Wiley Series on Methods and Applications in Data Mining

DATA MINING AND LEARNING ANALYTICS

Applications in Educational Research

Edited by
Samira ElAtia
Donald Ipperciel
Osmar R. Zaïane

WILEY

DATA MINING AND LEARNING ANALYTICS

WILEY SERIES ON METHODS AND APPLICATIONS IN DATA MINING

Series Editor: Daniel T. Larose

- **Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition**
  Daniel T. Larose and Chantal D. Larose
- **Data Mining for Genomics and Proteomics: Analysis of Gene and Protein Expression**
  Darius M. Dziuda
- **Knowledge Discovery with Support Vector Machines**
  Lutz Hamel
- **Data Mining on the Web: Uncovering Patterns in Web Content, Structure, and Usage**
  Zdravko Markov and Daniel T. Larose
- **Data Mining Methods and Models**
  Daniel T. Larose
- **Practical Text Mining with Perl**
  Roger Bilisoly
- **Data Mining and Predictive Analytics**
  Daniel T. Larose and Chantal D. Larose

DATA MINING AND LEARNING ANALYTICS

Applications in Educational Research

Edited by

SAMIRA ELATIA
DONALD IPPERCIEL
OSMAR R. ZAÏANE

WILEY

Published by John Wiley & Sons, Inc., Hoboken, New Jersey

Published simultaneously in Canada

Library of Congress Cataloguing-in-Publication Data

Names: ElAtia, Samira, 1973– editor. | Ipperciel, Donald, 1967– editor. | Zaiane, Osmar R., 1965– editor.
Title: Data mining and learning analytics : applications in educational research / edited by Samira ElAtia, Donald Ipperciel, Osmar R. Zaiane.
Description: Hoboken, New Jersey : John Wiley & Sons, Inc., [2016] | Includes bibliographical references and index.
Identifiers: LCCN 2016016549| ISBN 9781118998236 (cloth) | ISBN 9781118998212 (epub)
Subjects: LCSH: Education–Research–Statistical methods. | Educational statistics–Data processing. | Data mining.
Classification: LCC LB1028.43 .D385 2016 | DDC 370.72/7–dc23
LC record available at https://lccn.loc.gov/2016016549

Set in 10/12pt Times by SPi Global, Pondicherry, India

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CONTENTS

NOTES ON CONTRIBUTORS

Vincent Aleven, an associate professor in the Human–Computer Interaction Institute at Carnegie Mellon University, has 20 years of experience in research and development of educational software based on cognitive theory and self-regulated learning theory, with a focus on K–12 mathematics. He has created effective nonprogrammer authoring tools for intelligent tutoring systems (http://ctat.pact.cs.cmu.edu). He and his colleagues and students have created tutors that support self-regulated learning and collaborative learning and have even won seven best paper awards at international conferences. He has over 200 publications to his name and is the coeditor in chief of the *International Journal of Artificial Intelligence in Education*. He has been a PI on 8 major research grants and co-PI on 10 others.

Dennis Alonzo is a lecturer and applied statistician. He has been involved in various international and national research projects in a broad range of topics including student IT experiences, blended and online learning, and assessment. Also, he has received various scholarships from the Australian, Korean, and Philippine governments.

Alison L. Bailey is a professor of human development and psychology at the University of California, Los Angeles, focusing on the interdisciplinary development of language learning progressions for use in instruction and assessment with school-age students. Her most recent book is *Children's Multilingual Development and Education: Fostering Linguistic Resources in Home and School Contexts* (Cambridge University Press). She is also a faculty research partner at the National Center for Research on Evaluation, Standards, and Student Testing. She serves on the technical advisory boards of several states and consortia developing next-generation English language proficiency assessment systems.

Ryan S. Baker is an associate professor of cognitive studies and program coordinator for learning analytics at Teachers College, Columbia University. He earned his Ph.D. in human–computer interaction from Carnegie Mellon University. He was previously an assistant professor of psychology and learning sciences at Worcester Polytechnic Institute and served as the first technical director of the Pittsburgh Science of Learning Center DataShop, the largest public repository for data on the interaction between learners and educational software. He was the founding president of the International Educational Data Mining Society and is an associate editor of the *Journal of Educational Data Mining* and the *International Journal of Artificial Intelligence in Education*.

Tiffany Barnes is an associate professor of computer science at NC State University where she received her Ph.D. in 2003. She received an NSF CAREER Award for her novel work in using data to add intelligence to STEM learning environments. She is also a co-PI on the NSF STARS Computing Corps grants that engage college students in outreach, research, and service to broaden participation in computing. She researches effective ways to build serious games, promote undergraduate research, and develop new ways to teach computing. Dr. Barnes serves on the ACM SIGCSE, AIED, and IEDMS boards and has been on the organizing committees for several conferences, including Educational Data Mining and Foundations of Digital Games.

Bettina Berendt is a professor of computer science in the Declarative Languages and Artificial Intelligence group at KU Leuven, Belgium. Her research interests include web, text, and social and semantic mining, privacy and antidiscrimination and how data mining can contribute to this, teaching of and for privacy, and critical data science for computer scientists, digital humanists, and others. More information about Bettina Berendt can be found at http://people.cs.kuleuven.be/~bettina.berendt.

Yoav Bergner is a research scientist in the Computational Psychometrics Research Center at Educational Testing Service. He received his Ph.D. degree in theoretical physics from the Massachusetts Institute of Technology and B.A. degree in physics from Harvard University. His research combines methods from psychometrics and data mining with applications to data from collaborative problem-solving assessment, educational games, simulations, tutors, and MOOCs.

