

CHARLES ELKAN

Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139
(617) 258-7621
elkan@csail.mit.edu
September 1, 2005

ACADEMIC EMPLOYMENT

Visiting Scientist, Computer Science and Artificial Intelligence Laboratory, MIT, August 2005 to July 2006.

Professor, Department of Computer Science and Engineering, University of California, San Diego, since July 2004. Associate Professor, 1997 to 2004. Assistant Professor, 1990 to 1997.

Visiting Associate Professor, Department of Computer Science, Harvard University, 1998/99.

Postdoctoral Fellow, Knowledge Representation and Reasoning Group, Department of Computer Science, University of Toronto, 1989/90.

Lecturer, University of Pennsylvania, June 1989 and February 1990.

Lecturer, Cornell University, summer 1987.

EDUCATION

Ph.D. in computer science, Cornell University, conferred August 1990. Thesis title *Flexible Concurrency Control by Reasoning about Queries and Updates*. Committee members Prakash Panangaden, Dexter Kozen and George Staller (Department of Economics).

M.S. in computer science, Cornell University, January 1988.

Visiting scholar in artificial intelligence and economics, Stanford University, 1986.

B.A. (Honors) in mathematics, Fitzwilliam College, Cambridge University, June 1984. Concentration in optimization and statistics.

Double “maturité” with distinction in natural sciences and distinction in Latin, Collège Rousseau, Geneva, Switzerland, June 1981.

AWARDS

Award for first place out of 43 entries in the CoIL Challenge data mining competition, May 2000.

Award for first place out of 45 entries in the data mining competition at the International Conference on Knowledge Discovery in Databases, August 1997.

Best paper award, IEEE Conference on Artificial Intelligence for Applications (CAIA'93), March 1993, for *Categorization-Based Diagnostic Problem Solving in the VLSI Design Domain* with Amir Hekmatpour.

Honorable mention, best-written paper competition, National Conference on Artificial Intelligence

(AAAI'93), July 1993, for *The Paradoxical Success of Fuzzy Logic*.

RESEARCH GRANTS

Learning correct policies despite bias in available training data, California MICRO and Fair Isaac, Inc., \$59,616, 2004/05 (PI).

Gift to support UCSD data mining contests, Fair Isaac, Inc., \$50,000, 2004 and 2005 (PI).

Emergence and development of cognition, Matsushita Electric Industrial Corporation, Osaka, Japan, \$195,000, 2001 to 2004 (PI).

Cost-sensitive decision-making, California MICRO and Fair Isaac, Inc., \$64,750, 2003/04 (co-PI).

Data mining for reliability and availability, Sun Microsystems, Inc., \$50,000, 2003/04 (PI).

Early prediction of disk drive failures, Alfred P. Sloan Foundation, approximately \$40,000 of total \$2,140,000, 1998/2001 (co-investigator).

Motif-based hidden Markov modeling of protein families, National Science Foundation, \$50,000, 1996/97 (PI).

Equipment to support teaching and research in data mining, Silicon Graphics Inc., \$37,000, 1996 (PI).

Unsupervised learning using expectation-maximization, Hellman Faculty Fellowship, 1995/96, \$29,924 (PI).

Flexible distributed concurrency control by reasoning about transactions, National Science Foundation IRI-9110813, 1991/93, \$60,000 (PI).

Studentship to support Timothy Bailey, NIH Genome Analysis Training Grant HG00005, approximately \$100,000, 1991/95, (co-investigator).

Research initiation grant, Powell Foundation, 1990/91, \$52,725 (PI).

RECENT CONSULTING SERVICE

Consultant on research management, Systems Development Laboratory, Hitachi Ltd., Tokyo, 2003 and 2004.

Consultant on research management, Hewlett Packard Laboratories, Palo Alto, 1999 and 2000.

Senior Scientist, Knowledge Stream Partners, Inc., Boston, 1998 and 1999.

Consultant on data mining, Science Applications International Corporation (SAIC), San Diego, 1997.

Consultant on information system design, Sony (San Diego), IBM (San Jose), and Alcoa (San Diego), 1995.

Consultant on software intellectual property issues, Appx Inc. (Richmond, Virginia) and Morris, Manning & Martin, Attorneys at Law (Atlanta, Georgia), 1995.

