This volume offers state-of-the-art research in service science and its related research, education and practice areas. It showcases recent developments in smart service systems, operations management and analytics and their impact in complex service systems. The papers included in this volume highlight emerging technology and applications in fields including healthcare, energy, finance, information technology, transportation, sports, logistics, and public services. Regardless of size and service, a service organization is a service system. Because of the socio-technical nature of a service system, a systems approach must be adopted to design, develop, and deliver services, aimed at meeting end users' both utilitarian and socio-psychological needs. Effective understanding of service and service systems often requires combining multiple methods to consider how interactions of people, technology, organizations, and information create value under various conditions. The papers in this volume present methods to approach such technical challenges in service science and are based on top papers from the 2019 INFORMS International Conference on Service Science.

This two-volume set, LNAI 9651 and 9652, constitutes the thoroughly refereed proceedings of the 20th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, PAKDD 2016, held in Auckland, New Zealand, in April 2016. The 91 full papers were carefully reviewed and selected from 307 submissions. They are organized in topical sections named: classification; machine learning; applications; novel methods and algorithms; opinion mining and sentiment analysis; clustering; feature extraction and pattern mining; graph and network data; spatiotemporal and image data; anomaly detection and clustering; novel models and algorithms; and text mining and recommender systems.

This book addresses the problems that are encountered, and solutions that have been proposed, when we aim to identify people and to reconstruct populations under conditions where information is scarce, ambiguous, fuzzy and sometimes erroneous. The process from handwritten registers to a reconstructed digitized population consists of three major phases, reflected in the three main sections of this book. The first phase involves transcribing and digitizing the data while structuring the information in a meaningful and efficient way. In the second phase, records that refer to the same person or group of persons are identified by a process of linkage. In the third and final phase, the information on an individual is combined into a reconstruction of their life course. The studies and examples in this book originate from a range of countries, each with its own cultural and administrative characteristics, and from medieval charters through historical censuses and vital registration, to the modern issue of privacy preservation. Despite the diverse places and times addressed, they all share the study of fundamental issues when it comes to model reasoning for population reconstruction and the possibilities and limitations of information technology to support this process. It is thus not a single discipline that is involved in such an endeavor. Historians, social scientists, and linguists represent the humanities through their knowledge of the complexity of the past, the limitations of sources, and the possible interpretations of information. The availability of big data from digitized archives and the need for complex analyses to identify individuals calls for the involvement of computer scientists. With contributions from all these fields, often in direct cooperation, this book is at the heart of the digital humanities, and will hopefully offer a source of inspiration for future investigations.

This volume constitutes the refereed proceedings of the three workshops held at the 29th International Conference on Database and Expert Systems Applications.
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 organizing 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.

This two volume set LNCS 9642 and LNCS 9643 constitutes the refereed proceedings of the 21st International Conference on Database Systems for Advanced Applications, DASFAA 2016, held in Dallas, TX, USA, in April 2016. The 61 full papers presented were carefully reviewed and selected from a total of 183 submissions. The papers cover the following topics: crowdsourcing, data quality, entity identification, data mining and machine learning, recommendation, semantics computing and knowledge base, textual data, social networks, complex queries, similarity computing, graph databases, and miscellaneous, advanced applications. This book constitutes the refereed proceedings of the 10th International Conference on Machine Learning and Data Mining in Pattern Recognition, MLDM 2014, held in St. Petersburg, Russia in July 2014. The 40 full papers presented were carefully reviewed and selected from 128 submissions. The topics range from theoretical topics for classification, clustering, association rule and pattern mining to specific data mining methods for the different multimedia data types such as image mining, text mining, video mining and Web mining. This book constitutes the thoroughly refereed proceedings of the 20th East European Conference on Advances in Databases and Information Systems, ADBIS 2016, held in Prague, Czech Republic, in August 2016. The 21 full papers presented together with two keynote papers and one keynote abstract were carefully selected and reviewed from 85 submissions. The papers are organized in topical sections such as data quality, mining, analysis and clustering; model-driven engineering, conceptual modeling; data warehouse and multidimensional modeling, recommender systems; spatial and temporal data processing; distributed and parallel data processing; internet of things and sensor...
networks. This proceedings volume presents the latest on the theoretical approach of the contemporary issues evolved in strategic marketing and the integration of theory and practice. It highlights strategic research and innovative activities in marketing. The contributed chapters are concerned with using modern qualitative and quantitative techniques based on information technology used to manage and analyze business data, to discover hidden knowledge and to introduce intelligence into marketing processes. This allows for a focus on innovative applications in all aspects of marketing, of computerized technologies related to data analytics, predictive analytics and modeling, business intelligence and knowledge engineering, in order to demonstrate new ways of uncovering hidden knowledge and supporting marketing decisions with evidence-based intelligent tools. The chapters from the proceedings of the 5th International Conference on Strategic Innovative Marketing 2016 cover areas such as social media, marketing innovation, sustainable marketing, customer satisfaction strategies, customer relationship management, marketing research and analytics. The papers have been written by scientists, researchers, practitioners and students that demonstrate a special orientation in strategic marketing, all of whom aspire to be ahead of the curve based on the pillars of innovation. This proceedings volume shares their recent contributions to the field and showcases their exchange of insights on strategic issues in the science of innovation marketing.

