David M. Andrzejewski

Professional experience

Industry

2015 to present
Engineering Manager, Unified Logs and Metrics (backend),
Sumo Logic, Redwood City, CA.
Team responsible for new time-series data analysis platform

2014 - 2015
Tech Lead / Manager, Data Sciences Engineering,
Sumo Logic, Redwood City, CA.
Lead the development of advanced analytics for machine-generated log data

2013 - 2014
Lead Data Sciences Engineer,
Sumo Logic, Redwood City, CA.

2011 - 2013
Data Sciences Engineer,
Sumo Logic, Mountain View, CA.

2010 - 2011
Postdoctoral Research Staff Member,
Lawrence Livermore National Laboratory, Livermore, CA.
Applied statistical modeling to knowledge discovery in text corpora

2008
Research Intern, *Microsoft Research*, Redmond, WA.
Developed analysis techniques for investigating system performance anomalies
Collaborated with product team to deliver high-impact internal tool

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

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

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

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

Keith Stevens, Philip Kegelmeyer, **David Andrzejewski**, and David Buttler. Exploring topic coherence over many models and many topics. In *EMNLP-CoNLL 2012: Conference on Empirical Methods in Natural Language Processing and Natural Language Learning*. Association for Computational Linguistics, 2012. (18% of submissions accepted for oral presentation).

David Andrzejewski. Accelerated gibbs sampling for infinite sparse factor analysis. In *Lawrence Livermore National Laboratory Technical Report (LLNL-TR-499647)*, 2011.

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.

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 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).

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 Activities

Talks

- Sumo Logic Analytics Platform Integration. Illuminate - Sumo Logic User Conference (September 2017)
- Economical machine learning via functional programming. Big Data Scala by the Bay, Oakland (August 2015)
- Graph mining for log data. Strata + Hadoop World, San Jose (February 2015)
- Mining human-scale insights from log data with machine learning. Orange County Big Data Meetup (September 2014)
- Machine learning for machine data. Strata Conference, Santa Clara (February 2014)
- Scala type classes and machine learning. Bay Area Scala Enthusiasts lightning talk (January 2013)
- Latent Topic Feedback for Information Retrieval. ACM SIGKDD Conference on Knowledge Discovery and Data Mining (August 2011)
- A Framework for Incorporating General Domain Knowledge into Latent Dirichlet Allocation using First-Order Logic. International Joint Conference on Artificial Intelligence (July 2011)
- 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

Co-organizer   SF Bay Area Machine Learning Meetup (> 6000 members)

Reviewer	*International Conference on Artificial Intelligence and Statistics (AISTATS 2017), Hands-on Machine Learning with Scikit-Learn and TensorFlow (O'Reilly Media), Fundamentals of Deep Learning (O'Reilly Media), Fundamentals of Deep Learning (O'Reilly Media), Reactive Machine Learning Systems (Manning Publications), SoftwareMining-2015 (IEEE ASE 2015 Workshop), SoftwareMining-2014 (ICDM 2014 Workshop), Learning Spark (O'Reilly Media), Conference on Empirical Methods on Natural Language Processing (EMNLP 2014, 2015), Strata + Hadoop World New York 2014-2017, Strata + Hadoop World 2014-2017, Topic Models: Computation, Application, and Evaluation (NIPS 2013 workshop), Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013, 2012), International Conference on Machine Learning (ICML 2016, 2015, 2014, 2013, 2010), SIAM International Conference on Data Mining (SDM 2013), Data Mining and Knowledge Discovery, SoftwareMining-2013 (ASE 2013 Workshop), SoftwareMining-2012 (KDD 2012 Workshop), Neural Information Processing Systems (NIPS 2017, 2015, 2011), International Joint Conferences on Artificial Intelligence (IJCAI 2015, 2011), Scaling Up Machine Learning (book chapter), IEEE International Conference on Development and Learning (ICDL 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

Patents and 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 (US8464221)
  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)