Data Matching Concepts And Techniques For Record Linkage Entity Resolution And Duplicate Detection Data Centric Systems And Applications

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.

A survey of computational methods for understanding, generating, and manipulating human language, which offers a synthesis of classical representations and algorithms with contemporary machine learning techniques. This textbook provides a technical perspective on natural language processing—methods for building computer software that understands, generates, and manipulates human language. It emphasizes contemporary data-driven approaches, focusing on techniques from supervised and unsupervised machine learning. The first section establishes a foundation in machine learning by building a set of tools that will be used throughout the book and applying them to word-based textual analysis. The second section introduces structured representations of language, including sequences, trees, and graphs. The third section explores different approaches to the representation and analysis of linguistic meaning, ranging from formal logic to neural word embeddings. The final section offers chapter-length treatments of three transformative applications of natural language processing: information extraction, machine translation, and text generation. End-of-chapter exercises include both paper-and-pencil analysis and software implementation. The text synthesizes and distills a broad and diverse research literature, linking contemporary machine learning techniques with the field's linguistic and computational foundations. It is suitable for use in advanced undergraduate and graduate-level courses and as a reference for software engineers and data scientists. Readers should have a background in computer programming and college-level mathematics. After mastering the material presented, students will have the technical skill to build and analyze novel natural language processing systems and to understand the latest research in the field.

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.

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.

Graph data closes the gap between the way humans and computers view the world. While computers rely on static rows and columns of data, people navigate and reason about life through relationships. This practical guide demonstrates how graph data brings these two approaches together. By working with concepts from graph theory, database schema, distributed systems, and data analysis, you'll arrive at a unique intersection known as graph thinking. Authors Denise Koessler Gosnell and Matthias Broecheler show data engineers, data scientists, and data analysts how to solve complex problems with graph databases. You'll explore templates for building with graph technology, along with examples that demonstrate how teams think about graph data within an application. Build an example application architecture with relational and graph technologies Use graph technology to build a Customer 360 application, the most popular graph data pattern today Dive into hierarchical data and troubleshoot a new paradigm that comes from working with graph data Find paths in graph data and learn why your trust in different paths motivates and informs your preferences Use collaborative filtering to design a Netflix-inspired recommendation system

"Comprising more than 500 entries, the Encyclopedia of Research Design explains how to make decisions about research design, undertake research projects in an ethical manner, interpret and draw valid inferences from data, and evaluate experiment design strategies and results. Two additional features carry this encyclopedia far above other works in the field: bibliographic entries devoted to significant articles in the history of research design and reviews of contemporary tools, such as software and statistical procedures, used to analyze results. It covers the spectrum of research design strategies, from material presented in introductory classes to topics necessary in graduate research; it addresses cross- and multidisciplinary research needs, with many examples drawn from the social and behavioral sciences, neurosciences, and biomedical and life sciences; it provides summaries of advantages and disadvantages of often-used strategies; and it uses hundreds of sample tables, figures, and equations based on real-life cases."--Publisher's description.

Fully updated to reflect the most recent changes in the field, the Second Edition of Propensity Score Analysis provides an accessible, systematic review of the origins, history, and statistical foundations of propensity score analysis, illustrating how it can be used for solving evaluation and causal-inference problems. With a strong focus on practical applications, the authors explore various strategies for employing PSA, discuss the use of PSA with alternative types of data, and delineate the limitations of PSA under a variety of constraints. Unlike existing textbooks on program evaluation and causal inference, this book delves into

