Jay Pujara

CONTACT INFORMATION

3228 A.V. Williams Building
Computer Science Department
University of Maryland
College Park, MD 20742

202 567 7885
jay@cs.umd.edu
http://www.cs.umd.edu/~jay

RESEARCH INTERESTS

Machine Learning, Data Science, Scalable Learning, Statistical Relational Learning, Knowledge Graph Construction, Entity Resolution, Efficient Prediction, Cascade Classifiers, Active Learning.

EDUCATION

Ph.D. in Computer Science

Computer Science Department
University of Maryland, College Park, MD (Fall 2010 – Present)

- *Research Area*: Artificial Intelligence and Machine Learning
- *Dissertation Topic*: Scalable Construction of Knowledge Graphs
- *Candidacy Date*: Fall 2013
- *Expected Graduation*: Fall 2015
- *Advisor*: Prof. Lise Getoor
- *GPA*: 4.0

Visiting Student

Jack Baskin School of Engineering
University of California, Santa Cruz, CA (Spring 2014 – Present)

- *Advisor*: Prof. Lise Getoor

Visiting Research Scholar

Machine Learning Department
Carnegie Mellon University, Pittsburgh, PA (Fall 2014 – Winter 2015)

- *Advisor*: Prof. William Cohen

Master of Science in Computer Science

School of Computer Science
Carnegie Mellon University, Pittsburgh, PA (Summer 2004 – Spring 2005)

- *Dissertation Title*: Fundamental Properties of Feature Selection in fMRI Data
- *Advisor*: Prof. Tom Mitchell
- *GPA*: 3.8

Bachelor of Science in Computer Science

School of Computer Science
Carnegie Mellon University, Pittsburgh, PA (Fall 2000 – Spring 2004)

- *Distinction*: Graduated with University Honors and College Honors
- *Minors*: Robotics, Mathematical Sciences, and Logic and Computation.
- *Thesis Title*: Machine Learning Classification of fMRI Data
- *Advisor*: Prof. Tom Mitchell
- *GPA*: 3.6 (3.7 in Major)

Bachelor of Science in Cognitive Science

School of Humanities and Social Sciences
Carnegie Mellon University, Pittsburgh, PA (Fall 2001 – Spring 2004)

- *Distinction*: Graduated with University Honors
- *GPA*: 3.6 (4.0 in Major)

Bachelor of Science in Electrical and Computer Engineering

Carnegie Institute of Technology
Carnegie Mellon University, Pittsburgh, PA (Fall 2001 – Spring 2004)

- *Distinction*: Graduated with University Honors
- *GPA*: 3.6 (3.4 in Major)

Honors

Best Student Paper Award, International Semantic Web Conference, ISWC 2013.
Best Paper Award, Collaboration, Electronic Messaging, Anti-abuse, and Spam, CEAS 2011.
Dean's Fellowship Award, University of Maryland, College Park, 2010-2012.
Yahoo! FREP Award, "Active Feature Acquisition", Advisor: Martin Zinkevich, 2010-2012.
Lemonade Stand Multi-agent Competition, 2nd place, 2009; 3rd place, 2010.

Publications

Journal Articles

Jay Pujara, Hui Miao, Lise Getoor, William Cohen, "Using Semantics and Statistics to Turn Data into Knowledge." *AI Magazine* 36(1), 2015.

Conference Proceedings

Jay Pujara, Ben London, and Lise Getoor. "Budgeted Online Collective Inference." *In Conference on Uncertainty in Artificial Intelligence (UAI)*, Amsterdam, Netherlands, 2015.

Adam Grycner, Gerhard Weikum, **Jay Pujara**, James Foulds, Lise Getoor, "RELLY: Inferring Hypernym Relationships Between Relational Phrases." *In Emperical Methods in Natural Language Processing (EMNLP)*, Lisbon, Portugal, 2015.

