# Andrew McCallum

**Home Address**
1439 Walnut Street
Pittsburgh, PA 15218
(412) 243-1884

**Work Address**
WhizBang Labs
4616 Henry Street
Pittsburgh, PA 15213
mccallum@cs.cmu.edu
http://www.cs.cmu.edu/~mccallum
(412) 414-7913 (mobile)
(412) 683-9132 (office)

**Education**

Post-doctoral Research Fellow, School of Computer Science, Carnegie Mellon University. 1996. With Sebastian Thrun and Tom Mitchell. Machine Learning for Text, Language Modeling. Reinforcement Learning, Intelligent Control for Office Space.

Ph.D. in Computer Science, University of Rochester, Rochester New York, 1995. Thesis title: "Reinforcement Learning with Selective Perception and Hidden State". Adviser: Dana Ballard. Nominated for ACM Best Thesis Award.

M.S. in Computer Science, University of Rochester, Rochester, New York, 1992.

B.A. in Computer Science (*summa cum laude*, Phi Beta Kappa), Dartmouth College, Hanover, New Hampshire, 1989.

High School. The North Carolina School of Science and Mathematics. 1985.

**Professional Positions**

Vice President of Research & Development; Director of Pittsburgh office at WhizBang! Labs, Inc. January 2000-present.

Recruited and hired more than 30 people, including 8 Ph.D. research scientists from AT&T Labs Research, Carnegie Mellon University, IBM T. J. Watson Research, Australian National University, Xerox Research Center Europe, University of Edinburgh.

As Vice President of Research & Development set research direction, identified key research problems and technical approaches, managed large group of research scientists, performed research and implementation, published research papers, gave invited talks at universities and conferences, co-architected object-oriented foundation classes for machine learning software libraries, lead two large research-related software projects.

As Director of 30-person lab, built and maintained lab work culture, organized and lead lab-wide meetings, performed project team assignments and organization, planned multi-day technical retreats for entire research and development group.

Adjunct Faculty, Carnegie Mellon University, Center for Automated Learning and Discovery, and Language Technology Institute. 1998–present.

Co-taught classes, gave guest lectures, advised graduate and undergraduate research projects, gave invited academic research presentations.

Research Scientist and Research Coordinator, Just Research (Justsystem Pittsburgh Research Center). January 1997–December 1999.

Spearheaded research in technology for statistical text processing.

Leader of the project that created "Cora," a domain-specific search engine over computer science research papers, still maintained and in wide use. (With Kamal Nigam, Kristie Seymore and Jason Rennie.) This project resulted in eight refereed publications.

Created *libbow*, an extensive software library for document classification, document clustering, information retrieval and information extraction written in C. Placed both libbow and Cora code under the GPL.

Recruited and advised undergraduate and graduate summer interns, including Dayne Freitag (CMU), Kamal Nigam (CMU), Kristie Seymore (CMU), Thomas Minka (MIT), and many undergraduates.

Research Fellow, Computer Science Department, University of Rochester. January–June 1996. With Dana Ballard. Psychophysics experiments with human eye-movements during the performance of highway driving tasks in virtual reality: testing models of short-term memory.

Consultant, Net Community Inc., Portland, Oregon, 1996.

Research Assistant, Computer Science Department, University of Rochester. Spring 1991, Spring 1992, Spring, Summer & Fall 1993, Summer & Fall 1994, Spring & Summer 1995. Speeding reinforcement learning with techniques from Kohonen Maps. Learning Hidden Markov Models for tasks with incomplete perception. Memory-based techniques for fast learning of context-dependent tasks. Visual routines and deictic strategies on images generated from realistic images of a 3-D world. Human eye-tracking while chasing cars in virtual reality.

Research Staff, Biomedical Information Communications Center, Oregon Health Sciences University. Summer 1992, Fall 1989–Summer 1990. Using Genetic Algorithms to categorize diseases from their symptoms. Designing graphical user interfaces and novel graphical representations for complex medical data.

Programmer, IBM, Research Triangle Park, North Carolina. Summer & Winter 1988.