RECENT UNIVERSITY SERVICE

Chair, CSE M.S. committee, 2004/05.

Member, Academic Senate Committee on Educational Policy, 2004/05.

Member, Academic Senate Council, 2003/2004.

Chair, Academic Senate Committee on Affirmative Action and Diversity, 2003/04. Member, 2002/03.

Chair, Academic Senate Library Committee, 2001/02. Vice chair, 2000/01. Member, 1999/2000.

Organizer, interdisciplinary Ph.D. program in cognitive science seminar series on *Historical and Conceptual Foundations of Cognitive Science*, fall 2001.

Senior Fellow, San Diego Supercomputer Center, since 1995.

RECENT EDITORIAL SERVICE

Editorial board member: *IEEE Transactions on Data and Knowledge Engineering*, *Journal of Artificial Intelligence Research*, *Machine Learning*, and *Computational Intelligence* journals.

Program committee member, International Conference on Machine Learning (ICML, 2001, 2002, area chair 2005), SIAM International Conference on Data Mining (2005), IEEE International Conference on Data Mining (2003), International Conference on Knowledge Discovery and Data Mining (KDD, 2003, 2005), International Joint Conference on Artificial Intelligence (IJCAI, 2003), SIGIR Conference on Research and Development in Information Retrieval (2002, 2003, 2004), International Conference on Discovery Science (2000), and many other conferences before 2000.

Reviewer of papers submitted to *ACM Transactions on Computer Systems*, *ACM Transactions on Database Systems*, *Annals of Pure and Applied Logic*, *Artificial Intelligence*, *Computer Applications in the Biological Sciences*, *Computational Intelligence*, *IEEE Expert*, *IEEE Transactions on Data and Knowledge Engineering*, *Information Processing Letters*, *Journal of Artificial Intelligence Research*, *Journal of Automated Reasoning*, *Journal of Data Mining and Knowledge Discovery*, *Journal of Logic and Computing*, *Machine Learning*, *Neural Computation*.

Reviewer of book drafts submitted to Academic Press, Benjamin Cummings Publishing Co., McGraw-Hill Publishing Co., MIT Press, and Morgan Kaufmann Publishers.

Reviewer of journal proposals for Academic Press and Sage Periodicals Press.

RECENT PROFESSIONAL SERVICE

Referee for a promotion to full professor, Chinese University of Hong Kong, 2005.

Referee for a promotion to associate professor with tenure, University of Indiana, 2005.

Proposal reviewer, Netherlands Organization for Scientific Research (NWO), 2004.

Review panel member for EPSCoR Research Infrastructure Improvement proposals, state of Nebraska, 2003.

Referee for a promotion to associate professor with tenure, Northwestern University, 2003.

Review panel member for CAREER proposals, Knowledge and Cognitive Systems Program, Na-

tional Science Foundation, 2002.

Referee for a promotion to full professor, University of Iowa, 2000.

Referee for an appointment as full professor, UC Irvine, 1999.

Organizer, knowledge discovery contest and classifier learning contest, International Conference on Knowledge Discovery and Data Mining (KDD), 1999.

RECENT INVITED TALKS

Reinforcement learning for data mining, Fair Isaac Inc., November 5, 2004

Clustering with k-means: faster, smarter, cheaper, Workshop on Clustering High-Dimensional Data, SIAM International Conference on Data Mining, April 24, 2004.

What are the real challenges in data mining?, Workshop on Learning from Imbalanced Data Sets, International Conference on Machine Learning, August 21, 2003.

Measuring the quality of messages and documents automatically, Department of Computer Science, University of California, Santa Barbara, November 20, 2000.

Towards self-financing research, Santa Fe Institute seminar series, Tippie College of Business, University of Iowa, October 22, 1999.

PH.D. GRADUATES

Amir Hekmatpour, Ph.D. conferred June 1993. Thesis title *A methodology and architecture for interactive knowledge-based diagnostic problem-solving in VLSI manufacturing*. Now Senior Engineer, IBM Microelectronics. Senior Lecturer, Department of Electrical and Computer Engineering, University of Texas, Austin, 1996/98.

Timothy L. Bailey, Ph.D. conferred June 1995. Thesis title *Discovering motifs in DNA and protein sequences: The approximate common substring problem*. Now Senior Lecturer with tenure, University of Queensland, Australia.

