JAY PUJARA – RESEARCH STATEMENT

My research on large-scale machine learning for dynamic, uncertain environments addresses the challenges of building adaptive and collective models in today's landscape of large, heterogeneous data.

I solve the fundamental problems inherent in *noisy*, *large*, and *dynamic* data. To deal with uncertain, interdependent predictions I use **probabilistic models** that naturally capture complex dependencies and statistical uncertainty. Historically, a significant limitation that has prevented wider deployment of probabilistic models is **scalability**; I've designed rich probabilistic models that can jointly optimize millions of variables efficiently on even modest hardware, while still delivering state-of-the-art performance. However, modern problems are not only large, but also **dynamic**. Models must nimbly adapt to new information, and I have developed algorithms to update models while providing theoretical guarantees of quality.

In addition to addressing fundamental challenges in machine learning, my work is also driven by **important, practical questions** in artificial intelligence [1–5]. How can we use the wealth of knowledge on the Web to construct structured knowledge bases? When a user rates a new item, how can we update recommendations so other similar users benefit? What can we learn about the invisible influences of organizations from the social media activity of their followers? The central thread connecting these diverse questions is the need to exploit relationships and dependencies between instances – whether they are facts in a knowledge base, items in a product catalog, or users of a social network.

The marriage of theoretical challenges with practical questions has allowed me to build a versatile research agenda that both attracts **diverse collaborations** and positions me as a leader in the machine learning community. As a consequence, I have published expansively in venues spanning security [6], natural language processing [7, 8], social network analysis [5, 9], bioinformatics [10, 11], the Semantic Web [3, 12], and graphical models [4, 13, 14]. Complementing the breadth of my work is its depth. My twenty publications include **three best paper awards** [3, 6, 14] and an article featured in AI Magazine [2]. I am a **recognized leader in knowledge graph construction** [2, 3, 13–19], and have contributed to the community, both as an organizer of the immensely popular Automated Knowledge Base Construction (AKBC) workshop, and through a tutorial on knowledge graph construction offered at AAAI 2017. Moreover, my research has clear intersections with databases (entity resolution, query compilation and grounding), human computer interaction (explainable AI), scientific computing (optimization), systems (distributed computation), and theory (learning guarantees). In addition to my academic credentials, I also have significant industry experience solving large-scale machine learning problems in the wild. These include a senior engineering position at Yahoo!, research internships at LinkedIn and Google, and collaborative projects with Silicon Valley startup companies as a postdoc at UC Santa Cruz.

Probabilistic Models for Uncertain, Interdependent Predictions

The cornerstone of my research is designing sophisticated probabilistic models. Probabilistic graphical models are often a good match for real-world problems, where there are often *uncertain* input data, *constraints* on the predicted outputs, *dependencies* between variables, and a need for *collective inference* and *structured prediction*. For example, consider the problem of determining voter choices during an election. This problem requires integrating uncertain data, such as previous voting records or social media posts. Predictions are constrained so each voter selects one candidate for each position. Additionally, there are relational dependencies that a person will vote similarly to their friends, whose votes are also unknown. Ultimately, this model requires defining a joint probability distribution over all individual votes, and performing a global optimization to determine the most probable combination of all votes.

Historically, the barrier to harnessing the power and elegance of probabilistic approaches has been the difficulty of specifying and optimizing models. Given the complexity in even the simple voting example above, the task of building a statistical model and implementing efficient optimization is daunting. My contributions to **probabilistic soft logic** (PSL) [20], an accessible and scalable framework for probabilistic modeling, addresses these hurdles and enable solutions to large, noisy and complex problems. PSL allows model specification using a first-order logic syntax that is easily grasped by non-experts. These logical models are compiled into a convex optimization problem that can be solved efficiently while still offering optimality guarantees. My contributions include improving the data model, particularly for dynamic data [17], implementing a command-line interface to allow non-experts to run models, developing streaming inference algorithms [4], and modeling integrations, including the largest problem solved using PSL, knowledge graph construction.

Knowledge Graph Construction

My dissertation [1] demonstrated that rich probabilistic models both excel at constructing knowledge graphs and can scale to big, noisy and dynamic problems. Knowledge graphs power everything from search queries and Siri to decision-making systems. Yet despite their ubiquity, these tools are fairly limited because they rely on a narrow slice of curated knowledge. My work addresses a fundamental problem in artificial intelligence, the knowledge acquisition bottleneck, by combining the statistical signals from big datasets with the semantic constraints necessary to ensure meaningful, precise output. I define the problem of **knowledge graph identification** (KGI) [3], involving three interrelated problems: determining the entities in the knowledge graph, the attributes of each entity, and the relationships between entities. Each task inherently depends on the others. For example, entity types constrain their relations, while signals about co-referent entities can be pooled. My unique solution to KGI defines a probabilistic model using PSL that seamlessly and scalably combines statistical signals from extraction systems (like NLP pipelines or computer vision) with semantic constraints drawn from ontologies, rules and the Semantic Web. Integrating these in a collective model has dramatic results: on large problems involve tens of millions of statistical and semantic constraints, KGI **improves F1 measure over 25%** compared to information extraction systems, and completes in just two hours. When evaluated on benchmarks, KGI **beats state-of-the-art baselines using just seconds of computation.** KGI has had substantial impact in the field, winning a best paper award [3], Google Research Award, and being featured in AI Magazine [2]. My subsequent application of KGI enabling entity resolution in the Google Knowledge Graph won a second best paper award [14].