Anne Blackstock-Bernstein is a doctoral student in human development and psychology at the University of California, Los Angeles. As part of her work on the Dynamic Language Learning Progression Project, she has studied children's language and gesture use in the context of mathematics. She is interested in language assessment and oral language development during early childhood, particularly among English language learners. Prior to receiving her Master of Arts in Education from UCLA, she worked in preschool classrooms in Massachusetts and as a research assistant at Weill Cornell Medical College in New York City.

Alejandro Bogarín is an employee of Data and Statistics Section at the University of Córdoba in Spain and a member of the ADIR Research Group. At present, he is finishing his Ph.D. degree in computer science at the University of Córdoba, Spain. His research interests lie in applying educational process mining (EPM) techniques to extract knowledge from event logs recorded by an information system.

Dion Brocks is a professor and associate dean of undergraduate affairs at the Faculty of Pharmacy and Pharmaceutical Sciences at the University of Alberta. He has published over 110 peer-reviewed papers mostly in the area of pharmacokinetics. His more recent research interest besides that outlined in his chapter is related to pharmacokinetic changes in obesity. As part of his associate dean duties, he is in charge of the process for students desiring admission into the program, something he has been doing since 2003.

Rebecca Brown is a doctoral student in computer science at NC State University. Her research is focused on student interaction in online courses.

Meaghan Brugha completed her M.Ed. in Educational Administration and Comparative, International and Development Education at OISE, University of Toronto. Focusing her research on educational technology platforms such as MOOCs, she is fascinated by how educational innovation can act as a catalyst for a more equitable and accessible education for all.

Michael J. Cennamo is a doctoral student and instructor at Teachers College, Columbia University, studying instructional technology and media. His research is focused on "blended learning"; his passion lies in helping faculty find the perfect mix of online and face-to-face instruction for their particular classroom and teaching style. He has also worked at Columbia as an instructional technologist since 2008, first at the Columbia Center for New Media Teaching and Learning (CCNMTL) and currently at the School of Professional Studies (SPS). Throughout his career, he has had the opportunity to work with myriad faculty, allowing him to experiment, collaborate, and design various types of learning environments, ranging in size from 12 student seminars to 10,000 student MOOCs.

Professor Nick Cercone was a world-renowned researcher in the fields of artificial intelligence, knowledge-based systems, and human–machine interfaces. He served as dean of the Faculty of Science and Engineering at York University from 2006 to 2009. He joined York from Dalhousie University where he served as dean of computer science between 2002 and 2006. He cofounded *Computational Intelligence*, edited *Knowledge and Information Systems*, and served on editorial boards of six journals. He was president of the Canadian Society for the Computational Studies of Intelligence and of the Canadian Association for Computer Science. He was also a fellow of the IEEE and received a lifetime achievement award for his research on artificial intelligence in Canada.

The dean of the Lassonde School of Engineering, Janusz Kozinski, posted an obituary for Professor Cercone (http://lassonde.yorku.ca/nickcercone).

Rebeca Cerezo started to work as FPI scholarship researcher to the ADIR Research Group in 2007 and teaches in the Department of Psychology at the University of Oviedo since 2010, same year that she earned her Ph.D. in Psychology in that university. Her research interests are focused on metacognition, self-regulation, and educational data mining. She has transferred her work through a large number of projects, chapters, papers, and international conferences. She is an active member of the European Association for Research on Learning and Instruction (EARLI) and the Society for Learning Analytics Research (SoLAR). She is the managing editor of the JCR journal *Psicothema* and associate editor of *Aula Abierta* and *Magister*.

Dr. Hsin-liang (Oliver) Chen is an associate professor in the Palmer School of Library and Information Science at Long Island University. He received his

Ph.D. in Library and Information Science from the University of Pittsburgh, M.A. in Educational Communication and Technology from New York University, and B.A. in Library Science from Fu Jen Catholic University in Taiwan. His research interests focus on the application of information and communication technologies (ICTs) to assist users in accessing and using information in different environments.

Ellina Chernobilsky is an associate professor of education at Caldwell University. Prior to earning her Ph.D., she was a classroom teacher and used action research as means to study her own teaching in order to help herself and her students to become better learners. She teaches action research and other research courses regularly. Her areas of interest include, but are not limited to, the use of data in education, multilingualism, teaching English as second/foreign language, and caring in teaching.

Denise K. Comer is an associate professor of the practice of writing studies and director of First-Year Writing at Duke University. She teaches face-to-face and online writing courses and an MOOC. She earned the 2014 Duke University Teaching with Technology Award. Her scholarship has appeared in leading composition journals and explores writing pedagogy and writing program administration. She has written a textbook based on writing transfer, *Writing in Transit* (Fountainhead, 2015); a dissertation guide, *It's* Just *a Dissertation*, cowritten with Barbara Gina Garrett (Fountainhead, 2014); and a web text, *Writing for Success in College and Beyond* (Connect 4 Education, 2015). She currently lives in North Carolina with her husband and their three children.

Therese Condit holds an Ed.M. in International Education Policy from the Harvard University Graduate School of Education and a B.A. in Music and Rhetoric from Miami University. She has worked in educational technology and MOOC production with Harvard University, MIT, and Columbia University. She is currently an independent education consultant, specializing in program development and evaluation, with New York City public schools, BRIC Arts | Media in Brooklyn, and Wiseman Education in Hong Kong. In addition, she freelances as a film editor and postproduction specialist with Night Agency in New York City. She is also a performing jazz musician and classical accompanist and a former member of Gamelan Galak Tika, the first Balinese gamelan orchestra in the United States, led by Professor Evan Ziporyn at MIT.

