# Mining of Massive Datasets

## Third Edition

The Web, social media, mobile activity, sensors, Internet commerce, and many other modern applications provide many extremely large datasets from which information can be gleaned by data mining. This book focuses on practical algorithms that have been used to solve key problems in data mining and can be used on even the largest datasets.

It begins with a discussion of the MapReduce framework and related techniques for efficient parallel programming. The tricks of locality-sensitive hashing are explained. This body of knowledge, which deserves to be more widely known, is essential when seeking similar objects in a very large collection without having to compare each pair of objects. Stream-processing algorithms for mining data that arrives too fast for exhaustive processing are also explained. The PageRank idea and related tricks for organizing the Web are covered next. Other chapters cover the problems of finding frequent itemsets and clustering, each from the point of view that the data is too large to fit in main memory. Two applications: recommendation systems and Web advertising, each vital in e-commerce, are treated in detail. Later chapters cover algorithms for analyzing social-network graphs, compressing large-scale data, and machine learning.

This third edition includes new and extended coverage on decision trees, deep learning, and mining social-network graphs. Written by leading authorities in database and Web technologies, it is essential reading for students and practitioners alike.

# Mining of Massive Datasets

## Third Edition

JURE LESKOVEC
Stanford University

ANAND RAJARAMAN
Rocketship VC

JEFFREY DAVID ULLMAN
Stanford University

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,
New Delhi – 110025, India

79 Anson Road, #06-04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University’s mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781108476348
DOI: 10.1017/9781108684163

First published 2012
Second edition 2014
Third edition 2020

Printed in the United Kingdom by TJ International Ltd. Padstow Cornwall

*A catalogue record for this publication is available from the British Library*

ISBN 978-1-108-47634-8 Hardback

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

# Contents

CAMBRIDGE

Cambridge University Press
978-1-108-47634-8 — Mining of Massive Datasets
Jure Leskovec , Anand Rajaraman , Jeffrey David Ullman
Frontmatter
More Information

# Preface

This book evolved from material developed over several years by Anand Rajaraman and Jeff Ullman for a one-quarter course at Stanford. The course CS345A, titled “Web Mining,” was designed as an advanced graduate course, although it has become accessible and interesting to advanced undergraduates. When Jure Leskovec joined the Stanford faculty, we reorganized the material considerably. He introduced a new course CS224W on network analysis and added material to CS345A, which was renumbered CS246. The three authors also introduced a large-scale data-mining project course, CS341. The book now contains material taught in all three courses.

## What the Book Is About

At the highest level of description, this book is about data mining. However, it focuses on data mining of very large amounts of data, that is, data so large it does not fit in main memory. Because of the emphasis on size, many of our examples are about the Web or data derived from the Web. Further, the book takes an algorithmic point of view: data mining is about applying algorithms to data, rather than using data to “train” a machine-learning engine of some sort. The principal topics covered are:

(1) Distributed file systems and map-reduce as a tool for creating parallel algorithms that succeed on very large amounts of data.
(2) Similarity search, including the key techniques of minhashing and locality-sensitive hashing.
(3) Data-stream processing and specialized algorithms for dealing with data that arrives so fast it must be processed immediately or lost.
(4) The technology of search engines, including Google’s PageRank, link-spam detection, and the hubs-and-authorities approach.
(5) Frequent-itemset mining, including association rules, market-baskets, the A-Priori Algorithm and its improvements.
(6) Algorithms for clustering very large, high-dimensional datasets.
(7) Two key problems for Web applications: managing advertising and recommendation systems.

(8) Algorithms for analyzing and mining the structure of very large graphs, especially social-network graphs.
(9) Techniques for obtaining the important properties of a large dataset by dimensionality reduction, including singular-value decomposition and latent semantic indexing.
(10) Machine-learning algorithms that can be applied to very large data, such as perceptrons, support-vector machines, gradient descent, decision trees, and neural nets.
(11) Neural nets and deep learning, including the most important special cases: convolutional and recurrent neural networks, and long short-term memory networks.

## Prerequisites

To appreciate fully the material in this book, we recommend the following prerequisites:

(1) An introduction to database systems, covering SQL and related programming systems.
(2) A sophomore-level course in data structures, algorithms, and discrete math.
(3) A sophomore-level course in software systems, software engineering, and programming languages.

## Exercises

The book contains extensive exercises, with some for almost every section. We indicate harder exercises or parts of exercises with an exclamation point. The hardest exercises have a double exclamation point.

## Support on the Web

Go to `http://www.mmds.org` for slides, homework assignments, project requirements, and exams from courses related to this book.

## Gradiance Automated Homework

There are automated exercises based on this book, using the Gradiance root-question technology, available at `www.gradiance.com/services`. Students may enter a public class by creating an account at that site and entering the class with code `1EDD8A1D`. Instructors may use the site by making an account there and then emailing `support at gradiance dot com` with their login name, the name of their school, and a request to use the MMDS materials.

## Acknowledgements

Cover art is by Scott Ullman.

We would like to thank Foto Afrati, Arun Marathe, and Rok Sosic for critical readings of a draft of this manuscript.

Errors were also reported by Rajiv Abraham, Ruslan Aduk, Apoorv Agarwal, Aris Anagnostopoulos, Yokila Arora, Stefanie Anna Baby, Atilla Soner Balkir, Arnaud Belletoile, Robin Bennett, Susan Biancani, Richard Boyd, Amitabh Chaudhary, Leland Chen, Hua Feng, Marcus Gemeinder, Anastasios Gounaris, Clark Grubb, Shrey Gupta, Waleed Hameid, Saman Haratizadeh, Julien Hoachuck, Przemyslaw Horban, Hsiu-Hsuan Huang, Jeff Hwang, Rafi Kamal, Lachlan Kang, Ed Knorr, Haewoon Kwak, Ellis Lau, Greg Lee, David Z. Liu, Ethan Lozano, Yunan Luo, Michael Mahoney, Sergio Matos, Justin Meyer, Bryant Moscon, Brad Penoff, John Phillips, Philips Kokoh Prasetyo, Qi Ge, Harizo Rajaona, Timon Ruban, Rich Seiter, Hitesh Shetty, Angad Singh, Sandeep Sripada, Dennis Sidharta, Krzysztof Stencel, Mark Storus, Roshan Sumbaly, Zack Taylor, Tim Triche Jr., Wang Bin, Weng Zhen-Bin, Robert West, Steven Euijong Whang, Oscar Wu, Xie Ke, Christopher T.-R. Yeh, Nicolas Zhao, and Zhou Jingbo, The remaining errors are ours, of course.

J. L.
A. R.
J. D. U.
Palo Alto, CA
July, 2019