Scalable Machine Learning for Collective, Big Data Problems

My system for knowledge graph construction has an immense scalability advantage over competing systems. Yet as data grow exponentially, innovative approaches are needed to scale models, either by leveraging more computational resources or through discerning use of data. My research contributions involve two general scalability strategies: **distributed computation for probabilistic models** and **data-efficient models**.

One strategy for dealing with large problems is horizontal partitioning to **distribute computation across machines**. In the rich probabilistic models I've described, naive approaches ignore collective dependencies and perform poorly. I developed an innovative technique that performs a graph-cut on a data-annotated representation of the graphical model, rather than partitioning the data itself. Applying my **partitioning approach to knowledge graphs speeds up KGI 90%** with little loss in quality [16]. Even when model structure is initially unknown, my methods iteratively learn provisional structure and partition the data to refine the model [8].

A complementary strategy to distributing problems is **making models more data-efficient**. My application of data efficient classification to spam detection was **awarded a Yahoo! FREP award and a best paper award** [6]. I developed an adaptive classifier constrained by a data budget that **does more with less**, matching the predictions of more expensive models with 30% less data, or adhering to the budget of cheaper models while improving accuracy by 20%. Data efficiency can also reduce human effort. By leveraging unlabeled data to intelligently acquire labeled training instances, I improve predictive accuracy given a labeling budget [21].

Streaming Inference for Growing and Evolving Data

My research is scalable not only on big data, but also addresses the challenge of integrating dynamic data, producing models that grow and change over time. Incorporating new evidence is integral to solving real-world problems, but updating models is nontrivial, particularly in collective models where predictions are the result of joint optimization. My work on updating inference in collective, probabilistic models [4] answers two basic questions: (1) what are the consequences of partially updating predictions in a collective model? and (2) which predictions should be updated when new evidence arrives? Studying the first question, I proved **theoretical guarantees of solution quality after a partial update**, demonstrating that low-regret strategies for selectively updating models can succeed. I complemented this result by building a suite of **intelligent strategies for efficiently updating models**. One of the key novelties of my research is a deep integration into the optimizer: my algorithms track the most uncertain variables and most difficult optimization terms during inference, and exploit these patterns to channel effort in future epochs. My analysis of consensus optimization led to approaches that use Lagrange multipliers to select the most important variables to update when new evidence arrives. Adopting these strategies on evolving problem settings such as knowledge graph construction [13], collaborative filtering in recommender systems and preference classification in social networks **reduces the model update time by 70%** yet preserves predictive performance, outperforming full updates in some scenarios.

Future Research Directions

Building on this research foundation, I highlight my vision for enabling scalable machine learning to partner with humans to exploit massive, multi-modal datasets and solve challenges from sustainability to healthcare.

Curious ML: Optimization-driven Explainable and Interactive Learning

Machines can absorb the deluge of data, but ultimately humans must understand and act on algorithmic results. I envision curious ML systems that reflect on predictions and identify promising areas to explore. **My goal is to build systems that answer the questions, "What is interesting about the data?" and "Where does the model fail?"** To this end, I seek to formalize curious ML, a suite of algorithms that not only optimize an objective, but also identify important predictions for explaining and debugging models. My initial implementation of these ideas applied to biomedical domains [11] enabled models to identify promising new treatments using existing drugs.

Hybrid Architectures: Structured Probabilistic Models for Integrating Heterogeneous Data Streams

Structure is an important ingredient in many tasks, particularly when fusing data from different domains, such as text, vision, sound, and mechanical sensors. Recent advances in deep learning have produced excellent results in these domains, but struggle to incorporate structure. **My goal is developing novel, interpretable models that harness diverse data streams with meaningful semantics, while propagating valuable feedback to black-box feature extraction engines.** Results on knowledge graphs [3] demonstrate the power of this approach when combining different NLP extraction techniques.

