

## 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.

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

---

# 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  
978-1-108-47634-8 — Mining of Massive Datasets  
Jure Leskovec , Anand Rajaraman , Jeffrey David Ullman  
Frontmatter  
[More Information](#)

---

## 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](http://www.cambridge.org)

Information on this title: [www.cambridge.org/9781108476348](http://www.cambridge.org/9781108476348)

DOI: 10.1017/9781108684163

First edition © A. Rajaraman and J. D. Ullman 2012

Second and Third editions © J. Leskovec, A. Rajaraman, and J. D. Ullman 2014, 2020

This publication is in copyright. Subject to statutory exception  
and to the provisions of relevant collective licensing agreements,  
no reproduction of any part may take place without the written  
permission of Cambridge University Press.

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

	<i>Preface</i>	<i>page ix</i>
1	<b>Data Mining</b>	1
	1.1 What is Data Mining?	1
	1.2 Statistical Limits on Data Mining	5
	1.3 Things Useful to Know	8
	1.4 Outline of the Book	16
	1.5 Summary of Chapter 1	18
	1.6 References for Chapter 1	19
2	<b>MapReduce and the New Software Stack</b>	20
	2.1 Distributed File Systems	21
	2.2 MapReduce	23
	2.3 Algorithms Using MapReduce	29
	2.4 Extensions to MapReduce	40
	2.5 The Communication-Cost Model	52
	2.6 Complexity Theory for MapReduce	58
	2.7 Summary of Chapter 2	72
	2.8 References for Chapter 2	74
3	<b>Finding Similar Items</b>	78
	3.1 Applications of Set Similarity	79
	3.2 Shingling of Documents	83
	3.3 Similarity-Preserving Summaries of Sets	86
	3.4 Locality-Sensitive Hashing for Documents	96
	3.5 Distance Measures	101
	3.6 The Theory of Locality-Sensitive Functions	107
	3.7 LSH Families for Other Distance Measures	112
	3.8 Applications of Locality-Sensitive Hashing	118
	3.9 Methods for High Degrees of Similarity	125
	3.10 Summary of Chapter 3	133
	3.11 References for Chapter 3	136

<b>4</b>	<b>Mining Data Streams</b>	138
4.1	The Stream Data Model	138
4.2	Sampling Data in a Stream	142
4.3	Filtering Streams	145
4.4	Counting Distinct Elements in a Stream	148
4.5	Estimating Moments	151
4.6	Counting Ones in a Window	157
4.7	Decaying Windows	163
4.8	Summary of Chapter 4	165
4.9	References for Chapter 4	167
<b>5</b>	<b>Link Analysis</b>	169
5.1	PageRank	169
5.2	Efficient Computation of PageRank	183
5.3	Topic-Sensitive PageRank	189
5.4	Link Spam	193
5.5	Hubs and Authorities	197
5.6	Summary of Chapter 5	202
5.7	References for Chapter 5	205
<b>6</b>	<b>Frequent Itemsets</b>	206
6.1	The Market-Basket Model	206
6.2	Market Baskets and the A-Priori Algorithm	213
6.3	Handling Larger Datasets in Main Memory	222
6.4	Limited-Pass Algorithms	229
6.5	Counting Frequent Items in a Stream	235
6.6	Summary of Chapter 6	239
6.7	References for Chapter 6	241
<b>7</b>	<b>Clustering</b>	243
7.1	Introduction to Clustering Techniques	243
7.2	Hierarchical Clustering	247
7.3	K-means Algorithms	256
7.4	The CURE Algorithm	264
7.5	Clustering in Non-Euclidean Spaces	267
7.6	Clustering for Streams and Parallelism	271
7.7	Summary of Chapter 7	277
7.8	References for Chapter 7	280
<b>8</b>	<b>Advertising on the Web</b>	282
8.1	Issues in On-Line Advertising	282
8.2	On-Line Algorithms	285
8.3	The Matching Problem	288
8.4	The Adwords Problem	291

8.5	Adwords Implementation	300
8.6	Summary of Chapter 8	304
8.7	References for Chapter 8	305
<b>9</b>	<b>Recommendation Systems</b>	<b>307</b>
9.1	A Model for Recommendation Systems	307
9.2	Content-Based Recommendations	311
9.3	Collaborative Filtering	321
9.4	Dimensionality Reduction	327
9.5	The Netflix Challenge	336
9.6	Summary of Chapter 9	337
9.7	References for Chapter 9	338
<b>10</b>	<b>Mining Social-Network Graphs</b>	<b>340</b>
10.1	Social Networks as Graphs	340
10.2	Clustering of Social-Network Graphs	345
10.3	Direct Discovery of Communities	353
10.4	Partitioning of Graphs	358
10.5	Finding Overlapping Communities	365
10.6	Simrank	373
10.7	Counting Triangles	382
10.8	Neighborhood Properties of Graphs	388
10.9	Summary of Chapter 10	404
10.10	References for Chapter 10	408
<b>11</b>	<b>Dimensionality Reduction</b>	<b>410</b>
11.1	Eigenvalues and Eigenvectors of Symmetric Matrices	410
11.2	Principal-Component Analysis	417
11.3	Singular-Value Decomposition	423
11.4	CUR Decomposition	432
11.5	Summary of Chapter 11	438
11.6	References for Chapter 11	440
<b>12</b>	<b>Large-Scale Machine Learning</b>	<b>441</b>
12.1	The Machine-Learning Model	442
12.2	Perceptrons	449
12.3	Support-Vector Machines	462
12.4	Learning from Nearest Neighbors	474
12.5	Decision Trees	482
12.6	Comparison of Learning Methods	493
12.7	Summary of Chapter 12	494
12.8	References for Chapter 12	496

viii      **Contents**

---

<b>13</b>	<b>Neural Nets and Deep Learning</b>	498
13.1	Introduction to Neural Nets	498
13.2	Dense Feedforward Networks	504
13.3	Backpropagation and Gradient Descent	514
13.4	Convolutional Neural Networks	522
13.5	Recurrent Neural Networks	531
13.6	Regularization	538
13.7	Summary of Chapter 13	541
13.8	References for Chapter 13	542
	<i>Index</i>	544



## 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](http://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 Sasic 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 Rajana, 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

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

---