statistical concepts, formulas, and models within the context of a robust and engaging focus on application.
"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.
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.
This book starts with an introduction to process modeling and process paradigms, then explains how to query and analyze process models, and how to analyze the process execution data. In this way, readers receive a comprehensive overview of what is needed to identify, understand and improve business processes. The book chiefly focuses on concepts, techniques and methods. It covers a large body of knowledge on process analytics – including process data querying, analysis, matching and correlating process data and models – to help practitioners and researchers understand the underlying concepts, problems, methods, tools and techniques involved in modern process analytics. Following an introduction to basic business process and process analytics concepts, it describes the state of the art in this area before examining different analytics techniques in detail. In this regard, the book covers analytics over different levels of process abstractions, from process execution data and methods for linking and correlating process execution data, to inferring process models, querying process execution data and process models, and scalable process data analytics methods. In addition, it provides a review of commercial process analytics tools and their practical applications. The book is intended for a broad readership interested in business process management and process analytics. It provides researchers with an introduction to these fields by comprehensively classifying the current state of research, by describing in-depth techniques and methods, and by highlighting future research directions. Lecturers will find a wealth of material to choose from for a variety of courses, ranging from undergraduate courses in business process management to graduate courses in business process analytics. Lastly, it offers professionals a reference guide to the state of the art in commercial tools and techniques, complemented by many real-world use case scenarios.
Data Pipelines with Apache Airflow teaches you the ins-and-outs of the Directed Acyclic Graphs (DAGs) that power Airflow, and how to write your own DAGs to meet the needs of your projects. With complete coverage of both foundational and lesser-known features, when you're done you'll be set to start using Airflow for seamless data pipeline development and management. Pipelines can be challenging to manage, especially when your data has to flow through a collection of application components, servers, and cloud services. Airflow lets you schedule, restart, and backfill pipelines, and its easy-to-use UI and workflows with Python scripting has users praising its incredible flexibility. Data Pipelines with Apache Airflow takes you through best practices for creating pipelines for multiple tasks, including data lakes, cloud deployments, and data science. Data Pipelines with Apache Airflow teaches you the ins-and-outs of the Directed Acyclic Graphs (DAGs) that power Airflow, and how to write your own DAGs to meet the needs of your projects. With complete coverage of both foundational and lesser-known features, when you're done you'll be set to start using Airflow for seamless data pipeline development and management. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications.
As political, economic, and environmental issues increasingly spread across the globe, the science of geography is being rediscovered by scientists, policymakers, and educators alike. Geography has been made a core subject in U.S. schools, and scientists from a variety of disciplines are using analytical tools originally developed by geographers. Rediscovering Geography presents a broad overview of geography's renewed importance in a changing world. Through discussions and highlighted case studies, this book illustrates geography's impact on international trade, environmental change, population growth, information infrastructure, the condition of cities, the spread of AIDS, and much more. The committee examines some of the more significant tools for data collection, storage, analysis, and display, with examples of major contributions made by geographers. Rediscovering Geography provides a blueprint for the future of the discipline, recommending how to strengthen its intellectual and institutional foundation and meet the demand for geographic expertise among professionals and the public.
Managing Data in Motion describes techniques that have been developed for significantly reducing the complexity of managing system interfaces and enabling scalable architectures. Author April Reeve brings over two decades of experience to present a vendor-neutral approach to moving data between computing environments and systems. Readers will learn the techniques, technologies, and best practices for managing the passage of data between computer systems and integrating disparate data together in an enterprise environment. The average enterprise's computing environment is comprised of hundreds to thousands computer systems that have been built, purchased, and acquired over time. The data from these various systems needs to be integrated for reporting and analysis, shared for business transaction processing, and converted from one format to another when old systems are replaced and new systems are acquired. The management of the "data in motion" in organizations is rapidly becoming one of the biggest concerns for business and IT management. Data warehousing and conversion, real-time data integration, and cloud and "big data" applications are just a few of the challenges facing organizations and businesses today. Managing Data in Motion tackles these and other topics in a style easily understood by business and IT managers as well as programmers and architects. Presents a vendor-neutral overview of the different technologies and techniques for moving data between computer systems including the emerging solutions for unstructured as well as structured data types Explains, in non-technical terms, the architecture and components required to perform data integration Describes how to reduce the complexity of managing system interfaces and enable a scalable data architecture that can handle the dimensions of "Big Data"
The Book of R is a comprehensive, beginner-friendly guide to R, the world’s most popular programming language for statistical analysis. Even if you have no programming experience and little more than a grounding in the basics of mathematics, you’ll find everything you need

to begin using R effectively for statistical analysis. You'll start with the basics, like how to handle data and write simple programs, before moving on to more advanced topics, like producing statistical summaries of your data and performing statistical tests and modeling. You'll even learn how to create impressive data visualizations with R's basic graphics tools and contributed packages, like ggplot2 and ggvis, as well as interactive 3D visualizations using the rgl package. Dozens of hands-on exercises (with downloadable solutions) take you from theory to practice, as you learn: –The fundamentals of programming in R, including how to write data frames, create functions, and use variables, statements, and loops –Statistical concepts like exploratory data analysis, probabilities, hypothesis tests, and regression modeling, and how to execute them in R –How to access R's thousands of functions, libraries, and data sets –How to draw valid and useful conclusions from your data –How to create publication-quality graphics of your results Combining detailed explanations with real-world examples and exercises, this book will provide you with a solid understanding of both statistics and the depth of R's functionality. Make The Book of R your doorway into the growing world of data analysis.