Ken Cor has a Ph.D. in Educational Measurement and Evaluation from Stanford University. His areas of focus include educational assessment development, generalizability theory as a basis to inform performance assessment design, and quantitative educational research methods. He uses his measurement skills to support program evaluation efforts within the discipline-specific faculties and departments of higher education as well as to produce and support the production of scholarship in teaching and learning.

Scott Crossley is an associate professor of applied linguistics at Georgia State University. His primary research focus is on natural language processing and the

application of computational tools and machine learning algorithms in language learning, writing, and text comprehensibility. His main interest area is the development and use of natural language processing tools in assessing writing quality and text difficulty. He is also interested in the development of second language learner lexicons and the potential to examine lexical growth and lexical proficiency using computational algorithms.

Giorgio Maria Di Nunzio is an assistant professor of the Department of Information Engineering of the University of Padua, Italy. His main research interests are in interactive machine learning, evaluation of information retrieval systems, and digital geolinguistics. He has developed data visualization tools of probabilistic models for large-scale text analysis in R. His work has been published in journals and conference papers, as well as in books about data classification and data mining applications. Since 2011, he has been in charge of the database systems course of the Department of Information Engineering of the University of Padua; since 2006, he has also been in charge of the foundations of computer science course at the Faculty of Humanities of the same university.

José Diaz is a senior tech specialist at Columbia University's Center for Teaching and Learning, where he films and edits videos and develops massive open online courses (MOOCs). Prior to joining CTL, he worked at the Graduate School of Biomedical Sciences, Icahn School of Medicine at Mount Sinai, where he worked as a technical analyst. He has a bachelor's degree in business administration from Baruch College, CUNY, as well as in computer information systems at the same university and an M.A. in Educational Technology from Adelphi University.

Samira ElAtia is an associate professor of education and the director of graduate studies of Faculté Saint-Jean at the University of Alberta. She holds a Ph.D. from the University of Illinois at Urbana-Champaign. She specializes in the evaluation of competencies; her research interest focuses on issues of fairness in assessment. She is member of the board of directors of the Centre for Canadian Language Benchmarks in Ottawa. She has served as expert on several international testing agencies: Educational Testing Service in the United States, Pearson Education in the United Kingdom, the International Baccalaureate Organization, Chambre du commerce et de l'industrie of Paris, and the Centre international des études pédagogiques of the Ministry of Education in France. She is currently developing her own MOOC in French about assessment of learning in educational settings.

David Eubanks holds a Ph.D. in Mathematics from Southern Illinois University and currently serves as assistant vice president for assessment and institutional effectiveness at Furman University.

William Evers, Jr. is a senior analyst for institutional effectiveness at Eckerd College. He has Master of Arts in Organizational Leadership from Gonzaga University and Bachelor of Arts in Management from Eckerd College.

Rebecca Eynon is an associate professor and senior research fellow at the University of Oxford, where she holds a joint academic post between the Oxford Internet Institute (OII) and the Department of Education. Since 2000 her research has focused on education, learning, and inequalities, and she has carried out projects in a range of settings (higher education, schools, and the home) and life stages (childhood, adolescence, and late adulthood). Rebecca is the coeditor of *Learning, Media, and Technology*. Her work has been supported by a range of funders including the British Academy, the Economic and Social Research Council, the European Commission, Google, and the Nominet Trust. Prior to joining Oxford in 2005, she held positions as an ESRC postdoctoral fellow of the Department of Sociology, City University; as a research fellow of the Department of Education, University of Birmingham; and as a researcher for the Centre for Mass Communication Research, University of Leicester.

Oliver Ferschke is a postdoctoral researcher at Carnegie Mellon University in the Language Technologies Institute. He studies collaboration at scale and seeks to understand how collaboration works in communities through the lens of language and computational linguistics. He holds a Ph.D. in Computer Science from the Ubiquitous Knowledge Processing Lab at TU Darmstadt, Germany, as well as an M.A. in Linguistics and a teaching degree in computer science and English as a second language from the University of Würzburg, Germany. He is furthermore the codirector of the working group on Discussion Affordances for Natural Collaborative Exchange (DANCE).

Nabeel Gillani is currently a product analyst at Khan Academy, working with a passionate team of designers, engineers, and others to help deliver a free, world-class education to anyone, anywhere. Previously, he cofounded the digital internships platform Coursolve.org. He has worked with an interdisciplinary team at the University of Oxford, receiving grants from the Bill & Melinda Gates Foundation and Google to explore how social learning unfolds in online courses. He has an Sc.B. in Applied Mathematics and Computer Science from Brown University and two master's degrees from the University of Oxford (education and technology, machine learning), where he was a Rhodes Scholar.

Isis Hjorth is a researcher at the Oxford Internet Institute and a fellow at Kellogg College, University of Oxford. She is a cultural sociologist, who specializes in analyzing emerging practices associated with networked technologies. She completed her AHRC-funded DPhil (Ph.D.) at the OII in January 2014. Trained in the social sciences as well as the humanities, she holds a B.A. and M.A. in Rhetoric from the Department of Media, Cognition and Communication, University of Copenhagen, and an M.Sc. in Technology and Learning from the Department of Education, University of Oxford. Prior to joining the academic community, she worked in broadcast journalism and screenwriting in her native Copenhagen.