**Grant Proposals**

Co-PI with William Cohen on proposal to NSF, Division of Undergraduate Education, program on National Science, Technology, Engineering, and Mathematics Education Digital Library (NSDL). Grant was written by Tom Mitchell, and would go to WhizBang Labs if awarded. "Automated Extraction from Web-Based Sources for NSDL Collection Development."

**Professional Service Activities**

Will give invited tutorial at NIPS, "Information Extraction from the World Wide Web." Fall 2002.

Grant proposal reviewer. NSF ITR. (Invited by Michael Lesk.) June 2002.

Co-Organizer of IJCAI workshop "Text Learning: Beyond Supervision,", with Lillian Lee, Tony Jebara and Kamal Nigam). 2001.

Co-Organizer of IJCAI Workshop "Machine Learning for Information Filtering," with Thorsten Joachims, Mehran Sahami and Lyle Ungar. 1999.

Co-Organizer of NIPS*98 Workshop "Integrating Supervised and Unsupervised Learning," with Rich Caruana, Virginia de Sa and Michael Kearns. 1998.

Co-Organizer of ICML/AAAI Workshop "Learning for Text Categorization," with Mehran Sahami, Mark Craven, and Thorsten Joachims. 1998.

On the editorial board of the Journal of Machine Learning Research.

Reviewer for many conferences and journals, including: Machine Learning Journal, Journal of Machine Learning Research (JMLR), International Conference on Machine Learning (ICML), Neural Information Processing Systems (NIPS), Conference of the American Association for Artificial Intelligence (AAAI), Journal of Artificial Intelligence Research (JAIR), International Joint Conference on Artificial Intelligence (IJCAI), IEEE Systems, Man and Cybernetics (SMC), ACM SIGIR, Uncertainty in Artificial Intelligence (UAI), CVPR.

Chief Maintainer of GNUstep Project, appointed by Richard Stallman, January 1996-August 1998. GNUstep is the Free Software Foundation's effort to implement NeXT Computer Inc.'s OpenStep object-oriented GUI standard. Made small contributions to gcc and emacs.

**Software Packages**

C4. Co-architected and helped develop first stages of the foundation classes for WhizBang's machine learning software libraries. System eventually grew to over 1 million lines of Java code.

Libbow. Software package for statistical text classification, clustering and retrieval. Has hundreds of users from all over the world. Nearly 100k lines of C. Released under GPL.

RLkit. Software library that makes it easy to test various reinforcement learning algorithms in different environments with different sensory-motor systems. Objective-C and Guile.

Persia. Toolkit for building virtual reality environments on SGI Onyx RealityEngine. The kit is based on SGI's Performer library and an embeddable Scheme interpreter.

**Honors**

Nominated for ACM Best Ph.D. Thesis Award. 1995.

Rufus Choates Scholar. Graduation *summa cum laude*. Dartmouth College. 1989.

IBM T. J. Watson Scholar. 1985.

**Graduate Department Service**

Graduate Student Representative, Computer Science Department, University of Rochester. Attended faculty meetings, representing the graduate students. 1993–1994.

Curriculum Committee, Computer Science Department, University of Rochester. 1992–1993.

Admissions Committee, Computer Science Department, University of Rochester. 1991–1992.

**Graduate and Undergraduate Student Research Projects Supervised**

David Blei (UC Berkeley Ph.D. student) and Drew Bagnell (CMU Ph.D. student). Scoped learning for integrated local/global learning with language and layout. 2002.

Kamal Nigam (CMU Ph.D. student). Document classification and clustering with unlabeled data and EM, (with Tom Mitchell). 1998-2001.

Kary Meyers (CMU Ph.D. student). Language modeling with the Dirichlet-Multinomial (with Larry Wasserman). 2000.

Nicholas Roy (CMU Ph.D. student). Active learning with statistical sampling. 2000.

Alexandrin "Sasha" Popescul (U Penn Ph.D. student). Distributional clustering for integrating site-level information into document classification. 2000.

Andrew Ng (UC Berkeley Ph.D. student). More accurate posterior probabilities for naive Bayes classifiers. 2000.