This book highlights research in linking and mining data from across varied data sources. The authors focus on recent advances in this burgeoning field of multi-source data fusion, with an emphasis on exploratory and unsupervised data analysis, an area of increasing significance with the pace of growth of data vastly outpacing any chance of labeling them manually. The book looks at the underlying algorithms and technologies that facilitate the area within big data analytics, it covers their applications across domains such as smarter transportation, social media, fake news detection and enterprise search among others. This book enables readers to understand a spectrum of advances in this emerging area, and it will hopefully empower them to leverage and develop methods in multi-source data fusion and analytics with applications to a variety of scenarios. Includes advances on unsupervised, semi-supervised and supervised approaches to heterogeneous data linkage and fusion; Covers use cases of analytics over multi-view and heterogeneous data from across a variety of domains such as fake news, smarter transportation and social media, among others; Provides a high-level overview of advances in this emerging field and empowers the reader to explore novel applications and methodologies that would enrich the field.

This is the first monograph on the emerging area of linguistic linked data. Presenting a combination of background information on linguistic linked data and concrete implementation advice, it introduces and discusses the main benefits of applying linked data (LD) principles to the representation and publication of linguistic resources, arguing that LD does not look at a single resource in isolation but seeks to create a large network of resources that can be used together and uniformly, and so making more of the single resource. The book describes how the LD principles can be applied to modelling language resources. The first part provides the foundation for understanding the remainder of the book, introducing the data models, ontology and query languages used as
the basis of the Semantic Web and LD and offering a more detailed overview of the Linguistic Linked Data Cloud. The second part of the book focuses on modelling language resources using LD principles, describing how to model lexical resources using Ontolex-lemon, the lexicon model for ontologies, and how to annotate and address elements of text represented in RDF. It also demonstrates how to model annotations, and how to capture the metadata of language resources. Further, it includes a chapter on representing linguistic categories. In the third part of the book, the authors describe how language resources can be transformed into LD and how links can be inferred and added to the data to increase connectivity and linking between different datasets. They also discuss using LD resources for natural language processing. The last part describes concrete applications of the technologies: representing and linking multilingual wordnets, applications in digital humanities and the discovery of language resources. Given its scope, the book is relevant for researchers and graduate students interested in topics at the crossroads of natural language processing / computational linguistics and the Semantic Web / linked data. It appeals to Semantic Web experts who are not proficient in applying the Semantic Web and LD principles to linguistic data, as well as to computational linguists who are used to working with lexical and linguistic resources wanting to learn about a new paradigm for modelling, publishing and exploiting linguistic resources.

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.

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 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.

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.

Business intelligence initiatives have been dominating the technology priority list of many organizations. However, the lack of effective information quality and governance strategies and policies has been meeting these initiatives with some challenges. Information Quality and Governance for Business Intelligence presents the latest exchange of academic research on all aspects of practicing and managing information using a multidisciplinary approach that examines its quality for organizational growth. This book is an essential reference tool for researchers, practitioners, and university students specializing in business intelligence, information quality, and information systems.

