciplines to ensure the reliability of work, establish enforceable standards, and promote best practices with consistent application. Strengthening Forensic Science in the United States: A Path Forward provides a detailed plan for addressing these needs and suggests the creation of a new government entity, the National Institute of Forensic Science, to establish and enforce standards within the forensic science community. The benefits of improving and regulating the forensic science disciplines are clear: assisting law enforcement officials, enhancing homeland security, and reducing the risk of wrongful conviction and exoneration. Strengthening Forensic Science in the United States gives a full account of what is needed to advance the forensic science disciplines, including upgrading of systems and organizational structures, better training, widespread adoption of uniform and enforceable best practices, and mandatory certification and accreditation programs. While this book provides an essential call-to-action for congress and policy makers, it also serves as a vital tool for law enforcement agencies, criminal prosecutors and attorneys, and forensic science educators.

SQL Cookbook - Anthony Molinaro 2006

A guide to SQL covers such topics as retrieving records, metadata queries, working with strings, data arithmetic, date manipulation, reporting and warehousing, and hierarchical queries.

Geospatial Intelligence: Concepts, Methodologies, Tools, and Applications - Management Association, Information Resources 2019-03-01

Decision makers, such as government officials, need to better understand human activity in order to make

informed decisions. With the ability to measure and explore geographic space through the use of geospatial intelligence data sources including imagery and mapping data, they are better able to measure factors affecting the human population. As a broad field of study, geospatial research has applications in a variety of fields including military science, environmental science, civil engineering, and space exploration. Geospatial Intelligence: Concepts, Methodologies, Tools, and Applications explores multidisciplinary applications of geographic information systems to describe, assess, and visually depict physical features and to gather data, information, and knowledge regarding human activity. Highlighting a range of topics such as geovisualization, spatial analysis, and landscape mapping, this multi-volume book is ideally designed for data scientists, engineers, government agencies, researchers, and graduate-level students in GIS programs.

Data Mining: Concepts and Techniques - Jiawei Han 2011-06-09

Data Mining: Concepts and Techniques provides the concepts and techniques in processing gathered data or information, which will be used in various applications. Specifically, it explains data mining and the tools used in discovering knowledge from the collected data. This book is referred as the knowledge discovery from data (KDD). It focuses on the feasibility, usefulness, effectiveness, and scalability of techniques of large data sets. After describing data mining, this edition explains the methods of knowing, preprocessing, processing, and warehousing data. It then presents information about data warehouses, online analytical processing (OLAP), and data cube technology. Then, the methods involved in mining frequent patterns, associations, and correlations for large data sets are described. The book details the methods for data classification and introduces the concepts and methods for data clustering. The remaining chapters discuss the outlier detection and the trends, applications, and research frontiers in data mining. This book is intended for Computer Science students, application developers, business professionals, and researchers who seek information on data mining. Presents dozens of algorithms and implementation examples, all in pseudo-code and suitable for use in real-world, large-scale data mining projects Addresses advanced topics such as mining object-relational databases, spatial databases, multimedia databases, time-series databases, text databases, the World Wide Web, and applications in several fields Provides a comprehensive, practical look at the concepts and techniques you need to get the most out of your data

Knowledge Graphs and Big Data Processing - Valentina Janev 2020-01-01

This open access book is part of the LAMBDA Project (Learning, Applying, Multiplying Big Data Analytics), funded by the European Union, GA No. 809965. Data Analytics involves applying algorithmic processes to derive insights. Nowadays it is used in many industries to allow organizations and companies to make better decisions as well as to verify or disprove existing theories or models. The term data analytics is often used interchangeably with intelligence, statistics, reasoning, data mining, knowledge discovery, and others. The goal of this book is to introduce some of the definitions, methods, tools, frameworks, and solutions for big data processing, starting from the process of information extraction and knowledge representation, via knowledge processing and analytics to visualization, sense-making, and practical applications. Each chapter in this book addresses some pertinent aspect of the data processing chain, with a specific focus on understanding Enterprise Knowledge Graphs, Semantic Big Data Architectures, and Smart Data Analytics solutions. This book is addressed to graduate students from technical disciplines, to professional audiences following continuous education short courses, and to researchers from diverse areas following self-study courses. Basic skills in computer science, mathematics, and statistics are required.