Thomas Minka (MIT Ph.D. student). Learning distance metrics classifiers trained with labeled and unlabeled data. 1999.

Kristie Seymore (CMU Ph.D. student). Information Extraction with hidden Markov models. 1998-2000.

Jason Rennie (CMU undergraduate, now Ph.D. student at MIT) Efficient spidering with Reinforcement Learning. 1998-1999. Document classification with large negative classes. 1997.

Doug Baker (CMU Ph.D. student). Word merging for document classification. 1998.

Jessica Bayliss (Rochester Ph.D. student). Reinforcement learning with short-term memory. 1996.

Josh Richardson (Rochester undergraduate). Building graphical simulations of highway environments. 1995.

Matt Hohlfeld (Rochester undergraduate). A Scheme interface to SGI's 'Performer' Graphics 3D Library. 1995.

Jennifer Alexander (Rochester undergrad). Reinforcement learning for a cheetah learning to stalk and pounce. 1994–1996.

Chandler Chao (Rochester undergraduate). Learning to perform tasks with deictic primitives on bitmap-vision perception. 1994.

Mia Stern (Rochester undergraduate). Reinforcement learning exploration strategies. 1992.

Jason Anderson (Rochester undergraduate). A machine learning simulator implemented in Objective C. 1992.

**Teaching Experience**

Instructor, "Machine Learning for Text Mining", CMU graduate-level course. Semester-long course with class project. Co-taught with Tom Mitchell, Fernando Pereira and William Cohen. 2000.

Guest Lecturer, "Machine Learning", CMU graduate-level course. (Main instructor was Roni Rosenfeld). 2001.

Guest Lecturer, "Machine Learning", CMU undergraduate-level course. (Main instructor was Sebastian Thrun). 1999, 2000.

Guest Lecturer, "Information Retrieval", CMU graduate-level course. 1998, 1999. (Main instructor was Yiming Yang.)

Instructor, "Introduction to Computer Systems", University of Rochester, (CSC104). Fall 1995, Fall 1991. Introductory programming class for non-majors. Taught two lectures per week for entire semester. Redesigned course lecture format for greater interactivity and student involvement; new design was then used by other instructors.

Teaching Assistant and Lecturer, "Artificial Intelligence", University of Rochester. Spring 1992. Lectured one fourth of the classes. Assistant for Professor Chris Brown.

Teaching and Project Assistant, Rochester First-Year Graduate Student Boot-Camp Course (CSC400), 1992. Organized machine learning component.

Undergraduate teaching assistant, "Artificial Intelligence", "Operating Systems", Dartmouth College. 1988–1989. For Professors Peter Sandon and Matt Bishop.

**Journal Publications**

Andrew McCallum, Kamal Nigam, Jason Rennie, Kristie Seymore. "Automating the Construction of Internet Portals with Machine Learning." Information Retrieval Journal, volume 3, pages 127-163. Kluwer. 2000.

Kamal Nigam, Andrew McCallum, Sebastian Thrun and Tom Mitchell. "Text Classification from Labeled and Unlabeled Documents using EM." Machine Learning Journal, 2000.

Andrew McCallum. "Hidden State and Reinforcement Learning with Instance-Based State Identification". In *IEEE Transactions on Systems, Man and Cybernetics*, (Special Issue on Robot Learning), 26 (3), June 1996

**Refereed Conference Publications**

John Lafferty, Andrew McCallum and Fernando Pereira. "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data." In *The Proceedings of the Eighteenth International Machine Learning Conference*, (ICML-2001).

Nick Roy and Andrew McCallum. "Toward Optimal Active Learning through Sampling Estimation of Error Reduction." In *The Proceedings of the Eighteenth International Machine Learning Conference*, (ICML-2001).

Andrew McCallum, Dayne Freitag and Fernando Pereira. "Maximum Entropy Markov Models for Information Extraction and Segmentation." In *The Proceedings of the Seventeenth International Machine Learning Conference*, (ICML-2000).

