

## Foundations of Data Science

This book provides an introduction to the mathematical and algorithmic foundations of data science, including machine learning, high-dimensional geometry, and analysis of large networks. Topics include the counterintuitive nature of data in high dimensions, important linear algebraic techniques such as singular value decomposition, the theory of random walks and Markov chains, the fundamentals of and important algorithms for machine learning, algorithms, and analysis for clustering, probabilistic models for large networks, representation learning including topic modeling and nonnegative matrix factorization, wavelets, and compressed sensing. Important probabilistic techniques are developed including the law of large numbers, tail inequalities, analysis of random projections, generalization guarantees in machine learning, and moment methods for analysis of phase transitions in large random graphs. Additionally, important structural and complexity measures are discussed such as matrix norms and VC-dimension. This book is suitable for both undergraduate and graduate courses in the design and analysis of algorithms for data.

Avrim Blum is Chief Academic Officer at the Toyota Technological Institute at Chicago and formerly Professor at Carnegie Mellon University. He has over 25,000 citations for his work in algorithms and machine learning. He has received the AI Journal Classic Paper Award, ICML/COLT 10-Year Best Paper Award, Sloan Fellowship, NSF NYI award, and Herb Simon Teaching Award, and is a fellow of the Association for Computing Machinery.

John Hopcroft is the IBM Professor of Engineering and Applied Mathematics at Cornell University. He is a member National Academy of Sciences and National Academy of Engineering, and a foreign member of the Chinese Academy of Sciences. He received the Turing Award in 1986, was appointed to the National Science Board in 1992 by President George H. W. Bush, and was presented with the Friendship Award by Premier Li Keqiang for his work in China.

Ravindran (Ravi) Kannan is Principal Researcher for Microsoft Research, India. He was the recipient of the Fulkerson Prize in Discrete Mathematics (1991) and the Knuth Prize (ACM) in 2011. He is a distinguished alumnus of Indian Institute of Technology, Bombay, and his past faculty appointments include Massachusetts Institute of Technology, Carnegie-Mellon University, Yale University, and the Indian Institute of Science.

Cambridge University Press  
978-1-108-48506-7 — Foundations of Data Science  
Avrim Blum , John Hoperoft , Ravi Kannan  
Frontmatter  
[More Information](#)

---

# Foundations of Data Science

**Avrim Blum**

*Toyota Technological Institute at Chicago*

**John Hopcroft**

*Cornell University, New York*

**Ravindran Kannan**

*Microsoft Research, India*



**CAMBRIDGE**  
UNIVERSITY PRESS

Cambridge University Press  
978-1-108-48506-7 — Foundations of Data Science  
Avrim Blum, John Hopcroft, Ravi Kannan  
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/9781108485067](http://www.cambridge.org/9781108485067)

DOI: 10.1017/9781108755528

© Avrim Blum, John Hopcroft, and Ravindran Kannan 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 2020

Printed in the United Kingdom by TJ International, Padstow Cornwall

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

*Library of Congress Cataloging-in-Publication Data*

Names: Blum, Avrim, 1966– author. | Hopcroft, John E., 1939– author. | Kannan, Ravindran, 1953– author.

Title: Foundations of data science / Avrim Blum, Toyota Technological Institute at Chicago, John Hopcroft, Cornell University, New York, Ravindran Kannan, Microsoft Research, India.

Description: First edition. | New York, NY : Cambridge University Press, 2020. |

Includes bibliographical references and index.

Identifiers: LCCN 2019038133 (print) | LCCN 2019038134 (ebook) |

ISBN 9781108485067 (hardback) | ISBN 9781108755528 (epub)

Subjects: LCSH: Computer science. | Statistics. | Quantitative research.

Classification: LCC QA76 .B5675 2020 (print) | LCC QA76 (ebook) | DDC 004–dc23

LC record available at <https://lcn.loc.gov/2019038133>

LC ebook record available at <https://lcn.loc.gov/2019038134>

ISBN 978-1-108-48506-7 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

---

<b>1</b>	<b>Introduction</b>	<i>page</i> 1
<b>2</b>	<b>High-Dimensional Space</b>	4
2.1	Introduction	4
2.2	The Law of Large Numbers	4
2.3	The Geometry of High Dimensions	8
2.4	Properties of the Unit Ball	8
2.5	Generating Points Uniformly at Random from a Ball	13
2.6	Gaussians in High Dimension	15
2.7	Random Projection and Johnson-Lindenstrauss Lemma	16
2.8	Separating Gaussians	18
2.9	Fitting a Spherical Gaussian to Data	20
2.10	Bibliographic Notes	21
2.11	Exercises	22
<b>3</b>	<b>Best-Fit Subspaces and Singular Value Decomposition (SVD)</b>	29
3.1	Introduction	29
3.2	Preliminaries	31
3.3	Singular Vectors	31
3.4	Singular Value Decomposition (SVD)	34
3.5	Best Rank- $k$ Approximations	36
3.6	Left Singular Vectors	37
3.7	Power Method for Singular Value Decomposition	39
3.8	Singular Vectors and Eigenvectors	42
3.9	Applications of Singular Value Decomposition	42
3.10	Bibliographic Notes	53
3.11	Exercises	54
<b>4</b>	<b>Random Walks and Markov Chains</b>	62
4.1	Stationary Distribution	65
4.2	Markov Chain Monte Carlo	67
4.3	Areas and Volumes	71
4.4	Convergence of Random Walks on Undirected Graphs	73
4.5	Electrical Networks and Random Walks	81
4.6	Random Walks on Undirected Graphs with Unit Edge Weights	85