Jay Pujara, Hui Miao, Lise Getoor, William Cohen, "Knowledge Graph Identification." *International Semantic Web Conference (ISWC)*, Sydney, Australia, 2013. [winner of Best Student Paper award]

Jay Pujara, Hal Daumé III, Lise Getoor, "Using Classifier Cascades for Scalable E-Mail Classification." *In Proceedings of the 8th Conference on Collaboration, Electronic Messaging, Anti-abuse, and Spam (CEAS)*, Perth, Australia, 2011. [winner of Best Paper award]

Refereed Workshops

Jay Pujara, Ben London, Lise Getoor and William Cohen "Online Inference for Knowledge Graph Construction." *UAI 2015 Workshop on Statistical Relational Artificial Intelligence (StaRAI)*, Amsterdam, Netherlands, 2015.

Jay Pujara, Lise Getoor, "Building Dynamic Knowledge Graphs." *NIPS 2014 Workshop on Au-*

tomated Knowledge Base Construction, Montreal, QC, 2014.

Jay Pujara, Lise Getoor, "Building Dynamic Knowledge Graphs." *NIPS 2014 Workshop on Automated Knowledge Base Construction*, Montreal, QC, 2014.

Adam Grycner, Gerhard Weikum, **Jay Pujara**, James Foulds, Lise Getoor, "A Unified Probabilistic Approach for Semantic Clustering of Relational Phrases." *NIPS 2014 Workshop on Automated Knowledge Base Construction*, Montreal, QC, 2014.

Jay Pujara, Kevin Murphy, Xin Luna Dong, Curtis Janssen "Probabilistic Models for Collective Entity Resolution Between Knowledge Graphs." *Bay Area Machine Learning Symposium (BayLearn)*, Berkeley, CA, 2014.

Jay Pujara, Hui Miao, Lise Getoor, William Cohen, "Large-Scale Knowledge Graph Identification using PSL." *AAAI Symposium on Semantics for Big Data*, Arlington, VA, 2013. [contributed talk]

Jay Pujara, Hui Miao, Lise Getoor, William Cohen, "Ontology-Aware Partitioning for Knowledge Graph Identification." *CIKM 2013 Workshop on Automated Knowledge Base Construction*, Burlingame, CA, 2013. [spotlight talk]

Jay Pujara, Hui Miao, Lise Getoor, "Joint Judgments with a Budget: Strategies for Reducing the Cost of Inference." *ICML 2013 Workshop on Machine Learning with Test-Time Budgets*, Atlanta, GA, 2013.

Jay Pujara, Hui Miao, Lise Getoor, William Cohen, "Large-Scale Knowledge Graph Identification using PSL." *ICML 2013 Workshop on Structured Learning*, Atlanta, GA, 2013.

Jay Pujara, Pete Skomoroch, "Large-Scale Hierarchical Topic Models." *NIPS 2012 Workshop on Big Learning*, Lake Tahoe, NV, 2012.

Bert Huang, Stephen H. Bach, Eric Norris, **Jay Pujara**, Lise Getoor. "Social Group Modeling with Probabilistic Soft Logic." *NIPS 2012 Workshop - Social Network and Social Media Analysis: Methods, Models, and Applications*, Lake Tahoe, NV, 2012.

Jay Pujara, Ben London, Lise Getoor, "Reducing Label Cost by Combining Feature Labels and Crowdsourcing." *ICML 2011 Workshop on Combining Learning Strategies to Reduce Label Cost*, Seattle, WA, 2011. [selected for contributed talk]

Leo Claudino, Sameh Khamis, Ran Liu, Ben London, **Jay Pujara**, Catherine Plaisant, Ben Shneiderman, "Facilitating Medication Reconciliation with Animation and Spatial Layout." *In Proceedings of the Workshop on Interactive Healthcare Systems (WISH2011)*, Washington, DC, 2011.

Jay Pujara and Lise Getoor, "Coarse-to-Fine, Cost-Sensitive Classification of E-Mail." *NIPS 2010 Workshop on Coarse-to-Fine Processing*, Whistler, BC, Canada, 2010. [spotlight talk]

Patents

Jay Pujara, "Real-time Ad-Hoc Spam Filtering of E-Mail." Patent 8,069,128; awarded 2011.

Ke Wei, Hao Zheng, **Jay Pujara**, "Employing pixel density to detect a spam image." Patent 7,882,177; awarded 2011.

Jaesik Choi, **Jay Pujara**, Vishwanath Ramarao, Ke Wei, "Identifying IP addresses for spammers."

Patent 7,849,146; awarded 2010.