How do you approach answering queries when your data is stored in multiple databases that were designed independently by different people? This is first comprehensive book on data integration and is written by three of the most respected experts in the field. This book provides an extensive introduction to the theory and concepts underlying today's data integration techniques, with detailed, instruction for their application using concrete examples throughout to explain the concepts. Data integration is the problem of answering queries that span multiple data sources (e.g., databases, web pages). Data integration problems surface in multiple contexts, including enterprise information integration, query processing on the Web, coordination between government agencies and collaboration between scientists. In some cases, data integration is the key bottleneck to making progress in a field. The authors provide a working knowledge of data integration concepts and techniques, giving you the tools you need to develop a complete and concise package of algorithms and applications. Offers a range of data integration solutions enabling you to focus on what is most relevant to the problem at hand Enables you to build your own algorithms and implement your own data integration applications

Due to the growing use of web applications and communication devices, the use of data has increased throughout various industries, including business and healthcare. It is necessary to develop specific software programs that can analyze and interpret large amounts of data quickly in order to ensure adequate usage and predictive results. Cognitive Analytics: Concepts, Methodologies, Tools, and Applications provides emerging perspectives on the theoretical and practical aspects of data analysis tools and techniques. It also examines the incorporation of pattern management as well as decision-making and prediction processes through the use of data management and analysis. Highlighting a range of topics such as natural language processing, big data, and pattern recognition, this multi-volume book is ideally designed for information technology professionals, software developers, data analysts, graduate-level students, researchers, computer engineers, software engineers, IT specialists, and academicians.

Databases and information systems are the backbone of modern information technology and are crucial to the IT systems which support all aspects of our everyday life; from government, education and healthcare, to business processes and the storage of our personal photos and archives. This book presents 22 of the best revised papers accepted following stringent peer review for the 11th International Baltic Conference on Databases and Information Systems (Baltic DB&IS 2014), held in Tallinn, Estonia, in June 2014. The conference provided a forum for the exchange of scientific achievements between the research communities of the Baltic countries and the rest of the world in the area of databases and information systems, bringing together researchers, practitioners and Ph.D. students from many countries. The subject areas covered at the conference focused on big data processing, data warehouses, data integration and services, data and knowledge management, e-government, as well as e-services and e-learning.

Big Data and Social Science: Data Science Methods and Tools for Research and Practice, Second Edition shows how to apply data science to real-world problems, covering all stages of a data-intensive social science or policy project. Prominent leaders in the social sciences, statistics, and computer science as well as the field of data science provide a unique perspective on how to apply modern social science research principles and current analytical and computational tools. The text teaches you how to identify and collect appropriate data, apply data science methods and tools to the data, and recognize and respond to data errors, biases, and limitations. Features Takes an accessible, hands-on approach to handling new types of data in the social sciences Presents the key data science tools in a non-intimidating way to both social and data scientists while keeping the focus on research questions and purposes Illustrates social science and data science principles through real-world problems Links computer science concepts to practical social science research Promotes good scientific practice Provides freely available data and code as well as practical programming exercises through Binder and GitHub New to the Second Edition Increased use of examples from different areas of social sciences New chapter on dealing with Bias and Fairness in Machine Learning models Expanded chapters focusing on Machine Learning and Text Analysis Revamped hands-on Jupyter notebooks to reinforce concepts covered in each chapter This classroom-tested book fills a major gap in graduate- and professional-level data science and social science education. It can be used to train a new generation of social data scientists to tackle real-world problems and improve the skills and competencies of applied social scientists and public policy practitioners. It empowers you to use the massive and rapidly growing amounts of available data to interpret economic and social activities in a scientific and rigorous manner.