Andrew McCallum, Kamal Nigam and Lyle Ungar. "Efficient Clustering of High-Dimensional Data Sets with Application to Reference Matching" *The Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (KDD-2000).

Dayne Freitag and Andrew McCallum. "Information Extraction with HMM Structures Learned by Stochastic Optimization." *Proceedings of the Seventeenth National Conference on Artificial Intelligence*, (AAAI-2000).

Huan Chang, David Cohn and Andrew McCallum. "Creating Customized Authority Lists." In *The Proceedings of the Seventeenth International Machine Learning Conference*, (ICML-2000).

Andrew McCallum, Kamal Nigam, Jason Rennie and Kristie Seymore "Building Domain-Specific Search Engines with Machine Learning Techniques." *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, (IJCAI-99).

Jason Rennie and Andrew McCallum. "Using Reinforcement Learning to Spider the Web Efficiently." In *The Proceedings of the Sixteenth International Machine Learning Conference*, (ICML-99).

Andrew McCallum, Ronald Rosenfeld, Tom Mitchell and Andrew Ng. "Improving Text Classification by Shrinkage in a Hierarchy of Classes." In *The Proceedings of the Fifteenth International Machine Learning Conference*, (ICML-98).

Andrew McCallum and Kamal Nigam. "Employing EM in Pool-Based Active Learning for Text Classification." In *The Proceedings of the Fifteenth International Machine Learning Conference*, (ICML-98).

Doug Baker, Andrew McCallum "Distributional Clustering of Words for Text Classification." *Proceedings of the 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (SIGIR-98).

Kamal Nigam, Andrew McCallum, Sebastian Thrun and Tom Mitchell. "Learning to Classify Text from Labeled and Unlabeled Documents." AAAI-98.

Mark Craven, Dan DiPasquo, Dayne Freitag, Andrew McCallum, Tom Mitchell, Kamal Nigam, Sean Slattery. "Learning to Extract Knowledge from the World Wide Web." AAAI-98.

Andrew McCallum. "Learning to Use Selective Attention and Short-Term Memory in Sequential Tasks". *From Animals to Animats 5: Proceedings of the Fifth International Conference on Simulation of Adaptive Behavior*, (SAB-96).

Andrew McCallum. "Instance-Based Utile Distinctions for Reinforcement Learning". In *The Proceedings of the Twelfth International Machine Learning Conference*, (ICML-95). Lake Tahoe, Utah. Morgan Kaufmann Publishers, Inc. 1995.

Andrew McCallum. "Instance-Based State Identification for Reinforcement Learning". In *Advances of Neural Information Processing Systems*, (NIPS 7). 1995.

Andrew McCallum. "Overcoming Incomplete Perception with Utile Distinctions Memory". In *The Proceedings of the Tenth International Machine Learning Conference*, Amherst, Massachusetts. (ICML-93).

William Garrett, Ricardo Bianchini, Leonidas Kontothanassis, Andrew McCallum, Jeff Thomas, Robert Wisniewski and Michael L. Scott. "Linking Shared Segments". In *Proceedings of the USENIX Winter '93 Technical Conference*, pages 13–27, San Diego, CA, January 1993.

Andrew McCallum. "Using Transitional Proximity for Faster Reinforcement Learning". In *The Proceedings of the Ninth International Machine Learning Conference.* (ICML-92).

Andrew McCallum and Kent Spackman. "Using Genetic Algorithms to Learning Disjunctive Rules from Examples". In *The Proceedings of the Seventh International Machine Learning Conference.* (ICML-90).

**Invited Articles**

Dallan Quass, Andrew McCallum, William Cohen. "Unlocking the Information in Text." The Future of Software, Winter 2000/2001.

William Cohen, Andrew McCallum, Dallan Quass. "Learning to Understand the Web." IEEE Data Engineering Bulletin. September 2000, Vol. 23, No. 3. Pages 17-24.

**Refereed Workshop Publications**

Andrew McCallum. "Multi-Label Text Classification with a Mixture Model Trained by EM." AAAI'99 Workshop on Text Learning.