Donald Ipperciel is a professor of political philosophy at Glendon College, York University, Canada. He obtained his doctorate at Ruprecht-Karls-Universität in

Heidelberg in 1996. He held a Canadian research chair in political philosophy and Canadian studies between 2002 and 2012. After an 18-year career at the University of Alberta, where he held many administrative positions (including associate dean (research), associate dean (IT and innovation), vice-dean and director of the Canadian Studies Institute), he moved to Toronto to become the principal of Glendon College, York University. Aside from his philosophical work, he has dedicated many years to questions of learning technologies and big data in education. He has been the Francophone editor of the *Canadian Journal of Learning and Technology* since 2010.

Yutaka Ishii is a research associate at the Center for Higher Education Studies, Waseda University. He received a B.A. and M.Ed. from Waseda University. His main research interest is the data mining approach to learners' writing product and processes.

Joanne Jasmine is a professor of education at Caldwell University. She is a coordinator of the M.A. program in curriculum and instruction and cocoordinator of the Ed.D./Ph.D. program in educational leadership. Dr. Jasmine's recent work focuses on multiculturalism and social justice through literature, strategies for improving the teaching of language arts, and lessons to be learned from preschool children. She also teaches action research classes regularly.

Zhiyong Liu is an associate professor of the Software Institute of Northeast Normal University, China. His research interests include semantic web, knowledge discovery, and data analytics. He is author of 10+ papers and 2 books, held 4 projects, and supervised 12 postgraduates.

Liu obtained bachelor's degree in 2000, master's degree in 2003, and Ph.D. in 2010. He was accepted as a visiting scholar for 1 year in 2013 in the Department of Computer Science and Engineering at York University, Canada. He was awarded one of 100 young academic backbone scholars of Northeast Normal University in 2012 and the second prize bonus of the Higher Education Technology Outcomes by the Education Bureau of Jilin Province in 2010.

Collin F. Lynch is a research assistant professor of computer science at North Carolina State University. He received his Ph.D. in Intelligent Systems from the University of Pittsburgh. His research is focused on graph-based educational data mining and intelligent tutoring systems for ill-defined domains. Dr. Lynch also serves as the policy chair for the International Educational Data Mining Society.

Simon McIntyre is the director of Learning and Innovation at UNSW Australia | Art & Design. He is passionate about improving the effectiveness, quality, and relevance of the student learning experience through innovative and pedagogically driven integration of technology. After developing and teaching online courses in art and design for several years, he helped many other academics design and teach online through designing and convening a range of award-winning academic development programs. His research

explores how online pedagogies, open education and resources, and massive open online courses (MOOCs) can evolve education into a globally networked practice.

Danielle S. McNamara is a professor in cognitive science at Arizona State University. Her research interests include better understanding of the various processes involved in comprehension, learning, and writing in both real-world and virtual settings. She develops educational technologies (e.g., iSTART, Writing Pal) that help students improve their reading comprehension and writing skills and also works on the development of text analysis tools (e.g., Coh-Metrix, TERA, SiNLP, TAALES, TAACO) that provide a wide range of information about text, such as text difficulty and quality. Furthermore, she explores how these tools can be applied to other learning environments such as computer-supported collaborative learning environments and massive open online courses.

Negin Mirriahi has extensive experience in managing, implementing, and evaluating educational technology in higher education and in designing online and blended courses. She currently teaches postgraduate courses in learning and teaching and is a coinstructor of the Learning to Teach Online MOOC. Her research focuses on technology adoption, blended learning, and learning analytics.

Robin A. Moeller is an assistant professor of library science at Appalachian State University in Boone, North Carolina, United States, where she also serves as the director of the Library Science Program. She received her Ph.D. in Curriculum Studies from Indiana University, Bloomington. Before earning her doctorate, she was a school librarian. Her research interests include visual representations of information as they relate to youth and schooling, as well as exploring cultural facets of librarianship and materials for youth.

Stephanie Ogden is the lead digital media specialist at Columbia University's Center for Teaching and Learning. She manages a team of video specialists and influences the overall direction and role of digital video at the CTL. She also oversees all of the CTL video projects from developing productions for digital health interventions to producing interviews with world-renowned artists and intellectuals to directing scripted productions. She works closely with CTL's highly skilled technical team of videographers, editors, programmers, designers, and educational technologists and in partnership with faculty to produce videos for Columbia classes, hybrid courses, online programs, and massive open online courses.

Luc Paquette is an assistant professor of curriculum and instruction at the University of Illinois at Urbana-Champaign where he specializes in educational data mining and learning analytics. He earned a Ph.D. in Computer Science from the University of Sherbrooke. He previously worked as a postdoctoral research associate at Teachers College, Columbia University. One of his main research interests focused on the combination of knowledge engineering and educational data mining approaches to create better and more general models of students who disengage from digital learning environments by gaming the system.

Despina Pitsoulakis is a candidate in human development and psychology at the University of California, Los Angeles, working on the Dynamic Language Learning Progression Project. Her research interests include language and literacy development and assessment, with a particular focus on English language learners. A graduate of Georgetown University, she also holds a Master of Arts in Teaching from American University and a Master of Education from the Harvard Graduate School of Education. Prior to entering UCLA, she worked as an elementary school teacher and reading intervention specialist.

Patsie Polly is an associate professor in pathology and UNSW teaching fellow, UNSW, Australia. She is recognized for her medical research in gene regulation and higher education innovation. She also brings this experience to undergraduate science students with focus on using ePortfolios and virtual laboratories to develop professional and research practice skills. She has an extensive experience in authentic assessment as well as course-wide and program-wide ePortfolio use. She has also been recognized with multiple institutional and national teaching awards, with invited national and international presentations and peer-reviewed research outputs in research communication and ePortfolio use. She has attracted institutional and national funding to support development of e-learning resources.