Featuring a timely presentation of total survey error (TSE), this edited volume introduces valuable tools for understanding and improving survey data quality in the context of evolving large-scale data sets. This book provides an overview of the TSE framework and current TSE research as related to survey design, data collection, estimation, and analysis. It recognizes that survey data affects many public policy and business decisions and thus focuses on the framework for understanding and improving survey data quality. The book also addresses issues with data quality in official statistics and in social, opinion, and market research as these fields continue to evolve, leading to larger and messier data sets. This perspective challenges survey organizations to find ways to collect and process data more efficiently without sacrificing quality. The volume consists of the most up-to-date research and reporting from over 70 contributors representing the best academics and researchers from a range of fields. The chapters are broken out into five main sections: The Concept of TSE and the TSE Paradigm, Implications for Survey Design, Data Collection and Data Processing Applications, Evaluation and Improvement, and Estimation and Analysis. Each chapter introduces and examines multiple error sources, such as sampling error, measurement error, and nonresponse error, which often offer
the greatest risks to data quality, while also encouraging readers not to lose sight of the less commonly studied error sources, such as coverage error, processing error, and specification error. The book also notes the relationships between errors and the ways in which efforts to reduce one type can increase another, resulting in an estimate with larger total error. This book: • Features various error sources, and the complex relationships between them, in 25 high-quality chapters on the most up-to-date research in the field of TSE • Provides comprehensive reviews of the literature on error sources as well as data collection approaches and estimation methods to reduce their effects • Presents examples of recent international events that demonstrate the effects of data error, the importance of survey data quality, and the real-world issues that arise from these errors • Spans the four pillars of the total survey error paradigm (design, data collection, evaluation and analysis) to address key data quality issues in official statistics and survey research Total Survey Error in Practice is a reference for survey researchers and data scientists in research areas that include social science, public opinion, public policy, and business. It can also be used as a textbook or supplementary material for a graduate-level course in survey research methods.