Kamal Nigam, John Lafferty, Andrew McCallum. "Using Maximum Entropy for Text Classification." IJCAI'99 Workshop on Information Filtering.

Dayne Freitag and Andrew McCallum "Information Extraction with HMMs and Shrinkage. " AAAI'99 Workshop on Machine Learning for Information Extraction.

Kristie Seymore, Andrew McCallum, Roni Rosenfeld "Learning Hidden Markov Model Structure for Information Extraction." AAAI'99 Workshop on Machine Learning for Information Extraction.

Andrew McCallum and Kamal Nigam. "A Comparison of Event Models for Naive Bayes Text Classification." AAAI-98 Workshop on "Learning for Text Categorization".

Andrew McCallum. "Reduced Training Time for Reinforcement Learning with Hidden State". In *The Proceedings of the Eleventh International Machine Learning Workshop, "Reinforcement Learning"*, Rutgers University. 1994.

Andrew McCallum. "Short-Term Memory in Visual Routines". In *The Working Notes of the AAAI Spring Symposium Series, "Toward Physical Interactions and Manipulation"*, Stanford University. 1994.

Andrew McCallum. “Short-Term Memory for Visual Routines”. In *The Proceedings of the Intelligent Robotic Systems Workshop, (IRS’94).* Grenoble, France. 1994.

**Submitted Papers**

David Blei, Drew Bagnell and Andrew McCallum. “Learning with Scope, with Application to Information Extraction and Classification.” Submitted to UAI-2002.

**Unrefereed Papers**

Doug Baker, Thomas Hoffman, Andrew McCallum and Yiming Yang. “A Hierarchical Probabilistic Model for Novelty Detection in Text.”

Andrew McCallum. “Reinforcement Learning with Selective Perception and Hidden State”, (Ph.D. Thesis). Department of Computer Science, University of Rochester. December 1995.

Andrew McCallum. “Persia User’s Manual—A Scheme Interface to SGI’s Performer 3D Graphics Library”. Computer Science Department, University of Rochester. 1995.

Andrew McCallum. “Utile Suffix Memory for Reinforcement Learning with Hidden State”. TR 549, Department of Computer Science, University of Rochester. December 1994.

Andrew McCallum. “First Results with Instance-Based State Identification”. TR 502, Department of Computer Science, University of Rochester. 1994.

Andrew McCallum. “Learning with Incomplete Selective Perception”, (Thesis proposal). TR 453, Department of Computer Science, University of Rochester. March 1993.

Andrew McCallum. “First Results with Utile Distinctions Memory for Reinforcement Learning”. TR 446, Department of Computer Science, University of Rochester. December 1992.

William Garrett, Ricardo Bianchini, Leonidas Kontothanassis, Andrew McCallum, Jeff Thomas, Robert Wisniewski and Michael L. Scott. “Dynamic Sharing and Backward Compatibility on 64-Bit Machines”. TR 418, Department of Computer Science, University of Rochester. April 1992.

Andrew McCallum. “Using Transitional Proximity for Faster Reinforcement Learning”. Department of Computer Science, University of Rochester. 1992.

Andrew McCallum. “Faster Reinforcement Learning with On-Line Value Iteration”. In *The Proceedings of the Buffalo Graduate Conference on Computer Science*, SUNY Buffalo. 1992.

**Invited Talks and Presentations**

Google. (Contact: Peter Norvig) “Turning the Web into a Knowledge Base: Information Extraction with Finite State Models.” May 2002.

University of Texas at Austin. (Contact: Ray Mooney) “Information Extraction with Finite State Models.” February 2002.

IJCAI-2001 Workshop on Adaptive Text Extraction and Mining. Invited talk. (Contact: Nicholas Kushmerick) "Information Extraction with Machine Learning". Ralph Grisham gave the other invited talk. August 2001.

AT&T Shannon Labs. (Contact: Fernando Pereira and Michael Kearns) "Automatically Building Internet Portals using Machine Learning". September 1999.

IBM Almaden Research Lab. (Contact: Shivakumar Vaithaynathan) "Automatically Building Domain-Specific Search Engines using Machine Learning". March 1999.