Octav Popescu is a senior research programmer/analyst in Carnegie Mellon's Human–Computer Interaction Institute, where he is in charge of Tutor Shop, the learning management system part of the Cognitive Tutor Authoring Tools project. He has more than 25 years of experience working on various projects involving natural language understanding and intelligent tutoring systems. He holds an M.S. in Computational Linguistics and a Ph.D. in Language Technologies from Carnegie Mellon University.

Jean-Paul Restoule is an associate professor of aboriginal education at the Ontario Institute for Studies in Education of the University of Toronto (OISE/UT). He designed OISE's first MOOC, Aboriginal Worldviews and Education, which is launched in February 2013. The course continues to be viewed by approximately 60 new registrants a week.

Dr. Edith Ries is a professor of education at Caldwell University. Her recent presentations focus on the use of young adult literature as a vehicle for teaching social justice and global awareness. She teaches action research graduate-level classes at the university and has mentored several award-winning action research projects.

Geoffrey Rockwell is a professor of philosophy and humanities computing at the University of Alberta, Canada. He has published and presented papers in the area of big data, textual visualization and analysis, computing in the humanities, instructional technology, computer games, and multimedia including a book on humanities, *Defining Dialogue: From Socrates to the Internet*, and a forthcoming book from MIT Press, *Hermeneutica: Thinking Through Interpretative Text Analysis*. He collaborates with Stéfan Sinclair on Voyant Tools (http://voyant-tools.org), a suite

of text analysis tools, and leads the TAPoR (http://tapor.ca) project documenting text tools for humanists. He is currently the director of the Kule Institute for Advanced Study.

Cristóbal Romero received the B.Sc. and Ph.D. degrees in computer science from the University of Granada, Spain, in 1996 and 2003, respectively. He is currently an associate professor in the Department of Computer Science and Numerical Analysis, University of Cordoba, Spain. He has authored 2 books and more than 100 international publications, 33 of which have been published in journals with ISI impact factor. He is a member of the Knowledge Discovery and Intelligent Systems (KDIS) Research Group, and his main research interest is applying data mining and artificial intelligence techniques in e-learning systems. He is a member of IEEE, and he has served in the program committee of a great number of international conferences about education, artificial intelligence, personalization, and data mining.

Carolyn Rosé is an associate professor of language technologies and human–computer interaction in the School of Computer Science at Carnegie Mellon University. Her research program is focused on better understanding of the social and pragmatic nature of conversation and using this understanding to build computational systems that can improve the efficacy of conversation between people or between people and computers. In order to pursue these goals, she invokes approaches from computational discourse analysis and text mining, conversational agents, and computer-supported collaborative learning. She serves as president of the International Society of the Learning Sciences. She also serves as associate editor of the *International Journal of Computer-Supported Collaborative Learning* and the *IEEE Transactions on Learning Technologies*.

Eve Ryan is a Ph.D. candidate in human development and psychology at the University of California, Los Angeles, working on the Dynamic Language Learning Progression Project. She holds a master's degree in language testing from Lancaster University and has experience in the areas of language assessment and language teaching. Her research interests also include language and literacy development in the early years.

Miguel Sánchez-Santillán received his B.Sc. in Computer Science from the University of Oviedo in 2010, where he also got his master's degree in web engineering in 2012. Currently, he is a Ph.D. student at the research groups PULSO and ADIR at the same university. His main research interests are focused on educational data mining and adaptive hypermedia systems for e-learning.

Jonathan Sewall is a project director on the staff of the Human–Computer Interaction Institute at Carnegie Mellon University. He coordinates design and development work on the Cognitive Tutor Authoring Tools (CTAT), a software suite meant to aid in creation and use of intelligent tutoring systems (ITS). Prior to coming to CMU in 2004, he held various software development and testing positions in industry and government spanning a period of more than 20 years.

Nancy Frances Smith is a professor of marine science and biology at Eckerd College, where she has been a member of the faculty since 2000. Her teaching includes courses in introductory oceanography, marine invertebrate biology, ecology, and parasitology. She has also taught courses in Australia, Micronesia, and Latin America. Her research focuses on a broad range of topics in ecology, from the evolution of marine invertebrate life history to the interactions between marine parasites and their hosts. She advocates for initiating undergraduates in authentic research at the freshman level and has directed the marine science freshman research program at Eckerd. She has published in journals such as *Journal of Parasitology*, *Journal of Experimental Marine Biology and Ecology*, and *Biological Bulletin*.

Thuan Thai is a senior lecturer in the School of Education, University of Notre Dame Australia, where he teaches mathematics and science pedagogy in the teacher education programs. His research explores the use of technology to track and assess student learning and performance, as well as promote engagement, reflection, and professional development. He has over 10 years of experience as a medical researcher (cardiovascular disease) and previously taught pathology in the science, medical science, and health and exercise science programs at UNSW Australia.

Gaurav Singh Tomar is a graduate research assistant at the Language Technologies Institute in the School of Computer Science at Carnegie Mellon University.

Lorenzo Vigentini has a background in psychology, and his research interest is in individual differences in learning and teaching. His work focuses on the exploration of a variety of data sources and the techniques to make sense of such differences with a multidisciplinary, evidence-based perspective (including psychology, education, statistics, and data mining). He is currently the coordinator of the data analytics team in the Learning and Teaching Unit and is leading a number of initiatives in the learning analytics space at UNSW.