This two-volume set constitutes the refereed proceedings of the workshops which complemented the 19th Joint European Conference on Machine Learning and Knowledge Discovery in Databases, ECML PKDD, held in Würzburg, Germany, in September 2019. The 70 full papers and 46 short papers presented in the two-volume set were carefully reviewed and selected from 200 submissions. The two volumes (CCIS 1167 and CCIS 1168) present the papers that have been accepted for the following workshops: Workshop on Automating Data Science, ADS 2019; Workshop on Advances in Interpretable Machine Learning and Artificial Intelligence and eXplainable Knowledge Discovery in Data Mining, AIMLAI-XKDD 2019; Workshop on Decentralized Machine Learning at the Edge, DMLE 2019; Workshop on Advances in Managing and Mining Large Evolving Graphs, LEG 2019; Workshop on Data and Machine Learning Advances with Multiple Views; Workshop on New Trends in Representation Learning with Knowledge Graphs; Workshop on Data Science for Social Good, SoGood 2019; Workshop on Knowledge Discovery and User Modelling for Smart Cities, UMCIT 2019; Workshop on Data Integration and Applications Workshop, DINA 2019; Workshop on Machine Learning for Cybersecurity, MLCS 2019; Workshop on Sports Analytics: Machine Learning and Data Mining for Sports Analytics, MLSA 2019; Workshop on Categorising Different Types of Online Harassment Languages in Social Media; Workshop on IoT Stream for Data Driven Predictive Maintenance, IoTStream 2019; Workshop on Machine Learning and Music, MML 2019; Workshop on Large-Scale Biomedical Semantic Indexing and Question Answering, BioASQ 2019. This book collects articles presented at the 13th International Conference on Information Technology- New Generations, April, 2016, in Las Vegas, NV USA. It includes over 100 chapters on critical areas of IT including Web Technology,
Communications, Security, and Data Mining. This book aims to provide an international forum for scholarly researchers, practitioners and academic communities to explore the role of information and communication technologies and its applications in technical and scholarly development. The conference attracted a total of 464 submissions, of which 152 submissions (including 4 poster papers) have been selected after a double-blind review process. Academic pioneering researchers, scientists, industrial engineers and students will find this series useful to gain insight into the current research and next-generation information science and communication technologies. This book discusses the aspects of communication, data science, ambient intelligence, networking, computing, security and Internet of things, from classical to intelligent scope. The authors hope that readers find the volume interesting and valuable; it gathers chapters addressing state-of-the-art intelligent methods and techniques for solving real-world problems along with a vision of the future research. Survey Methodology is becoming a more structured field of research, deserving of more and more academic attention. The SAGE Handbook of Survey Methodology explores both the increasingly scientific endeavour of surveys and their growing complexity, as different data collection modes and information sources are combined. The handbook takes a global approach, with a team of international experts looking at local and national specificities, as well as problems of cross-national, comparative survey research. The chapters are organized into seven major sections, each of which represents a stage in the survey life-cycle: Surveys and Societies Planning a Survey Measurement Sampling Data Collection Preparing Data for Use Assessing and Improving Data Quality The SAGE Handbook of Survey Methodology is a landmark and essential tool for any scholar within the social sciences. The 4 volume set LNCS 12112-12114 constitutes the papers of the 25th International Conference on Database Systems for Advanced Applications which will be held online in September 2020. The 119 full papers presented together with 19 short papers plus 15 demo papers and 4 industrial papers in this volume were carefully reviewed and selected from a total of 487 submissions. The conference program presents the state-of-the-art R&D activities in database systems and their applications. It provides a forum for technical presentations and discussions among database researchers, developers and users from academia, business and industry. Data warehouses consolidate various activities of a business and often form the backbone for generating reports that support important business decisions. Errors in data tend to creep in for a variety of reasons. Some of these reasons include errors during input data collection and errors while merging data collected independently across different databases. These errors in data warehouses often result in erroneous upstream reports, and could impact business decisions negatively. Therefore, one of the critical challenges while maintaining large data warehouses is that of ensuring
the quality of data in the data warehouse remains high. The process of maintaining high data quality is commonly referred to as data cleaning. In this book, we first discuss the goals of data cleaning. Often, the goals of data cleaning are not well defined and could mean different solutions in different scenarios. Toward clarifying these goals, we abstract out a common set of data cleaning tasks that often need to be addressed. This abstraction allows us to develop solutions for these common data cleaning tasks. We then discuss a few popular approaches for developing such solutions. In particular, we focus on an operator-centric approach for developing a data cleaning platform. The operator-centric approach involves the development of customizable operators that could be used as building blocks for developing common solutions. This is similar to the approach of relational algebra for query processing. The basic set of operators can be put together to build complex queries. Finally, we discuss the development of custom scripts which leverage the basic data cleaning operators along with relational operators to implement effective solutions for data cleaning tasks. This book constitutes the refereed proceedings of the 12th Extended Semantic Web Conference, ESWC 2014, held in Anissaras, Portoroz, Slovenia, in May/June 2015. The 43 revised full papers presented together with three invited talks were carefully reviewed and selected from 164 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 12 contributions, selected out of 16 submissions. The core tracks of the research conference were complemented with new tracks focusing on linking machine and human computation at web scale (cognition and Semantic Web, Human Computation and Crowdsourcing) beside the following subjects Vocabularies, Schemas, Ontologies, Reasoning, Linked Data, Semantic Web and Web Science, Semantic Data Management, Big data, Scalability, Natural Language Processing and Information Retrieval, Machine Learning, Mobile Web, Internet of Things and Semantic Streams, Services, Web APIs and the Web of Things, Cognition and Semantic Web, Human Computation and Crowdsourcing and In-Use Industrial Track as well. This book focuses on novel and state-of-the-art scientific work in the area of detection and prediction techniques using information found generally in graphs and particularly in social networks. Community detection techniques are presented in diverse contexts and for different applications while prediction methods for structured and unstructured data are applied to a variety of fields such as financial systems, security forums, and social networks. The rest of the book focuses on graph-based techniques for data analysis such as graph clustering and edge sampling. The research presented in this volume was selected based on solid reviews from the IEEE/ACM International Conference on Advances in Social Networks, Analysis, and Mining (ASONAM ‘17). Chapters were then improved and extended substantially, and the final versions were rigorously reviewed and revised to meet the series standards. This book will appeal to practitioners,
researchers and students in the field. This book constitutes the refereed proceedings of the 28th International Conference on Data Analytics and Management in Data Intensive Domains, DAMDID/RCDL 2016, held in Ershovo, Moscow, Russia, in October 2016. The 16 revised full papers presented together with one invited talk and two keynote papers were carefully reviewed and selected from 57 submissions. The papers are organized in topical sections on semantic modeling in data intensive domains; knowledge and learning management; text mining; data infrastructures in astrophysics; data analysis; research infrastructures; position paper. The BIRS Workshop “Advances in Interactive Knowledge Discovery and Data Mining in Complex and Big Data Sets” (15w2181), held in July 2015 in Banff, Canada, was dedicated to stimulating a cross-domain integrative machine-learning approach and appraisal of “hot topics” toward tackling the grand challenge of reaching a level of useful and useable computational intelligence with a focus on real-world problems, such as in the health domain. This encompasses learning from prior data, extracting and discovering knowledge, generalizing the results, fighting the curse of dimensionality, and ultimately disentangling the underlying explanatory factors in complex data, i.e., to make sense of data within the context of the application domain. The workshop aimed to contribute advancements in promising novel areas such as at the intersection of machine learning and topological data analysis. History has shown that most often the overlapping areas at intersections of seemingly disparate fields are key for the stimulation of new insights and further advances. This is particularly true for the extremely broad field of machine learning. The two-volume set LNAI 12084 and 12085 constitutes the thoroughly refereed proceedings of the 24th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2020, which was due to be held in Singapore, in May 2020. The conference was held virtually due to the COVID-19 pandemic. The 135 full papers presented were carefully reviewed and selected from 628 submissions. The papers present new ideas, original research results, and practical development experiences from all KDD related areas, including data mining, data warehousing, machine learning, artificial intelligence, databases, statistics, knowledge engineering, visualization, decision-making systems, and the emerging applications. They are organized in the following topical sections: recommender systems; classification; clustering; mining social networks; representation learning and embedding; mining behavioral data; deep learning; feature extraction and selection; human, domain, organizational and social factors in data mining; mining sequential data; mining imbalanced data; association; privacy and security; supervised learning; novel algorithms; mining multi-media/multi-dimensional data; application; mining graph and network data; anomaly detection and analytics; mining spatial, temporal, unstructured and semi-structured data; sentiment analysis; statistical/graphical model; multi-
source/distributed/parallel/cloud computing.