## CONTENTS

4.7	Random Walks in Euclidean Space	92
4.8	The Web as a Markov Chain	95
4.9	Bibliographic Notes	98
4.10	Exercises	99
<b>5</b>	<b>Machine Learning</b>	<b>109</b>
5.1	Introduction	109
5.2	The Perceptron Algorithm	110
5.3	Kernel Functions and Nonlinearly Separable Data	111
5.4	Generalizing to New Data	113
5.5	VC-Dimension	118
5.6	VC-Dimension and Machine Learning	126
5.7	Other Measures of Complexity	127
5.8	Deep Learning	128
5.9	Gradient Descent	134
5.10	Online Learning	138
5.11	Boosting	145
5.12	Further Current Directions	148
5.13	Bibliographic Notes	152
5.14	Exercises	152
<b>6</b>	<b>Algorithms for Massive Data Problems: Streaming, Sketching, and Sampling</b>	<b>159</b>
6.1	Introduction	159
6.2	Frequency Moments of Data Streams	160
6.3	Matrix Algorithms Using Sampling	169
6.4	Sketches of Documents	177
6.5	Bibliographic Notes	178
6.6	Exercises	179
<b>7</b>	<b>Clustering</b>	<b>182</b>
7.1	Introduction	182
7.2	$k$ -Means Clustering	185
7.3	$k$ -Center Clustering	189
7.4	Finding Low-Error Clusterings	189
7.5	Spectral Clustering	190
7.6	Approximation Stability	197
7.7	High-Density Clusters	199
7.8	Kernel Methods	201
7.9	Recursive Clustering Based on Sparse Cuts	202
7.10	Dense Submatrices and Communities	202
7.11	Community Finding and Graph Partitioning	205
7.12	Spectral Clustering Applied to Social Networks	208
7.13	Bibliographic Notes	210
7.14	Exercises	210
<b>8</b>	<b>Random Graphs</b>	<b>215</b>
8.1	The $G(n, p)$ Model	215

## CONTENTS

8.2	Phase Transitions	222
8.3	Giant Component	232
8.4	Cycles and Full Connectivity	235
8.5	Phase Transitions for Increasing Properties	239
8.6	Branching Processes	241
8.7	CNF-SAT	246
8.8	Nonuniform Models of Random Graphs	252
8.9	Growth Models	254
8.10	Small-World Graphs	261
8.11	Bibliographic Notes	266
8.12	Exercises	266
<b>9</b>	<b>Topic Models, Nonnegative Matrix Factorization, Hidden Markov Models, and Graphical Models</b>	<b>274</b>
9.1	Topic Models	274
9.2	An Idealized Model	277
9.3	Nonnegative Matrix Factorization	279
9.4	NMF with Anchor Terms	281
9.5	Hard and Soft Clustering	282
9.6	The Latent Dirichlet Allocation Model for Topic Modeling	283
9.7	The Dominant Admixture Model	285
9.8	Formal Assumptions	287
9.9	Finding the Term-Topic Matrix	290
9.10	Hidden Markov Models	295
9.11	Graphical Models and Belief Propagation	298
9.12	Bayesian or Belief Networks	299
9.13	Markov Random Fields	300
9.14	Factor Graphs	301
9.15	Tree Algorithms	301
9.16	Message Passing in General Graphs	303
9.17	Warning Propagation	310
9.18	Correlation between Variables	311
9.19	Bibliographic Notes	315
9.20	Exercises	315
<b>10</b>	<b>Other Topics</b>	<b>318</b>
10.1	Ranking and Social Choice	318
10.2	Compressed Sensing and Sparse Vectors	322
10.3	Applications	325
10.4	An Uncertainty Principle	327
10.5	Gradient	330
10.6	Linear Programming	332
10.7	Integer Optimization	334
10.8	Semi-Definite Programming	334
10.9	Bibliographic Notes	336
10.10	Exercises	337

## CONTENTS

<b>11 Wavelets</b>	341
11.1 Dilation	341
11.2 The Haar Wavelet	342
11.3 Wavelet Systems	345
11.4 Solving the Dilation Equation	346
11.5 Conditions on the Dilation Equation	347
11.6 Derivation of the Wavelets from the Scaling Function	350
11.7 Sufficient Conditions for the Wavelets to Be Orthogonal	353
11.8 Expressing a Function in Terms of Wavelets	355
11.9 Designing a Wavelet System	356
11.10 Applications	357
11.11 Bibliographic Notes	357
11.12 Exercises	357
<b>12 Background Material</b>	360
12.1 Definitions and Notation	360
12.2 Useful Relations	361
12.3 Useful Inequalities	365
12.4 Probability	372
12.5 Bounds on Tail Probability	380
12.6 Applications of the Tail Bound	386
12.7 Eigenvalues and Eigenvectors	387
12.8 Generating Functions	400
12.9 Miscellaneous	404
12.10 Exercises	407
<i>References</i>	411
<i>Index</i>	421