Yuan "Elle" Wang is doctoral research fellow in cognitive and learning sciences in the Department of Human Development at Teachers College, Columbia University. As an MOOC researcher, her research focuses on MOOC learner motivation, course success metrics, and postcourse career development measurement. As an MOOC practitioner, she has been a key member in the instructors' team for three MOOCs offered via both *Coursera* and *edX*. She received her M.A. in Communication, Technology, and Education in the Department of Mathematics, Sciences, and Technology also from Columbia University. She has previously published in peer-reviewed scientific journals such as *Journal of Learning Analytics*, *MERLOT Journal of Online Learning and Teaching*, and *InSight: A Journal of Scholarly Teaching*.

Taha Yasseri is a research fellow in computational social science at the Oxford Internet Institute (OII), University of Oxford. He graduated from the Department of Physics at the Sharif University of Technology, Tehran, Iran, in 2005, where he also obtained his M.Sc. in 2006, working on localization in scale-free complex networks. In 2007, he moved to the Institute of Theoretical Physics at the University of

Göttingen, Germany, where he completed his Ph.D. in Complex Systems Physics in 2010. Prior to coming to the OII, he spent two years as a postdoctoral researcher at the Budapest University of Technology and Economics, working on the sociophysical aspects of the community of Wikipedia editors.

Osmar R. Zaïane is a professor in computing science at the University of Alberta, Canada, and the scientific director of the Alberta Innovates Centre for Machine Learning (AICML). He obtained his Ph.D. from Simon Fraser University, Canada, in 1999. He has published more than 200 papers in refereed international conferences and journals. He is associate editor of many international journals on data mining and data analytics and served as program chair and general chair for scores of international conferences in the field of knowledge discovery and data mining. He received numerous awards including the 2010 ACM SIGKDD Service Award from the ACM Special Interest Group on Data Mining, which runs the world's premier data science, big data, and data mining association and conference.

Jing Zhang is a master's student in cognitive studies in education at Teachers College, Columbia University. Her master thesis is on using educational data mining methods to predict student's retention in an MOOC learning environment. Before that, she obtained an M.A. in Instructional Technology and Media at Teachers College. At that time, her master thesis was on motivational theories that were related to MOOCs.

INTRODUCTION: EDUCATION AT COMPUTATIONAL CROSSROADS

Samira ElAtia[1], Donald Ipperciel[2], and Osmar R. Zaïane[3]
[1] Campus Saint-Jean, University of Alberta, Edmonton, Alberta, Canada
[2] Glendon College, York University, Toronto, Ontario, Canada
[3] Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada

For almost two decades, data mining (DM) has solidly grounded its place as a research tool within institutions of higher education. Defined as the "analysis of observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owners (Han, Kamber, and Pei, 2006)," DM is a multidisciplinary field that integrates methods at the intersection of artificial intelligence (AI), machine learning, natural language processing (NLP), statistics, and database systems. DM techniques are used to analyze large-scale data and discover meaningful patterns such as natural grouping of data records (cluster analysis), unusual records (anomaly and outlier detection), and dependencies (association rule mining). It has made major advances in biomedical, medical, engineering, and business fields. Educational data mining (EDM) emerged in the last few years from computer sciences as a field in its own right that uses DM techniques to advance teaching, learning, and research in higher education. It has matured enough to have its own international conference (http://www.educationaldatamining.org). In 2010, Marc Parry in an article in *The Chronicles of Higher Education* suggested that academia is "at a computational crossroads" when it comes to big data and analytics in education. DM, learning analytics (LA), and big data offer a new way of looking, analyzing, using, and studying data generated from various educational settings, be it for admission; for program development, administration, and evaluation; within the classroom and e-learning environments, to name a few.

This novel approach to pedagogy does not make other educational research methodologies obsolete but far from it. The richness of available methodologies will continue to shed light on the complex processes of teaching and learning, adjusting as required to the object of study. However, DM and LA are providing educational researchers with additional tools to afford insight into circumstances that were previously obscured either because methodological approaches were confined to a small number of cases, making any generalization problematic, or because available data sources were so massive that analyzing them and extracting information from them was far too challenging. Today, with the computational tools at our disposal,

educational research is poised to make a significant contribution to the understanding of teaching and learning.

Yet, most of the advances in EDM so far are, to a large extent, led by computing sciences. Educators from "education fields per se," unfortunately, play a minor role in EDM, but the potential for a collaborative initiative between the two would open doors to new researches and new insights into higher education in the twenty-first century. We believe that advances in pedagogical and educational research have remained tangential and not exploited as it should be in EDM and have thus far played a peripheral role in this strongly emerging field that could greatly benefit and shape education and educational research for various stakeholders.

This book showcases the intersection between DM, LA, EDM, and education from a social science perspective. The chapters in this book collectively address the impacts of DM on education from various perspectives: insights, challenges, issues, expectations, and practical implementation of DM within educational mandates. It is a common interdisciplinary platform for both scientists at the cutting edge of EDM and educators seeking to use and integrate DM and LA in addressing issues, improving education, and advancing educational research. Being at the crossroads of two intertwined disciplines, this book serves as a reference in both fields with implementation and understanding of traditional educational research and computing sciences.