Research Experience

University of Maryland, College Park, MD

Research Assistant **Fall 2010 – Present**

My research focused on scalable machine learning to address scenarios where billions of predictions are necessary in a limited amount of time and large, noisy corpora of training data are available.

Dissertation topic: *Scalable Construction of Knowledge Graphs*
Abstract: In the past decade, systems that extract information from millions of Internet documents have become commonplace. One approach to better understand and organize this vast amount of information is constructing a knowledge graph – a structured knowledge base that captures entities, their attributes and the relationships between them. I examine the problem of constructing a knowledge graph from the noisy output of an information extraction system. My work introduces an approach called knowledge graph identification (KGI), which resolves the entities, attributes and relationships in the knowledge graph by incorporating uncertain extractions from multiple sources, entity co-references, and ontological constraints. I define a probability distribution over possible knowledge graphs and infer the most probable knowledge graph. I extend my model for knowledge graph construction by introducing a powerful system for entity resolution, developing a partitioning approach to distribute KGI over many machines, defining a problem formulation for the streaming setting, and developing techniques for online inference in probabilistic graphical models enabling streaming knowledge graph construction.

Google Inc, Mountain View, CA

Software Engineering Intern, Knowledge Vault **Summer 2014**

The Google Knowledge Graph and Knowledge Vault are built using structured knowledge from extraction tools and third-party feeds. A key challenge in integrating this knowledge is reconciliing the entities between different sources to determine whether to modify an existing entity or add a new entity. My goal was to build a collective entity resolution system that uses the relational structure of entities in the knowledge graph to improve entity resolution performance while scaling to millions of entities. My project involved using Google infrastructure to extract relational data from the Knowledge Graph and other sources, generate entity resolution candidates, and experiment with different resolution models. I used a probabilistic graphical modeling framework (PSL) to build a collective model that outperformed the non-collective baseline. This work was presented at the BayLearn symposium in 2014.

LinkedIn Corporation, Mountain View, CA

Data Scientist Intern, Skills **Summer 2012**

LinkedIn extracts and generates data about skills its members possess, such as Machine Learning and Computer Science. I undertook a summer research project to find structure in the skills data. The approaches I experimented with included spectral clustering of skill co-occurrences, pairwise predictors for parent-child relationships and hierarchical topic models. Hierarchical topic models proved to be the most promising of the approaches, so I implemented a novel method for creating hierarchical topic models on MapReduce, suitable for tens of millions of documents and tens of thousands of terms. This work was presented at the NIPS BigLearning workshop in 2012.

Carnegie Mellon University, Pittsburgh, PA

Research Assistant **Fall 2003 – Spring 2005**

Machine Learning Approaches for fMRI Data: Functional Magnetic Resonance Imaging has enabled neuroscientists to measure brain activity at millimeter resolution, resulting in data from tens of

thousands of regions of the brain (voxels) on a per-second basis. Traditional analysis often relies on averaging over time and across brain regions, missing some of the unique opportunities this data provides. Machine learning offers the ability to perform analysis at a much finer scale. In order to build classifiers that predicted functional differences between tasks, I initially focused on statistically-sound models for comparing fMRI data from disparate experiments. Some interesting results from these classification experiments resulted in my Masters research: I designed a statistical model of the voxel measurements taken in fMRI experiments, analyzed the parameters to the model that best explained our empirical results, and using these models created a better feature selection procedure for fMRI data.

Teaching Experience

University of Maryland, College Park, MD

Teaching Assistant **Fall 2011**

CMSC 421: Introduction to Artificial Intelligence

Professional Experience

Google Inc, Mountain View, CA

Software Engineering Intern, Knowledge Vault **Summer 2014**

Scalable collective entity reconciliation for the Knowledge Graph and structured knowledge sources.

LinkedIn Corp., Mountain View, CA

Data Science Intern, Skills **Summer 2012**

Hadoop implementation of hierarchical topic models to discover the structure of LinkedIn skills.

Yahoo! Inc, Sunnyvale, CA [remote]

Data Researcher, Yahoo! Mail **Fall 2010 – Spring 2012**

Research on cost-sensitive approaches for e-mail classification.