Web Database Applications with PHP and MySQL - Hugh E. Williams 2002

Combines language tutorials with application design advice to cover the PHP server-side scripting language and the MySQL database engine.

Principles of Data Wrangling - Tye Rattenbury 2017-06-29

A key task that any aspiring data-driven organization needs to learn is data wrangling, the process of converting raw data into something truly useful. This practical guide provides business analysts with an overview of various data wrangling techniques and tools, and puts the practice of data wrangling into context by asking, "What are you trying to do and why?" Wrangling data consumes roughly 50-80% of an analyst’s time before any kind of analysis is possible. Written by key executives at Trifacta, this book walks you through the wrangling process by exploring several factors—time, granularity, scope, and structure—that you need to consider as you begin to work with data. You’ll learn a shared language and a comprehensive understanding of data wrangling, with an emphasis on recent agile analytic processes used by many of today’s data-driven organizations. Appreciate the importance—and the satisfaction—of wrangling data the right way. Understand what kind of data is available Choose which data to use and at what level of detail Meaningfully combine multiple sources of data Decide how to distill the results to a size and shape that can drive downstream analysis

Data Matching - Peter Christen 2012-07-04

Data matching (also known as record or data linkage, entity resolution, object identification, or field matching) is the task of identifying, matching and merging records that correspond to the same entities from several databases or even within one database. Based on research in various domains including applied statistics, health informatics, data mining, machine learning, artificial intelligence, database management, and digital libraries, significant advances have been achieved over the last decade in all aspects of the data matching process, especially on how to improve the accuracy of data matching, and its scalability to large databases. Peter Christen’s book is divided into three parts: Part I, “Overview”, introduces the subject by presenting several sample applications and their special challenges, as well as a general overview of a generic data matching process. Part II, “Steps of the Data Matching Process”, then details its main steps like pre-processing, indexing, field and record comparison, classification, and quality evaluation. Lastly, part III, “Further Topics”, deals with specific aspects like privacy, real-time matching, or matching unstructured data. Finally, it briefly describes the main features of many research and open source systems available today. By providing the reader with a broad range of data matching concepts and techniques and touching on all aspects of the data matching process, this book helps researchers as well as students specializing in data quality or data matching aspects to familiarize themselves with recent research advances and to identify open research challenges in the area of data matching. To this end, each chapter of the book includes a final section that provides pointers to further background and research material. Practitioners will better understand the current state of the art in data matching as well as the internal workings and limitations of current systems. Especially, they will learn that it is often not feasible to simply implement an existing off-the-shelf data matching system without substantial adaption and customization. Such practical considerations are discussed for each of the major steps in the data matching process.

Data Mining for Business Intelligence - Galit Shmueli 2006-12-11

Learn how to develop models for classification, prediction, and customer segmentation with the help of Data Mining for Business Intelligence In today's world, businesses are becoming more capable of accessing their ideal consumers, and an understanding of data mining contributes to this success. Data Mining for Business Intelligence, which was developed from a course taught at the Massachusetts Institute of Technology's Sloan School of Management, and the University of Maryland's Smith School of Business, uses real data and actual cases to illustrate the applicability of data mining intelligence to the development of successful business models. Featuring XLMiner, the Microsoft Office Excel add-in, this book allows readers to follow along and implement algorithms at their own speed, with a minimal learning curve. In addition, students and practitioners of data mining techniques are presented with hands-on, business-oriented applications. An abundant amount of exercises and examples are provided to motivate learning and understanding. Data Mining for Business Intelligence: Provides both a theoretical and practical understanding of the key methods of classification, prediction, reduction, exploration, and affinity analysis Features a business decision-making context for these key methods Illustrates the application and interpretation of these methods using real business cases and data This book helps readers understand the beneficial relationship that can be established between data mining and smart business practices, and is an excellent learning tool for creating valuable strategies and making wiser business decisions.

