David Andrzejewski

http://people.llnl.gov/andrzejewski1
✉ david.andrzej@gmail.com
510-730-0362
Lawrence Livermore National Laboratory
Livermore, CA 94551-0808

Education

2007–2010 **PhD**, *University of Wisconsin–Madison*, Madison, WI.
Computer Sciences
Research focus: Machine Learning
Advisors: Mark Craven and Xiaojin Zhu
Thesis: *Incorporating Domain Knowledge in Latent Topic Models*

2005–2007 **MS**, *University of Wisconsin–Madison*, Madison, WI.
Computer Sciences

2000–2005 **BS**, *University of Wisconsin–Madison*, Madison, WI.
Computer Engineering, Mathematics, Computer Sciences

Publications

David Andrzejewski and David Buttler. Latent topic feedback for information retrieval. In *KDD '11: Proceedings of the 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, 2011. (8% of submissions accepted for oral presentation).

David Andrzejewski, Xiaojin Zhu, Mark Craven, and Benjamin Recht. A framework for incorporating general domain knowledge into latent Dirichlet allocation using first-order logic. In *IJCAI '11: Proceedings of the 22nd International Joint Conference on Artificial Intelligence*. AAAI Press, 2011. (17% of submissions accepted).

David Andrzejewski, David G. Stork, Xiaojin Zhu, and Ron Spronk. Inferring compositional style in the neo-plastic paintings of Piet Mondrian by machine learning. In David G. Stork, Jim Coddington, and Anna Bentkowska-Kafel, editors, *Computer Vision and Image Analysis of Art*, volume 7531, page 75310G. SPIE, 2010.

David Andrzejewski and Xiaojin Zhu. Latent Dirichlet allocation with topic-in-set knowledge. In *SemiSupLearn '09: Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, pages 43–48. Association for Computational Linguistics, 2009.

David Andrzejewski, Xiaojin Zhu, and Mark Craven. Incorporating domain knowledge into topic modeling via Dirichlet forest priors. In *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, pages 25–32. Association for Computing Machinery, 2009. (25% of submissions accepted).

Andrew B. Goldberg, Nathanael Fillmore, David Andrzejewski, Zhiting Xu, Bryan Gibson, and Xiaojin Zhu. May all your wishes come true: a study of wishes and how to recognize them. In *HLT-NAACL 2009: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 263–271. Association for Computational Linguistics, 2009. (29% of submissions accepted).

David Andrzejewski, Anne Mulhern, Ben Liblit, and Xiaojin Zhu. Statistical debugging using latent topic models. In *ECML '07: Proceedings of the 18th European conference on Machine Learning*, pages 6–17. Springer-Verlag, 2007. (9% of submissions accepted).

Xiaojin Zhu, Andrew B. Goldberg, Jurgen Van Gael, and David Andrzejewski. Improving diversity in ranking using absorbing random walks. In *HLT-NAACL 2007: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 97–104. The Association for Computational Linguistics, 2007. (24% of submissions accepted).

Andrew B. Goldberg, David Andrzejewski, Jurgen Van Gael, Burr Settles, Xiaojin Zhu, and Mark Craven. Ranking biomedical passages for relevance and diversity: University of Wisconsin–Madison at TREC Genomics 2006. In Ellen M. Voorhees and Lori P. Buckland, editors, *TREC 2006: Proceedings of the Fifteenth Text REtrieval Conference*, volume Special Publication 500-272. National Institute of Standards and Technology (NIST), 2006.

Lam Raga A. Markely, David Andrzejewski, Erick Butzlaff, and Alexander Kiselev. Enhancement of combustion by drift in a coupled reaction-diffusion model. *Communications in Mathematical Sciences*, 4(1):213–225, 2006.