When we first started working on this project, the MOOC was the new kid on the block and was all the rage. Many were claiming it would revolutionize education. While all the hype about the MOOC is fading, a new life has been breathed into the MOOC with some substantial contributions to research on big data, something that has become clear as our work on this volume progressed. Indeed, the MOOC has opened a new window of research on large educational data. It is thus unsurprising that in each of the three parts of this book, there is a chapter that uses a MOOC delivery system as the basis for their enquiry and data collection. In a sense, MOOCs are indeed the harbinger of a new, perhaps even revolutionary, educational approach, but not for the reasons put forward at the height of the craze. Education will probably not be a "massive" enterprise in the future, aside from niche undertakings; it will probably not be entirely open, as there are strong forces—both structural and personal—working against this, and it is unlikely that it will have a purely online presence, the human element of face-to-face learning being and remaining highly popular among learners. However, the MOOC does point to the future in that it serves as a laboratory and study ground for a renewed, data-driven pedagogy. This becomes especially evident in EDM.

On a personal note, we would like to pay homage to one of the authors of this volume, the late Nick Cercone. At the final stages of editing and reviewing the chapters, Professor Nick Cercone passed away. Considered one of the founding fathers of machine learning and AI in the 1960s, Professor Cercone's legacy spans six decades with an impressive record of research in the field. He witnessed the birth of DM and LA and we were honored to count him among the contributors. He was not only an avid researcher seeking to deepen our understanding in this complex field but also an extraordinary educator who worked hard to solve issues relating to higher education as he took on senior administrative positions across Canada. Prof. Cercone's legacy and his insight live on in this book as a testimony to this great educator.

This edited volume contains 15 chapters grouped into three parts. The contributors of these chapters come from all over the world and from various disciplines. They need not be read in the order in which they appear, although the first part lays the conceptual ground for the following two parts. The level of difficulty and complexity varies from one article to the other and from the presentation of learning technology environment that makes DM possible (e.g., Thai and Polly, 2016) to mathematical and probabilistic demonstration of DM techniques (e.g., Di Nunzio, 2016). They all present a different aspect of EDM that is relevant to beginners and experts.

The articles were selected not only in the field of DM per se but also in propaedeutic and grounding areas that build up to the more complex techniques of DM. Level 1 of this structure is occupied by *learning systems*. They are foundationally important insofar as they represent the layer in which educational data is gathered and preorganized. Evidently, there is no big data without data collection and data warehousing. Chapters relating to this level present ideas on types of data and information that can be collected in an educational context. Level 2 pertains to *LA* stricto sensu. LA uses descriptive methods mainly drawn from statistics. Here, information is produced from data by organizing and, as it were, "massaging" it through statistical and probabilistic means. Level 3 of the structure is home to *DM* in the narrow sense in which machine learning techniques and algorithmics are included. These techniques allow for true knowledge discovery, that is, pattern recognition that one did not foresee within a massive data set. This layer builds on the previous, as statistical tools and methods are also used in this context, and is dependent on the first layer, where relevant data is first collected and preorganized. To be sure, DM is also commonly used to refer to levels 2 and 3 in a looser sense. And some authors in this book utilize at times the term in this way. Nonetheless, it makes sense to distinguish these concepts in a more rigorous context.

I.1 PART I: AT THE INTERSECTION OF TWO FIELDS: EDM

Articles in the first part present a general overview and definitions of DM, LA, and data collection models in the context of educational research. The goal is to share information about EDM as a field, to discuss technical aspects and algorithm development in DM, and to explain the four guiding principles of DM: prediction, clustering, rule association, and outlier detection. Throughout this part, readers will deepen their understanding not only of DM and LA and how they operate but also of the type of data and the organization of data needed for carrying EDM studies within an educational context at both the macro- (e.g., programs, large-scale studies) and microlevels (e.g., classroom, learner centered).

In the first chapter, **Romero et al.** present the emblematic exploratory analysis one could do on data using off-the-shelf open-source software. They present a study of the learning activity process passing and failing students follow in an online course, and they indeed use existing free data analysis tools such as ProM for process mining and Weka (Witten and Frank, 2005), a machine learning and DM tool, to do

their analysis. By combining these tools, they obtain models of student learning behaviors for different cohorts.

From the humanities perspectives, **Rockwell and Berendt** discuss the important role and potential that DM, and especially text mining, can have in the study of large collections of literary and historical texts now made available through projects such as Google Books, Project Gutenberg, etc. They present a historical perspective on the development of the field of text mining as a research tool in EDM and in the humanities.

Eubanks et al. compare the use of traditional statistical summaries to using DM in the context of higher education and finding predictors ranging from enrollment and retention indicators, financial aid, and revenue predictions to learning outcomes assessment. They showcase the significance of EDM in managing large data generated by universities and its usefulness to better predict the success of the learning experience as a whole.

Baker is one of the pioneers in EDM and one of the innovators in student modeling and intelligent tutoring systems. With his colleagues he has developed a MOOC on EDM and LA, big data in education, that went through two iterations. Baker et al. recount their experience setting up this MOOC, with the goal to use EDM methods to answer educational research questions. They first describe the tools and content, and the lessons learned, but later highlight how this MOOC and the data it provided supported research in EDM, such as predicting dropouts, analyzing negativity toward instructors, studying participations in online discussions, etc.

Chernobilsky et al. examine two different approaches to using DM within action research studies in a purely educational sense. Action research is a widely used approach to study various phenomena in which changes are made to a learning/teaching context as research is being conducted and in which the context adapts as results are analyzed. Because of its qualitative nature, action research would at first glance seem incompatible with EDM. However, Chernobilsky et al. attempt to bridge the two fields by exploring ways in which these two investigative approaches can be made compatible and complementary to one another in order to guide teachers and researchers in making informed decision for improving teaching practice.