Data Analytics in Medicine: Concepts, Methodologies, Tools, and Applications - Management Association, Information Resources 2019-12-06

Advancements in data science have created opportunities to sort, manage, and analyze large amounts of data more effectively and efficiently. Applying these new technologies to the healthcare industry, which has vast quantities of patient and medical data and is increasingly becoming more data-reliant, is crucial for

refining medical practices and patient care. Data Analytics in Medicine: Concepts, Methodologies, Tools, and Applications is a vital reference source that examines practical applications of healthcare analytics for improved patient care, resource allocation, and medical performance, as well as for diagnosing, predicting, and identifying at-risk populations. Highlighting a range of topics such as data security and privacy, health informatics, and predictive analytics, this multi-volume book is ideally designed for doctors, hospital administrators, nurses, medical professionals, IT specialists, computer engineers, information technologists, biomedical engineers, data-processing specialists, healthcare practitioners, academicians, and researchers interested in current research on the connections between data analytics in the field of medicine.

Data Mining with Rattle and R - Graham Williams 2011-08-04

Data mining is the art and science of intelligent data analysis. By building knowledge from information, data mining adds considerable value to the ever increasing stores of electronic data that abound today. In performing data mining many decisions need to be made regarding the choice of methodology, the choice of data, the choice of tools, and the choice of algorithms. Throughout this book the reader is introduced to the basic concepts and some of the more popular algorithms of data mining. With a focus on the hands-on end-to-end process for data mining, Williams guides the reader through various capabilities of the easy to use, free, and open source Rattle Data Mining Software built on the sophisticated R Statistical Software. The focus on doing data mining rather than just reading about data mining is refreshing. The book covers data understanding, data preparation, data refinement, model building, model evaluation, and practical deployment. The reader will learn to rapidly deliver a data mining project using software easily installed for free from the Internet. Coupling Rattle with R delivers a very sophisticated data mining environment with all the power, and more, of the many commercial offerings.

Data-Driven Policy Impact Evaluation - Nuno Crato 2018-10-02

In the light of better and more detailed administrative databases, this open access book provides statistical tools for evaluating the effects of public policies advocated by governments and public institutions. Experts from academia, national statistics offices and various research centers present modern econometric methods for an efficient data-driven policy evaluation and monitoring, assess the causal effects of policy measures and report on best practices of successful data management and usage. Topics include data confidentiality, data linkage, and national practices in policy areas such as public health, education and employment. It offers scholars as well as practitioners from public administrations, consultancy firms and nongovernmental organizations insights into counterfactual impact evaluation methods and the potential of data-based policy and program evaluation.

Data Wrangling with Python - Jacqueline Kazil 2016-02-04

How do you take your data analysis skills beyond Excel to the next level? By learning just enough Python to get stuff done. This hands-on guide shows non-programmers like you how to process information that’s initially too messy or difficult to access. You don't need to know a thing about the Python programming language to get started. Through various step-by-step exercises, you’ll learn how to acquire, clean, analyze, and present data efficiently. You’ll also discover how to automate your data process, schedule file- editing and clean-up tasks, process larger datasets, and create compelling stories with data you obtain. Quickly learn basic Python syntax, data types, and language concepts Work with both machine-readable and human-consumable data Scrape websites and APIs to find a bounty of useful information Clean and format data to eliminate duplicates and errors in your datasets Learn when to standardize data and when to test and script data cleanup Explore and analyze your datasets with new Python libraries and techniques Use Python solutions to automate your entire data-wrangling process