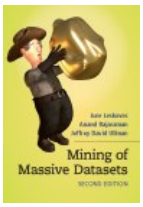
Mining of Massive Datasets

Jure Leskovec, Anand Rajaraman, Jeff Ullman

Big-data is transforming the world. Here you will learn data mining and machine learning techniques to process large datasets and extract valuable knowledge from them.

The book

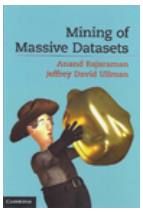


The book is based on Stanford Computer Science course CS246: Mining Massive Datasets (and CS345A: Data Mining).

The book, like the course, is designed at the undergraduate computer science level with no formal prerequisites. To support deeper explorations, most of the chapters are supplemented with further reading references.

The Mining of Massive Datasets book has been published by Cambridge University Press. You can get a 20% discount by applying the code MMDS20 at checkout.

By agreement with the publisher, you can download the book for free from this page. Cambridge University Press does, however, retain copyright on the work, and we expect that you will obtain their permission and acknowledge our authorship if you republish parts or all of it.



We welcome your feedback on the manuscript.

The MOOC (Massive Open Online Course)

We are running the third edition of an online course based on the Mining Massive Datases book:

Mining Massive Datasets MOOC

The course starts September 12 2015 and will run for 9 weeks with 7 weeks of lectures. Additional information and registration.

The 3rd edition of the book (v3.0 beta)

We are developing the third edition of the book.
You can see the current state of the new edition, along with a description of the changes so far here.

The 2nd edition of the book (v2.1)

The following is the second edition of the book. There are three new chapters, on mining large graphs, dimensionality reduction, and machine learning. There is also a revised Chapter 2 that treats map-reduce programming in a manner closer to how it is used in practice.

Together with each chapter there is aslo a set of lecture slides that we use for teaching Stanford CS246: Mining Massive Datasets course. Note that the slides do not necessarily cover all the material convered in the corresponding chapters.

Chapter	Title	Book	Slides	Videos
	Preface and Table of Contents	PDF		
Chapter 1	Data Mining	PDF	PDF PPT	
Chapter 2	Map-Reduce and the New Software Stack	PDF	PDF PPT	1 2 3 4 5 6 7 8
Chapter 3	Finding Similar Items	PDF	PDF PPT	1 2 3 4 5 6 7 8 9 10 11 12 13
Chapter 4	Mining Data Streams	PDF	Part 1: PDF PPT Part 2: PDF PPT	1 2 3 4 5
Chapter 5	Link Analysis	PDF	Part 1: PDF PPT Part 2: PDF PPT	1 2 3 4 5 6 7 8 9 10 11 12 13 14
Chapter 6	Frequent Itemsets	PDF	PDF PPT	1 2 3 4
Chapter 7	Clustering	PDF	PDF PPT	1 2 3 4 5
Chapter 8	Advertising on the Web	PDF	PDF PPT	1 2 3 4
Chapter 9	Recommendation Systems	PDF	Part 1: PDF PPT Part 2: PDF PPT	1 2 3 4 5
Chapter 10	Mining Social-Network Graphs	PDF	Part 1: PDF PPT PDF PPT	1 2 3 4 5 6 7 8 9 10 11 12

			Part 2:			
Chapter 11	Dimensionality Reduction	PDF		PDF	PPT	1 2 3 4 5 6 7 8 9 10 11 12
Chapter 12	Large-Scale Machine Learning	PDF	Part 1: Part 2:	PDF PDF	PPT PPT	1 2 3 4 5 6 7 8 9 10 11 12
	Index	PDF				
	Errata	HTML				

Download the latest version of the book as a single big PDF file (511 pages, 3 MB).

Download the full version of the book with a hyper-linked table of contents that make it easy to jump around: PDF file (513 pages, 3.69 MB).

The Errata for the second edition of the book: HTML.

Download slides (PPT) in French: Chapter 4, Chapter 5, Chapter 8, Chapter 9, Chapter 10. Courtesy of Richard Khoury.

Note to the users of provided slides: We would be delighted if you found this our material useful in giving your own lectures. Feel free to use these slides verbatim, or to modify them to fit your own needs. PowerPoint originals are available. If you make use of a significant portion of these slides in your own lecture, please include this message, or a link to our web site: http://www.mmds.org/.

Comments and corrections are most welcome. Please let us know if you are using these materials in your course and we will list and link to your course.

Stanford big data courses

CS246

CS246: Mining Massive Datasets is graduate level course that discusses data mining and machine learning algorithms for analyzing very large amounts of data. The emphasis is on Map Reduce as a tool for creating parallel algorithms that can process very large amounts of data.

CS341

CS341 Project in Mining Massive Data Sets is an advanced project based course. Students work on data mining and machine learning algorithms for analyzing very large amounts of data. Both interesting big datasets as well as computational infrastructure (large MapReduce cluster) are provided by course staff. Generally, students first take CS246 followed by CS341.

CS341 is generously supported by Amazon by giving us access to their EC2 platform.

CS224W

CS224W: Social and Information Networks is graduate level course that covers recent research on the structure and analysis of such large social and information networks and on models and algorithms that abstract their basic properties. Class explores how to practically analyze large scale network data and how to reason about it through models for network structure and evolution.

You can take Stanford courses!

If you are not a Stanford student, you can still take CS246 as well as CS224W or earn a Stanford Mining Massive Datasets graduate certificate by completing a sequence of four Stanford Computer Science courses. A graduate certificate is a great way to keep the skills and knowledge in your field current. More information is available at the Stanford Center for Professional Development (SCPD).

Supporting materials

If you are an instructor interested in using the Gradiance Automated Homework System with this book, start by creating an account for yourself here. Then, email your chosen login and the request to become an instructor for the MMDS book to support@gradiance.com. You will then be able to create a class using these materials. Manuals explaining the use of the system are available here.

Students who want to use the Gradiance Automated Homework System for self-study can register here. Then, use the class token `1EDD8A1D` to join the "omnibus class" for the MMDS book. See The Student Guide for more information.

Previous versions of the book

Version 1.0

The following materials are equivalent to the published book, with errata corrected to July 4, 2012.

Chapter	Title	Book
	Preface and Table of Contents	PDF
Chapter 1	Data Mining	PDF
Chapter 2	Large-Scale File Systems and Map-Reduce	PDF
Chapter 3	Finding Similar Items	PDF
Chapter 4	Mining Data Streams	PDF
Chapter 5	Link Analysis	PDF
Chapter 6	Frequent Itemsets	PDF
Chapter 7	Clustering	PDF
Chapter 8	Advertising on the Web	PDF

Chapter 9	Recommendation Systems	PDF
	Index	PDF
	Errata	HTML

Download the book as published here (340 pages, 2 MB).