The contributors to Best Practices in Quantitative Methods envision quantitative methods in the 21st century, identify the best practices, and, where possible, demonstrate the superiority of their recommendations empirically. Editor Jason W. Osborne designed this book with the goal of providing readers with the most effective, evidence-based, modern quantitative methods and quantitative data analysis across the social and behavioral sciences. The text is divided into five main sections covering select best practices in Measurement, Research Design, Basics of Data Analysis, Quantitative Methods, and Advanced Quantitative Methods. Each chapter contains a current and expansive review of the literature, a case for best practices in terms of method, outcomes, inferences, etc., and broad-ranging examples along with any empirical evidence to show why certain techniques are better. Key Features: Describes important implicit knowledge to readers: The chapters in this volume explain the important details of seemingly mundane aspects of quantitative research, making them accessible to readers and demonstrating why it is important to pay attention to these details. Compares and contrasts analytic techniques: The book examines instances where there are multiple options for doing things, and make recommendations as to what is the "best" choice—or choices, as what is best often depends on the circumstances. Offers new procedures to update and explicate traditional techniques: The featured scholars present and explain new options for data analysis, discussing the advantages and disadvantages of the new procedures in depth, describing how to perform them, and demonstrating their use. Intended Audience: Representing the vanguard of research methods for the 21st century, this book is an invaluable resource for graduate students and researchers who want a comprehensive, authoritative resource for practical and sound advice from leading experts in quantitative methods.

"This book explores multidisciplinary applications of geographic information systems and technologies in addition to the latest trends and developments in the field. It also examines land administration, encompassing both cadastral systems and land registration, as well as the methods of land governance strategies. Highlighting a range of topics such as geovisualization, spatial analysis, and landscape mapping"--

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.

"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"--

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.

With the increased use of technology in modern society, high volumes of multimedia information exists. It is important for businesses, organizations, and individuals to understand how to optimize this data and new methods are emerging for more efficient information management and retrieval. Information Retrieval and Management: Concepts, Methodologies, Tools, and Applications is an innovative reference source for the latest academic material in the field of information and communication technologies and explores how complex information systems interact with and affect one another. Highlighting a range of topics such as knowledge discovery, semantic web, and information resources management, this multi-volume book is ideally designed for researchers, developers, managers, strategic planners, and advanced-level students.

Developments in technologies have evolved in a much wider use of technology throughout science, government, and business; resulting in the expansion of geographic information systems. GIS is the academic study and practice of presenting geographical data through a system designed to capture, store, analyze, and manage geographic information. Geographic Information Systems: Concepts, Methodologies, Tools, and Applications is a collection of knowledge on the latest advancements and research of geographic information systems. This book aims to be useful for academics and practitioners involved in geographical data.

The environment for obtaining information and providing statistical data for policy makers and the public has changed significantly in the past decade, raising questions about the fundamental survey paradigm that underlies federal statistics. New data sources provide opportunities to develop a new paradigm that can improve timeliness, geographic or subpopulation detail, and statistical efficiency. It also has the potential to reduce the costs of producing federal statistics. The panel's first report described federal statistical agencies' current paradigm, which relies heavily on sample surveys for producing national statistics, and challenges agencies are facing; the legal frameworks and mechanisms for protecting the privacy and confidentiality of statistical data and for providing researchers access to data, and challenges to those frameworks and mechanisms; and statistical agencies access to alternative sources of data. The panel recommended a new approach for federal statistical programs that would combine diverse data sources from government and private sector sources and the creation of a new entity that would provide the foundational elements needed for this new approach, including legal authority to access data and protect privacy. This second of the panel's two reports builds on the analysis, conclusions, and recommendations in the first one. This report assesses alternative methods for implementing a new approach that would combine diverse data sources from government and private sector sources, including describing statistical models for combining data from multiple sources; examining statistical and computer science approaches that foster privacy protections; evaluating frameworks for assessing the quality and utility of alternative data sources; and various models for implementing the recommended new entity. Together, the two reports offer ideas and recommendations to help federal statistical agencies examine and evaluate data from alternative sources and then combine them as appropriate to provide the country with more timely, actionable, and useful information for policy makers, businesses, and individuals.

Missing data pose challenges to real-life data analysis. Simple ad-hoc fixes, like deletion or mean imputation, only work under highly restrictive conditions, which are often not met in practice. Multiple imputation replaces each missing value by multiple plausible values. The variability between these replacements reflects our ignorance of the true (but missing) value. Each of the completed data set is then analyzed by standard methods, and the results are pooled to obtain unbiased estimates with correct confidence intervals. Multiple imputation is a general approach that also inspires novel solutions to old problems by reformulating the task at hand as a missing-data problem. This is the second edition of a popular book on multiple imputation, focused on explaining the application of methods through detailed worked examples using the MICE package as developed by the author.