Data Matching

Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection

Springer Science & Business Media

This three-volume set, LNAI 10937, 10938, and 10939, constitutes the thoroughly refereed proceedings of the 22nd Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, PAKDD 2018, held in Melbourne, VIC, Australia, in June 2018. The 164 full papers were carefully reviewed and selected from 592 submissions. The volumes present papers focusing on new ideas, original research results and practical development experiences from all KDD related areas, including data mining, data warehousing, machine learning, artificial intelligence, databases, statistics, knowledge engineering, visualization, decision-making systems and the emerging applications.

This book presents a contemporary view of the role of information quality in information fusion and decision making, and provides a formal foundation and the implementation strategies required for dealing with insufficient information quality in building fusion systems for decision making. Information fusion is the process of gathering, processing, and combining large amounts of information from multiple and diverse sources, including physical sensors to human intelligence reports and social media. That data and information may be unreliable, of low fidelity, insufficient resolution, contradictory, fake and/or redundant. Sources may provide unverified reports obtained from other sources resulting in correlations and biases. The success of the fusion processing depends on how well knowledge produced by the processing chain represents reality, which in turn depends on how adequate data are, how good and adequate are the models used, and how accurate, appropriate or applicable prior and contextual knowledge is. By offering contributions by leading experts, this book provides an unparalleled understanding of the problem of information quality in information fusion and decision-making for researchers and professionals in the field.