Michael Sussna, Ph.D. conferred June 1997. Thesis title *Text retrieval using inference in semantic metanetworks*.

Alvaro Monge, Ph.D. conferred August 1997. Thesis title *Adaptive detection of approximately duplicate database records and the database integration approach to information discovery*. Now Associate Professor with tenure, Department of Computer Science, California State University, Long Beach.

William N. Grundy, Ph.D. conferred June 1998. Thesis title *A Bayesian approach to motif-based protein modeling*. Now Assistant Professor (tenure-track), Department of Genome Sciences, University of Washington (Seattle). Formerly Assistant Professor (tenure-track), Department of Computer Science, Columbia University.

Greg Hamerly, Ph.D. conferred June 2003. Thesis title *Learning structure and concepts in data through data clustering*. Now Assistant Professor (tenure-track), Department of Computer Science, Baylor University.

Bianca Zadrozny, Ph.D. conferred August 2003. Thesis title: *Policy mining: Learning decision policies from fixed sets of data*. Now research staff member (permanent), IBM Research, Yorktown

Heights, NY.

REFEREED PAPERS

Charles Elkan, David Lubinsky, and Daryl Pregibon. Automated descriptions of data. In *Proceedings of the Fifth International Symposium on Data Analysis and Computer Science*, pp. 165–170. Versailles, France, September 1987.

Charles Elkan and David McAllester. Automated inductive reasoning about logic programs. In *Proceedings of the Fifth International Conference Symposium on Logic Programming*, pp. 876–892. Seattle, August 1988. MIT Press.

Charles Elkan. A decision procedure for conjunctive query disjointness. In *Proceedings of the Eighth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS'89)*, pp. 134–139. Philadelphia, March 1989. ACM Press.

Charles Elkan. Conspiracy numbers and caching for searching and/or trees and theorem-proving. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence (IJCAI'89)*, pp. 341–346. Detroit, August 1989. Morgan Kaufmann Publishers, Inc.

Charles Elkan. Logical characterizations of nonmonotonic TMSs. In *Proceedings of the Symposium on Mathematical Foundations of Computer Science*, pp. 218–224. Porąbka-Kozubnik, Poland, August 1989. Springer Verlag Lecture Notes in Computer Science, no. 379.

Charles Elkan. Independence of logic database updates and queries. In *Proceedings of the Ninth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS'90)*, pp. 154–160. Nashville, Tennessee, April 1990. ACM Press.

Charles Elkan. A rational reconstruction of nonmonotonic truth maintenance systems. *Artificial Intelligence* **43(2)**219–234, May 1990.

Charles Elkan. Incremental, approximate planning. In *Proceedings of the National Conference on Artificial Intelligence (AAAI-90)*, pp. 145–150. Boston, August 1990. MIT Press.

Alberto Segre, Charles Elkan, and Alex Russell. A critical look at experimental evaluations of EBL. *Machine Learning* **6(2)**183–195, February 1991.

Charles Elkan. Formalizing causation in first-order logic: Lessons from an example. In *Working Notes of the AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*, pp. 41–47. Stanford, California, March 1991.

Russell Greiner and Charles Elkan. Measuring and improving the effectiveness of representations. In *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence (IJCAI'91)*, pp. 518–524. Sydney, Australia, August 1990. Morgan Kaufmann Publishers, Inc.

Alberto M. Segre, Charles Elkan, Daniel Scharstein, Geoffrey J. Gordon, and Alexander Russell. Adaptive inference. In *Foundations of Knowledge Acquisition, Vol. II: Machine Learning: Induction, Analogy, and Discovery*, edited by Susan Chipman and Alan Meyrowitz, pp. 43–81. Kluwer Academic Publishers, 1992.

Charles Elkan. Reasoning about action in first-order logic. In *Proceedings of the Ninth Biennial Conference of the Canadian Society for Computational Studies of Intelligence (CSCSI'92)*, pp. 221–

227. Vancouver, Canada, May 1992. Morgan Kaufmann Publishers, Inc.

Amir Hekmatpour and Charles Elkan. Categorization-based diagnostic problem solving in the VLSI design domain. In *Proceedings of the Ninth IEEE Conference on Artificial Intelligence for Applications (CAIA'93)*, pp. 121–127. Orlando, Florida, March 1993. IEEE Press. (Winner, best paper award.)