This new edition incorporates the recent developments in this fast-moving field. This class-tested book avoids mathematical and technical details as much as possible: formulas are accompanied by verbal statements that explain the formula in accessible terms. The book sharpens the reader's intuition on how to think about missing data, and provides all the tools needed to execute a well-grounded quantitative analysis in the presence of missing data.

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.

Now in its second edition, this book focuses on practical algorithms for mining data from even the largest datasets.

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.

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.

This comprehensive reference consists of 18 chapters from prominent researchers in the field. Each chapter is self-contained, and synthesizes one aspect of frequent pattern mining. An emphasis is placed on simplifying the content, so that students and practitioners can benefit from the book. Each chapter contains a survey describing key research on the topic, a case study and future directions. Key topics include: Pattern Growth Methods, Frequent Pattern Mining in Data Streams, Mining Graph Patterns, Big Data Frequent Pattern Mining, Algorithms for Data Clustering and more. Advanced-level students in computer science, researchers and practitioners from industry will find this book an invaluable reference.

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.

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.

With the immense amount of data that is now available online, security concerns have been an issue from the start, and have grown as new technologies are increasingly integrated in data collection, storage, and transmission. Online cyber threats, cyber terrorism, hacking, and other cybercrimes have begun to take advantage of this information that can be easily accessed if not properly handled. New privacy and security measures have been developed to address this cause for concern and have become an essential area of research within the past few years and into the foreseeable future. The ways in which data is secured and privatized should be discussed in terms of the technologies being used, the methods and models for security that have been

developed, and the ways in which risks can be detected, analyzed, and mitigated. The Research Anthology on Privatizing and Securing Data reveals the latest tools and technologies for privatizing and securing data across different technologies and industries. It takes a deeper dive into both risk detection and mitigation, including an analysis of cybercrimes and cyber threats, along with a sharper focus on the technologies and methods being actively implemented and utilized to secure data online. Highlighted topics include information governance and privacy, cybersecurity, data protection, challenges in big data, security threats, and more. This book is essential for data analysts, cybersecurity professionals, data scientists, security analysts, IT specialists, practitioners, researchers, academicians, and students interested in the latest trends and technologies for privatizing and securing data.

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.

Image Correlation for Shape, Motion and Deformation Measurements provides a comprehensive overview of data extraction through image analysis. Readers will find and in-depth look into various single- and multi-camera models (2D-DIC and 3D-DIC), two- and three-dimensional computer vision, and volumetric digital image correlation (VDIC). Fundamentals of accurate image matching are described, along with presentations of both new methods for quantitative error estimates in correlation-based motion measurements, and the effect of out-of-plane motion on 2D measurements. Thorough appendices offer descriptions of continuum mechanics formulations, methods for local surface strain estimation and non-linear optimization, as well as terminology in statistics and probability. With equal treatment of computer vision fundamentals and techniques for practical applications, this volume is both a reference for academic and industry-based researchers and engineers, as well as a valuable companion text for appropriate vision-based educational offerings.

Data MatchingConcepts and Techniques for Record Linkage, Entity Resolution, and Duplicate DetectionSpringer Science & Business Media

The big data era is upon us: data are being generated, analyzed, and used at an unprecedented scale, and data-driven decision making is sweeping through all aspects of society. Since the value of data explodes when it can be linked and fused with other data, addressing the big data integration (BDI) challenge is critical to realizing the promise of big data. BDI differs from traditional data integration along the dimensions of volume, velocity, variety, and veracity. First, not only can data sources contain a huge volume of data, but also the number of data sources is now in the millions. Second, because of the rate at which newly collected data are made available, many of the data sources are very dynamic, and the number of data sources is also rapidly exploding. Third, data sources are extremely heterogeneous in their structure and content, exhibiting considerable variety even for substantially similar entities. Fourth, the data sources are of widely differing qualities, with significant differences in the coverage, accuracy and timeliness of data provided. This book explores the progress that has been made by the data integration community on the topics of schema alignment, record linkage and data fusion in addressing these novel challenges faced by big data integration. Each of these topics is covered in a systematic way: first starting with a quick tour of the topic in the context of traditional data integration, followed by a detailed, example-driven exposition of recent innovative techniques that have been proposed to address the BDI challenges of volume, velocity, variety, and veracity. Finally, it presents merging topics and opportunities that are specific to BDI, identifying promising directions for the data integration community.

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

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