Stanford University, Computer Science Department. (Contact: Pat Langley and Daphne Koller) "Automatically Building Domain-Specific Search Engines using Machine Learning". March 1999.

SRI. (Contact: Moises Goldszmidt and Andreas Stolcke) "Automatically Building Domain-Specific Search Engines using Machine Learning". March 1999.

MIT AI Lab. (Contact: Paul Viola) "Text Classification with Limited Labeled Data". November 1998.

MIT Media Lab. (Contact: Pattie Maes and Tony Jebara) "Text Classification with Limited Labeled Data". November 1998.

Carnegie Mellon University, AI Seminar. (Contact: Tai Sing Lee) "Two Methods for Improving Text Classification when there is Sparse Training Data". April 1998.

Media Lab, MIT. Invited talk. "Learning Visual Routines for Highway Driving—Goldilocks meets Reinforcement Learning". November 1996.

AAAI Fall Symposium, (Symposium on 'Learning Complex Behaviors in Adaptive Intelligent Systems'). "Learning Task-Relevant State Spaces with a Utile Distinction Test". November 1996.

Simulation of Adaptive Behavior. "Learning to Use Selective Attention and Short-Term Memory in Sequential Tasks". September 1996.

Center for Visual Science, Department of Psychology, University of Rochester. Invited talk. "A Model for Learning Visual Routines with Short Term Memory". April 1996.

Carnegie Mellon University, Reinforcement Learning Group. Invited talk. "Addressing Selective Attention and Hidden State with Utile Distinctions in Feature-Space and History-Space". April 1996.

SUNY Geneseo, Department of Computer Science. Invited talk. "Learning Where to Look When Driving on a Crowded Highway". March 1996.

Carnegie Mellon University, Reinforcement Learning Group. Invited talk. "First Results with Instance-Based State Identification for Reinforcement Learning". April 1994.

AAAI Spring Symposium, ('Toward Physical Interaction and Manipulation'), Stanford University. "Short-Term Memory in Visual Routines". March 1994.

Xerox PARC. Invited talk. "A Nearest-Neighbor Approach to Short-Term Context While Performing Tasks with Visual Routines". March 1994.

Brown University. Invited talk. "Learning Hidden Markov Models for Short-Term Memory in Reinforcement Learning". June 1993.

University of Buffalo Graduate Conference on Computer Science. "Speeding Reinforcement Learning with On-line Value Iteration". Fall 1992.

**Languages**

English: Native language

French: 9 years of experience; lived six months in Toulouse, France

Greek: Limited knowledge

**Personal Information**

Born: April 2, 1967

Nationality: United States of America

Marital Status: Married, two children

**Reference Contacts**

Tom M. Mitchell, Fredkin Professor of AI and Learning, Director, Center for Automated Learning and Discovery, School of Computer Science, Carnegie Mellon University. `tom.mitchell@cs.cmu.edu`

Fernando Pereira, Andrew and Debra Rachleff Professor, Chair Department of Computer and Information Science, University of Pennsylvania. `pereira@cis.upenn.edu`

Lyle Ungar, Associate Professor of Computer and Information Science, Associate Professor of Chemical Engineering, Associate Professor of Systems Engineering, in the School of Engineering and Applied Science; and Associate Professor of Operations and Information Management in the Wharton School at the University of Pennsylvania. `ungar@cis.upenn.edu`

Supplemental References

Raymond Mooney, Professor, Computer Science Department, University of Texas at Austin. `mooney@cs.utexas.edu`

Supplemental References regarding Teaching

Roni Rosenfeld, Associate Professor, School of Computer Science (Language Technology Institute, Center for Automated Learning and Discover, Computer Science Department), Carnegie Mellon University. `roni@cs.cmu.edu`

Supplemental References regarding Student Advising

Andrew Ng, Ph.D. UC Berkeley. `ang@cs.berkeley.edu`

Kamal Nigam, Ph.D. Carnegie Mellon University. `knigam@cs.cmu.edu`

May, 2002