Charles Elkan. The paradoxical success of fuzzy logic. In *Proceedings of the Eleventh National Conference on Artificial Intelligence (AAAI-93)*, pp. 698–703. Boston, Massachusetts, July 1993. MIT Press. (Honorable mention, best written paper award.)

Timothy L. Bailey and Charles Elkan. Estimating the accuracy of learned concepts. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence (IJCAI'93)*, pp. 895–900. Chambéry, France, September 1993. Morgan Kaufmann Publishers, Inc.

Alberto M. Segre and Charles Elkan. A high-performance explanation-based learning algorithm, *Artificial Intelligence* **69(1)**1–50, September 1994.

Timothy L. Bailey and Charles Elkan. Cross-validation and modal theories. In *Computational Learning Theory and Natural Learning Systems, Volume III: Selecting Good Models*, ed. Thomas Petsche, Stephen J. Hanson, and Jude W. Shavlik, pp. 345–359. MIT Press, 1995.

Charles Elkan. The paradoxical success of fuzzy logic. *IEEE Expert* **6(6)**3–8, August 1994. With fifteen responses by other researchers on pp. 9–46.

Charles Elkan. The paradoxical controversy over fuzzy logic. *IEEE Expert* **6(6)**47–49, August 1994.

Timothy L. Bailey and Charles Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology (ISMB'94)*, pp. 28–36. Stanford, California, August 1994. AAAI Press.

Timothy L. Bailey and Charles Elkan. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning* **21(1-2)**51–80, October 1995.

Timothy L. Bailey and Charles Elkan. The value of prior knowledge in discovering motifs with MEME. In *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology (ISMB'95)*, pp. 21–29. Cambridge, England, July 1995. AAAI Press.

Alberto M. Segre, Geoffrey J. Gordon, and Charles Elkan. Exploratory analysis of speedup learning data using expectation maximization. *Artificial Intelligence*, **85(1-2)**301–319, August 1996.

Karan Bhatia and Charles Elkan. LPMEME: A statistical method for inductive logic programming. In *Proceedings of the Eleventh Biennial Conference of the Canadian Society for Computational Studies of Intelligence (CSCSI'96)*, pp. 227–239. Toronto, Canada, May 1996. Springer Verlag.

Charles Elkan. Reasoning about unknown, counterfactual, and nondeterministic actions in first-order logic. In *Proceedings of the Eleventh Biennial Conference of the Canadian Society for Computational Studies of Intelligence (CSCSI'96)*, pp. 54–68. Toronto, Canada, May 1996. Springer Verlag.

William N. Grundy, Timothy L. Bailey, and Charles Elkan. ParaMEME: a parallel implementation and a web interface for a DNA and protein motif discovery tool. *Computer Applications in the*

Biosciences, **12(4)**:303–310, 1996.

Alvaro E. Monge and Charles Elkan. The field matching problem: Algorithms and applications. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*, pp. 267–270. Portland, Oregon, August 1996. AAAI Press.

Timothy L. Bailey, Michael E. Baker, and Charles Elkan. An artificial intelligence approach to motif discovery in protein sequences: Application to steroid dehydrogenases. *Journal of Steroid Biochemistry and Molecular Biology*, **62(1)**:29–44, 1997.

William N. Grundy, Timothy L. Bailey, Charles Elkan, and Michael E. Baker. Hidden Markov model analysis of motifs in steroid dehydrogenases and their homologs. *Biochemical and Biophysical Research Communications*, **231(3)**:760–766, 1997.

William N. Grundy, Timothy L. Bailey, Charles Elkan, and Michael E. Baker. Meta-MEME: Motif-based hidden Markov models of protein families. *Computer Applications in the Biosciences*, **13(4)**:397–406, 1997.

Alvaro E. Monge and Charles Elkan. An efficient domain-independent algorithm for detecting approximately duplicate database records. Published at the *SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD'97)*. Tucson, Arizona, May 1997.

Timothy L. Bailey, Michael E. Baker, Charles Elkan, and William N. Grundy. MEME, MAST, and Meta-MEME: New tools for motif discovery in protein sequences. In *Pattern Discovery in Biomolecular Data: Tools, Techniques, and Applications*, J. Wang, B. Shapiro and D. Shasha, editors. Oxford University Press, 1999.