I.2 PART II: PEDAGOGICAL APPLICATIONS OF EDM

The five chapters of this part address issues relating to the applications of and challenges to using DM and LA in pedagogical settings. They aim to highlight effective classroom practices in which EDM can advance learning and teaching. In order to ensure a broad representation of various educational settings, we sought studies mainly outside of the field of computing sciences. Social networking in a classroom setting, students' interactions, feedback, response analyses, and assessment are some of the teaching tools through which EDM has been proven effective within the classroom.

In the opening chapter of this part, **Liu and Cercone** present their work on developing and using an adaptive learning system (ALS) within an e-learning environment that can intelligently adapt to the learner's needs by evaluating these needs and presenting only the suitable learning contents, learning paths, learning instruction,

and feedback solely based on their unique individual characteristics. From an AI computing perspective, they focus on the dimensionalities of the user model, which define and organize the personal characteristics of users.

Di Nunzio shifts to presenting a more technical aspect of EDM in engineering. He focuses on DM as in interdisciplinary fields in which students study foundations of machine learning and probabilistic models for classification. He presents interactive and dynamic geometric interpretations of probabilistic concepts and designs a tool that allows for data collection, which in turn can be used to improve the learning process and student performance using EDM.

Using their MOOC "Aboriginal Worldviews and Education," **Brugha and Restoule** study the effectiveness of online networks in promoting learning, particularly for "traditionally marginalized learners of higher education." They look into how to set up the online networks and discuss how they use data analytics to explore the big data generated from e-learning educational environment. Ultimately, their goal is to ensure that good and sound pedagogical practices are being addressed in online educational directives.

Thai and Polly, both from the Department of Pathology at the University of New South Wales, present a unique way in using DM in e-portfolios deployed for educational purposes for students in the medical sciences. Turning their backs on the static and theory-oriented educational software previously used in medical education, they take advantage of virtual labs as dynamic learning spaces, which allow them to showcase several opportunities for DM during the learning process of medical education.

EDM can have various applications within the social sciences as the chapters in Part I attest. This is also confirmed by the work of **Yutaka Ishii**, who focuses on the analysis of grammatical errors in the written production of university students learning another language, Japanese students learning English in this case. He demonstrates the usefulness of rule association as it applies to the co-occurrence of patterns in learners' grammatical errors in order to explain and further advance research in second/foreign language acquisition. Using DM on large data sets, Ishii conducts association analysis in order to discover correlations and patterns in the production of errors.

I.3 PART III: EDM AND EDUCATIONAL RESEARCH

In this part, the articles will exclusively focus on EDM in educational research. An important aspect and use of EDM is the potential role it can play in providing new research tools for advancing educational research, as well as for exploring, collecting, and analyzing data. EDM is an innovative research method that has the potential to revolutionize research in education: instead of following predetermined research questions and predefined variables, EDM can be used as a means to look at data holistically, longitudinally, and transversally and to "let" data speak for itself, thus revealing more than if it were restricted to specific variables within a time constraint.

Vigentini et al., using data collected from a MOOC, explore the effect of course design on learner engagement. Their key hypothesis is that the adaptive and

flexible potential of a MOOC, designed to meet the varying intents and diverse learning needs of participants, could enable personally meaningful engagement and learning, as long as the learners are given the flexibility to choose their learning paths. They delved into the pedagogical implication of motivation, engagement, and self-directed learning in e-learning environments.

Eynon et al. use EDM also within a MOOC environment to understand the communication patterns among users/students, combining both DM on large data and qualitative methods. While qualitative methods had been used in the past with small sample sizes, Eynon and her Oxford team assign an important role for EDM in carrying qualitative studies on large-scale longitudinal data.

In institutions of higher education, EDM goes beyond research on learning. It can also be an extremely useful tool for administrative purposes. **Brocks and Cor**, using longitudinal data from over 262,000 records from a pharmacy program, investigate, within very competitive programs that have a set admission quota, the relationship between applicant attributes and academic measures of success. They mined a large data set in order to look into the admission process of a pharmacy program and the impact predetermined courses and other criteria for admission have on the success of students.

Moeller and Chen, in a two-step study, use textual analysis of online discussions and the most circulated books in selected schools to investigate how children view the concept of difference as it relates to race and ethnicity. Although both race and ethnicity have been extensively studied, for television, online forums, and children's picture books, this study uses DM as a new approach to research the issues of race, ethnicity, and education from a different perspective.

In the last chapter, **Bailey et al.** showcase the usefulness of DM and NLP in research in elementary education. From an interdisciplinary perspective, they aim to build a digital data system that uses DM, corpus linguistics, and NLP and that can be queried to access samples of typical school-age language uses and to formulate customizable learning progressions. This system will help educators make informed assessment about children language progress.

REFERENCES

Di Nunzio, G. M. (2016). "The 'Geometry' of Naïve Bayes: Teaching Probabilities by 'Drawing' Them," *Data Mining and Learning Analytics: Applications in Educational Research*. Hoboken, NJ, John Wiley & Sons, Inc.

Han, J., M. Kamber, and J. Pei (2011). *Data Mining: Concepts and Techniques*, 2nd ed. San Francisco, CA, Morgan Kaufmann Publishers.

Parry, M. (2010). "The Humanities Go Google," *The Chronicles of Higher Education*, May 28, 2010. On-line: http://chronicle.com/article/The-Humanities-Go-Google/65713/. Accessed April 22, 2016.

Thai, T. and P. Polly (2016). "Exploring the Usefulness of Adaptive eLearning Laboratory Environments in Teaching Medical Science," *Data Mining and Learning Analytics: Applications in Educational Research*. Hoboken, NJ, John Wiley & Sons, Inc.

Witten, I. H. and E. Frank (2005). *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. San Francisco, CA, Elsevier.