This book constitutes the refereed conference proceedings of the 2nd International Workshop on Cryptocurrencies and Blockchain Technology, CBT 2018, and the 13th International Workshop on Data Privacy Management, DPM 2018, on conjunction with the 23nd European Symposium on Research in Computer Security, ESORICS 2018, held in Barcelona, Spain, in September 2018. From the CBT Workshop 7 full and 8 short papers out of 39 submissions are included. The selected papers cover aspects of identity management, smart contracts, soft- and hardforks, proof-of-works and proof of stake as well as on network layer aspects and the application of blockchain technology for secure connect event ticketing. The DPM Workshop received 36 submissions from which 11 full and 5 short papers were selected for presentation. The papers focus on challenging problems such as translation of high-level business goals into system level privacy policies, administration of sensitive identifiers, data integration and privacy engineering.
This, the 38th issue of Transactions on Large-Scale Data- and Knowledge-Centered Systems, contains extended and revised versions of six papers selected from the 68 contributions presented at the 27th International Conference on Database and Expert Systems Applications, DEXA 2016, held in Porto, Portugal, in September 2016. Topics covered include query personalization in databases, data anonymization, similarity search, computational methods for entity resolution, array-based computations in big data analysis, and pattern mining.

This book constitutes the refereed proceedings at PAKDD Workshops 2013, affiliated with the 17th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD) held in Gold Coast, Australia in April 2013. The 47 revised full papers presented were carefully reviewed and selected from 92 submissions. The workshops affiliated with PAKDD 2013 include: Data Mining Applications in Industry and Government (DMAapps), Data Analytics for Targeted Healthcare (DANTH), Quality Issues, Measures of Interestingness and Evaluation of Data Mining Models (QIMIE), Biologically Inspired Techniques for Data Mining (BDM), Constraint Discovery and Application (CDA), Cloud Service Discovery (CloudSD).

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.

This book constitutes the refereed proceedings of the 14th International Conference entitled Beyond Databases, Architectures and Structures, BDAS 2018, held in Poznań, Poland, in September 2018, during the IFIP World Computer Congress. It consists of 38 carefully reviewed papers selected from 102 submissions. The papers are organized in topical sections, namely big data and cloud computing; architectures, structures and algorithms for efficient data processing; artificial intelligence, data mining and knowledge discovery; text mining, natural language processing, ontologies and semantic web; image analysis and multimedia mining.

The two-volume set LNAI 8443 + LNAI 8444 constitutes the refereed proceedings of the 18th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2014, held in Tainan, Taiwan, in May 2014. The 40 full papers and the 60 short papers presented within these proceedings were carefully reviewed and selected from 371 submissions. They
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 EIMS 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. Resource Description Framework (RDF) is a graph-based data model used to publish data as a Web of Linked Data. RDF is an emergent foundation for large-scale data integration, the problem of providing a unified view over multiple data sources. An Entity Name System (ENS) is a thesaurus for entities, and is a crucial component in a data integration architecture. Populating a Linked Data ENS is equivalent to solving an Artificial Intelligence problem called instance matching, which concerns identifying pairs of entities referring to the same underlying entity. This publication presents an instance matcher with 4 properties, namely automation, heterogeneity, scalability and domain independence. Automation is addressed by employing inexpensive but well-performing heuristics to automatically generate a training set, which is employed by other machine learning algorithms in the pipeline. Data-driven alignment algorithms are adapted to deal with structural heterogeneity in RDF graphs. Domain independence is established by actively avoiding prior assumptions.
about input domains, and through evaluations on 10 RDF test cases. The full system is scaled by implementing it on cloud infrastructure using MapReduce algorithms. Resource Description Framework (RDF) is a graph-based data model used to publish data as a Web of Linked Data. RDF is an emergent foundation for large-scale data integration, the problem of providing a unified view over multiple data sources. An Entity Name System (ENS) is a thesaurus for entities, and is a crucial component in a data integration architecture. Populating a Linked Data ENS is equivalent to solving an Artificial Intelligence problem called instance matching, which concerns identifying pairs of entities referring to the same underlying entity. This publication presents an instance matcher with 4 properties, namely automation, heterogeneity, scalability and domain independence. Automation is addressed by employing inexpensive but well-performing heuristics to automatically generate a training set, which is employed by other machine learning algorithms in the pipeline. Data-driven alignment algorithms are adapted to deal with structural heterogeneity in RDF graphs. Domain independence is established by actively avoiding prior assumptions about input domains, and through evaluations on 10 RDF test cases. The full system is scaled by implementing it on cloud infrastructure using MapReduce algorithms.