Michael E. Baker, William N. Grundy, and Charles Elkan. Spinach CSP41, an mRNA-binding protein and ribonuclease, is homologous to nucleotide-sugar epimerases and hydroxysteroid dehydrogenases. *Biochemical and Biophysical Research Communications*, **248(2)**:250–254, 1998.

Michael E. Baker, William N. Grundy, and Charles Elkan. A common ancestor for a subunit in the mitochondrial proton-translocating NADH:ubiquinone oxidoreductase (complex I) and short-chain dehydrogenases/reductases. *Cellular and Molecular Life Sciences*, **55(3)**:450–455, 1999.

Fredrik Farnstrom, James Lewis, and Charles Elkan. Scalability for clustering algorithms revisited. *ACM SIGKDD Explorations*, **2(1)**:51–57, August 2000.

Charles Elkan. Paradoxes of fuzzy logic, revisited. *International Journal of Approximate Reasoning*, vol. 26, no. 2, pp. 153–155, 2001.

Koji Morikawa, Sameer Agarwal, Charles Elkan, and Garrison W. Cottrell. A taxonomy of computational and social learning. In *Proceedings of the Workshop on Developmental Embodied Cognition (DECO-2001)*, Edinburgh, July 2001, pp. 46–50.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML'01)*, July 2001, pp. 609–616.

Greg Hamerly and Charles Elkan. Bayesian approaches to failure prediction for disk drives. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML'01)*, July 2001, pp. 202–209.

Charles Elkan. The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth Inter-*

- national Joint Conference on Artificial Intelligence (IJCAI'01)*, August 2001, pp. 973–978.
- Charles Elkan. Magical thinking in data mining: Lessons from CoIL Challenge 2000. In *Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining (KDD'01)*, September 2001, pp. 426–431.
- Bianca Zadrozny and Charles Elkan. Learning and making decisions when costs and probabilities are both unknown. In *Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining (KDD'01)*, September 2001, pp. 204–213.
- Charles Elkan. Method and system for selecting documents by measuring document quality. United States Patent Application 20020055940, filed November 2, 2001.
- Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth International Conference on Knowledge Discovery and Data Mining (KDD'02)*, August 2002, pp. 694–699.
- Gordon F. Hughes, Joseph F. Murray, Kenneth Kreutz-Delgado, and Charles Elkan. Improved disk-drive failure warnings. *IEEE Transactions on Reliability*, vol. 51, no. 3, pp. 350–357, September 2002.
- Greg Hamerly and Charles Elkan. Alternatives to the k -means algorithm that find better clusterings. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM'02)*, November 2002, pp. 600–607.
- Charles Elkan. Using the triangle inequality to accelerate k -means. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML'03)*, Washington, DC, August 2003, pp. 147–153.
- Eric Wiewiora, Garrison Cottrell, and Charles Elkan. Principled methods for advising reinforcement learning agents. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML'03)*, Washington, DC, August 2003, pp. 792–799.
- David Kauchak and Charles Elkan. Learning rules to improve a machine translation system. In *Proceedings of the Fourteenth European Conference on Machine Learning (ECML'03)*, Dubrovnik, Croatia, September 2003, pp. 205–216.
- Greg Hamerly and Charles Elkan. Learning the k in k -means. In *Proceedings of the Seventeenth Annual Conference on Neural Information Processing Systems (NIPS)*, December 2003.
- Andrew Smith and Charles Elkan. A Bayesian network framework for reject inference. In *Proceedings of the Tenth International Conference on Knowledge Discovery and Data Mining (KDD'04)*, August 2004, pp. 286–295.
- David Kauchak, Joseph Smarr, and Charles Elkan. Sources of success for boosted wrapper induction. *Journal of Machine Learning*, vol. 5, pp. 499–527, 2004.
- Doug Turnbull and Charles Elkan. Fast recognition of musical genre using RBF networks. *IEEE Transactions on Data and Knowledge Engineering*, vol. 17, no. 4, pp. 580–584, 2005.
- Rasmus Madsen, David Kauchak, and Charles Elkan. Modeling word burstiness using the Dirichlet distribution. In *Proceedings of the Twenty-Second International Conference on Machine Learning (ICML'05)*, Bonn, Germany, August 2005.