Yahoo! Inc, Sunnyvale, CA

Senior Software Engineer, Yahoo! Mail **Fall 2007 – Fall 2010**

Software Engineer, Yahoo! Mail **Fall 2006 – Fall 2007**

As a member of the Yahoo! Mail AntiSpam team, my key contributions included:

- Implementing components of the Antispam data distribution platform that handle billions of transactions daily, including a robust caching tier and an interface to backend data sources.
- Improving systems that track the reputation of URLs, IP addresses, message tokens, and users.
- Moving Antispam data and analysis to a Hadoop cluster, and subsequently creating scripts and analytics that identified spam trends on an hourly basis.
- Redesigning, rewriting, and improving the system that handles user feedback on messages.
- Leading the Trusted User Project developed to track user reputation across Yahoo! properties.

Oracle Corp, Redwood Shores, CA

Member of Technical Staff, Business Intelligence **Fall 2005 – Fall 2006**

Contributed to a new Business Intelligence platform that leveraged the database backend for analytics. Implemented caching systems that allowed analytics queries to be fulfilled from cached data in middleware rather than making expensive database queries.

University of Pittsburgh, Pittsburgh, PA

Research Programmer, Learning R&D Center **Summer 2003**

I worked with Professor Walt Schneider to implement a proposed cognitive model in Matlab and ran experiments to evaluate the performance and plausibility of the model.

InternalDrive Corporation, Stanford, CA

Camp Instructor, Game Programming and C++ **Summer 2002**

I taught courses to children ages 8-18 on Game Programming and C++.

Carnegie Mellon University, Pittsburgh, PA

Research Programmer, Robotics Institute **Summer 2001**

Working with Dr. Henry Schneiderman, I produced a web application demonstrating face detection algorithms developed by the research group. I also built a data crawling utility to collect images from web pages and a simple ground-truthing application for marking faces in these images.

West Virginia State Legislature, Charleston, WV

Web Designer and Developer **Spring 2000**

I programmed ColdFusion/SQL queries to interface Access databases for dynamic webpages.

Activities

Member, Computer Science Executive Council
University of Maryland, College Park Fall 2010-Present

- Volunteer at department social activities such as weekly coffee hour and semester picnic

President, Computer Science Executive Council
University of Maryland, College Park Fall 2011-Winter 2013

- Organize volunteers and coordinate budget and finances for department social activities
- Increase communication between department leadership, faculty and students
- Run weekly and yearly department social events, organize new events

Student Organizer, Machine Learning Seminar
University of Maryland, College Park Spring 2012

- Help publicize talks by visiting researchers as part of Yahoo!-sponsored ML Seminar
- Coordinate scheduling and faculty meetings for visitors

Elected Student Representative, Computer Science Department Council
University of Maryland, College Park Fall 2012-Spring 2013

- Participated in faculty-student panels to improve Visit Day and Grad Student orientation
- Work to create new collaboration spaces, such as conference rooms and student meeting spaces

Volunteer for Prospective Student Events
University of Maryland, College Park Spring 2011-Spring 2014

- Student contact for admitted students, answering questions about the department
- Volunteer at Visit Day events, such as open Q&A session and student dinners
- Responsible for AI area dinner for prospective students, Spring 2011.

Elected Representative, Graduate Student Government
University of Maryland, College Park Fall 2012-Spring 2014

- Represent Computer Science students in the Graduate Student Government

- Serve on the Sustainability Committee to improve environmental awareness across campus.

Presenter at National Youth Science Camp,
Camp Pocahontas, Barstow, WV Summer 2011, 2012, 2013

- Presented a 90-minute lecture about Computer Science and Machine Learning geared towards new high school graduates
- Conducted a three-day workshop on Machine Learning for a class of ten students

Computer Skills

Programming: C/C++, Java, Perl, Python, Matlab, Pig
Development Platforms: Linux, FreeBSD, Mac OSX, Windows

Relevant Graduate Coursework

Machine Learning, Computational Linguistics, Algorithmic Game Theory, Information Visualization, Computer Networks, Database Management Systems, Computational Models of Neural Systems, Cognitive Neuropsychology, Mobile Robot Programming, Link Mining, Social Networks and Privacy, Advanced Machine Learning: Computational Methods for High-Throughput Analysis of Biological Systems