Critical Real-World Applications: Knowledge Graphs, Bioinformatics, Sustainability

My research agenda is aimed at solving large, difficult problems, and I describe three areas of ongoing work where applied ML can produce dramatic improvements. Knowledge graphs are increasingly the interface between vast repositories of knowledge and humans making vital decisions. I am tackling the problem of **open-world concept learning**, constructing hierarchical taxonomies of previously unseen relations and entities, such as new diseases and regulation pathways in medical literature. Additional bioinformatics collaborations include working with the NIH to perform entity resolution in **family health histories** [10], and with UCSC computational biologists to understand **cellular pathways involved in breast cancer**. Several of my projects apply ML to find sustainable solutions, including **analyzing smart meter data** to reduce energy consumption (at UCSC) and **understanding commute patterns** to recommend green alternatives (with UC Berkeley).

References

[1] Jay Pujara. *Probabilistic Models for Scalable Knowledge Graph Construction*. PhD thesis, University of Maryland, College Park, 2016.

[2] Jay Pujara, Hui Miao, Lise Getoor, and William W. Cohen. Using Semantics & Statistics to Turn Data into Knowledge. *AI Magazine*, 36(1):65–74, 2015.

[3] Jay Pujara, Hui Miao, Lise Getoor, and William W. Cohen. Knowledge Graph Identification. In *International Semantic Web Conference (ISWC)*, 2013. **Winner of Best Student Paper award**.

[4] Jay Pujara, Ben London, and Lise Getoor. Budgeted Online Collective Inference. In *Uncertainty and Artificial Intelligence (UAI)*, 2015.

[5] Shachi Kumar, Jay Pujara, Lise Getoor, David Mares, Dipak Gupta, and Ellen Riloff. Unsupervised Models for Predicting Strategic Relations between Organizations. In *International Conference on Advances in Social Networks Analysis and Mining*, 2016.

[6] Jay Pujara, Hal Daumé III, and Lise Getoor. Using Classifier Cascades for Scalable E-Mail Classification. In *Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference*, 2011. **Winner of Best Paper Award**.

[7] Adam Grycner, Gerhard Weikum, Jay Pujara, James Foulds, and Lise Getoor. RELLY: Inferring Hypernym Relationships Between Relational Phrases. In *Conference on Empirical Methods in Natural Language Processing*, 2015.

[8] Jay Pujara and Peter Skomoroch. Large-Scale Hierarchical Topic Models. In *NIPS Workshop on Big Learning*, 2012.

[9] Bert Huang, Stephen H. Bach, Eric Norris, Jay Pujara, and Lise Getoor. Social Group Modeling with Probabilistic Soft Logic. In *NIPS Workshop on Social Network and Social Media Analysis: Methods, Models, and Applications*, 2012.

[10] Pigi Kouki, Jay Pujara, Christopher Marcum, Laura Koehly, and Lise Getoor. Collective Entity Resolution in Familial Networks. *Under Review*.

[11] Shobeir Fakhraei, Dhanya Sridhar, Jay Pujara, and Lise Getoor. Adaptive Neighborhood Graph Construction for Inference in Multi-Relational Networks. In *KDD Workshop on Mining and Learning with Graphs (MLG)*, 2016.

[12] Jay Pujara, Hui Miao, Lise Getoor, and William W. Cohen. Extended Abstract: Large-Scale Knowledge Graph Identification using PSL. In *AAAI Fall Symposium on Semantics for Big Data*, 2013.

[13] Jay Pujara, Ben London, Lise Getoor, and William W. Cohen. Online Inference for Knowledge Graph Construction. In *Fifth International Workshop on Statistical Relational AI*, 2015.

[14] Jay Pujara and Lise Getoor. Generic Statistical Relational Entity Resolution in Knowledge Graphs. In *Sixth International Workshop on Statistical Relational AI*, 2016. **Winner of Best Paper Award**.

[15] Jay Pujara, Hui Miao, Lise Getoor, and William W. Cohen. Large-Scale Knowledge Graph Identification using PSL. In *ICML Workshop on Structured Learning*, 2013.

[16] Jay Pujara, Hui Miao, Lise Getoor, and William W. Cohen. Ontology-Aware Partitioning for Knowledge Graph Identification. In *CIKM Workshop on Automatic Knowledge Base Construction*, 2013.

[17] Jay Pujara and Lise Getoor. Building Dynamic Knowledge Graphs. In *NIPS Workshop on Automated Knowledge Base Construction*, 2014.

[18] Jay Pujara, Kevin Murphy, Xin Luna Dong, and Curtis Janssen. Probabilistic Models for Collective Entity Resolution Between Knowledge Graphs. In *Bay Area Machine Learning Symposium*, 2014.

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

[20] LINQS Research Group. Probabilistic Soft Logic. `https://github.com/linqs/psl/`, 2016.

[21] Jay Pujara, Ben London, and Lise Getoor. Reducing Label Cost by Combining Feature Labels and Crowdsourcing. In *ICML Workshop on Combining Learning Strategies to Reduce Label Cost*, 2011.