1.

Record Nr.	UNINA9910838222203321
Autore	Ilyas Ihab F
Titolo	Data Cleaning
Pubbl/distr/stampa	San Rafael : , : Morgan & Claypool Publishers, , 2019 ©2019
ISBN	1-4503-7155-8
Descrizione fisica	1 online resource (284 pages)
Altri autori (Persone)	ChuXu
Disciplina	005.74
Soggetti	Data editing Database management Electronic data processing
Lingua di pubblicazione	Inglese
Formato	Materiale a stampa
Livello bibliografico	Monografia
Nota di contenuto	Intro -- Contents -- Preface -- Figure and Table Credits -- 1. Introduction -- 2. Outlier Detection -- 3. Data Deduplication -- 4. Data Transformation -- 5. Data Quality Rule Definition and Discovery -- 6. Rule-Based Data Cleaning -- 7. Machine Learning and Probabilistic Data Cleaning -- 8. Conclusion and Future Thoughts -- References -- Index -- Author Biographies -- Blank Page.
Sommario/riassunto	This is an overview of the end-to-end data cleaning process. Data quality is one of the most important problems in data management, since dirty data often leads to inaccurate data analytics results and incorrect business decisions. Poor data across businesses and the U.S. government are reported to cost trillions of dollars a year. Multiple surveys show that dirty data is the most common barrier faced by data scientists. Not surprisingly, developing effective and efficient data cleaning solutions is challenging and is rife with deep theoretical and engineering problems. This book is about data cleaning, which is used to refer to all kinds of tasks and activities to detect and repair errors in the data. Rather than focus on a particular data cleaning task, this book describes various error detection and repair methods, and attempts to anchor these proposals with multiple taxonomies and views. Specifically, it covers four of the most common and important data cleaning tasks, namely, outlier detection, data transformation, error

repair (including imputing missing values), and data deduplication. Furthermore, due to the increasing popularity and applicability of machine learning techniques, it includes a chapter that specifically explores how machine learning techniques are used for data cleaning, and how data cleaning is used to improve machine learning models. This book is intended to serve as a useful reference for researchers and practitioners who are interested in the area of data quality and data cleaning. It can also be used as a textbook for a graduate course. Although we aim at covering state-of-the-art algorithms and techniques, we recognize that data cleaning is still an active field of research and therefore provide future directions of research whenever appropriate.