Professional experience

Industry

Fall 2010 to present — **Postdoctoral Research Staff Member**,
Lawrence Livermore National Laboratory, Livermore, CA.
Apply statistical modeling to knowledge discovery in text corpora

Summer 2008 — **Research Intern**, *Microsoft Research*, Redmond, WA.
Mentored by Alice Zheng in the Knowledge Tools research group
Developed analysis techniques and software (patent applied for)
Collaborated with product team to deliver high-impact internal tool

Summer 2004 Summer 2005 — **Research & Development Engineer**, *GE Healthcare*, Madison, WI.
Developed software for drug identification system prototype (patent applied for)
Created clinical pharmacokinetic modeling system prototype

Fall 2003 — **Software Engineer**, *GE Healthcare*, Menomonee Falls, WI.
Developed testing tools for cardiac image analysis software
Identified and resolved bugs in cardiac image analysis software

Academic

2008-2010 — **Research Assistant (Professors Mark Craven and Xiaojin Zhu)**,
UW–Madison, Madison, WI.
Knowledge-augmented topic models
Developed new latent topic models to allow prior knowledge and user feedback
Proposed, implemented, and conducted experiments on new models and techniques

2005-2008 — **Computation and Informatics in Biology and Medicine predoctoral trainee**,
UW–Madison, Madison, WI.
Biomedical text mining
Applied text mining to assist biological researchers in understanding experimental results
Incorporated structured knowledge sources into biomedical text analysis

Fall 2004 — **Undergraduate Researcher**, *UW–Madison*, Madison, WI.
Conducted computational experiments on reaction-diffusion equations

Professional Activities

Talks

- Machine Learning: An Overview. LLNL Global Security Tech Talks (May 2011)
- Inferring compositional style in the neo-plastic paintings of Piet Mondrian by machine learning. SPIE Computer Vision and Image Analysis of Art (January 2010)
- Incorporating domain knowledge into topic modeling via Dirichlet forest priors. International Conference on Machine Learning (June 2009)
- Data analysis with latent topic models: genes, bugs, and art. UW-Madison CIBM Seminar (March 2008)
- Statistical debugging using latent topic models. European Conference on Machine Learning (September 2007)
- Extracting information from the scientific literature to aid in uncovering gene-regulatory networks. NSF Symposium on Cyber-Enabled Discovery and Innovation (September 2007)

Service

Reviewer — *Neural Information Processing Systems* (NIPS 2011), *International Joint Conferences on Artificial Intelligence* (IJCAI 2011), *Scaling Up Machine Learning* (book chapter), *IEEE International Conference on Development and Learning* (ICDL 2010), *International Conference on Machine Learning* (ICML 2010), *Journal of Computer Science and Technology*, *Journal of the American Society for Information Science and Technology*, *Open Information Systems Journal*, *Machine Learning*

Organizer — Math for Machine Learning reading group (Spring 2010)

Coordinator — AI reading group (Fall 2009–Spring 2010)

Volunteer — UW–Madison Computer Sciences graduate admissions committee (2009)

Other professional accomplishments

Released research software

- **Dirichlet Forest LDA** (Python C++ extension module)
  Topic model with user-defined constraints over words (Must-Link and Cannot-Link)
  More than 300 tracked downloads on `mloss.org`
- **Delta LDA** (Python C extension module)
  Topic model with special "restricted" topics that only appear in certain documents
  More than 600 tracked downloads on `mloss.org`
- See other projects at
  `http://pages.cs.wisc.edu/~andrzeje/software`
  `https://github.com/davidandrzej`

Patent applications

- **System and method of drug identification through radio frequency identification (RFID)**
  United States Patent Application (11/465993)
  Ronald Makin, Kyle Jansson, Silas Zirn, David Andrzejewski,Timothy Flink

- **Visualization tool for system tracing infrastructure events**
  United States Patent Application (12/485726)
  Alice X. Zheng, Trishul A. Chilimbi, Shuo-Hsien Hsiao, Danyel A. Fisher, David M. Andrzejewski

Awards

- ICML student travel award (2009)
- Computation and Informatics in Biology and Medicine (CIBM) traineeship (2005-2008 NIH/NLM doctoral training award)

Additional information

Programming languages	Python, Clojure, Java, C, MATLAB, Scala, C++, R, C#
Scientific Python	NumPy, SciPy, matplotlib, scikits.learn, multiprocessing, NetworkX, Cython, CVXOPT
Selected technologies	SQL, HBase, MongoDB, Hadoop, Condor High Throughput Computing
Selected coursework	Machine Learning, Natural Language Processing, Computer Vision, Nonlinear Optimization, Database Management Systems, Approximation Algorithms, Statistical Learning Theory, Genetics, Biochemistry, Prokaryotic Molecular Biology, Bioinformatics, Statistics for Biosciences