Research Statement

William Yang Wang, CMU

I study the theoretical foundation and practical algorithms for **Artificial Intelligence**. To build intelligent machines that can tackle challenging reasoning problems under uncertainty, I have pursued answers via studies of **Machine Learning**, **Natural Language Processing**, and **Interdisciplinary Data Science**. More specifically, I am interested in designing scalable inference and learning algorithms to analyze massive datasets with complex structures. In particular, I advance methods in the following research areas:

- Statistical Relational Learning
- Knowledge Representation and Reasoning
- Natural Language Processing, Speech, and Computational Social Science

The central focus of my dissertation research is to bring together all areas above and design scalable algorithms for large scale inference problems on knowledge graphs (§1). Meanwhile, I enjoy collaborating with scientists and domain experts of different background for interdisciplinary research in data science (§2). In future work, I look forward to advancing fundamental problems in artificial intelligence, while creating impacts of data analysis across all fields of science (§3).

1 Scalable Inference for Statistical Relational Learning

Learning to reason and understand the world's knowledge is a fundamental problem in Artificial Intelligence (AI). Traditional symbolic AI methods are popular in the 1980s, when first-order logic rules are mostly handwritten, and reasoning algorithms are built on top of them. In the 90s, more and more researchers are interested in statistical methods that deal with the uncertainty of the data, using probabilistic models. While it is always hypothesized that both the symbolic and statistical approaches are necessary for building intelligent systems, in practice, bridging the two in a combined framework might bring intractability—most probabilistic first-order logics are simply not efficient enough for real-world sized tasks. Therefore, building scalable probabilistic logic is a core research task in modern statistical AI research (see Fig. 1). In my PhD thesis, I aim to build a new paradigm to combine the best of the two worlds by leveraging the complementarities of first-order logic reasoning and scalable machine learning and inference algorithms.

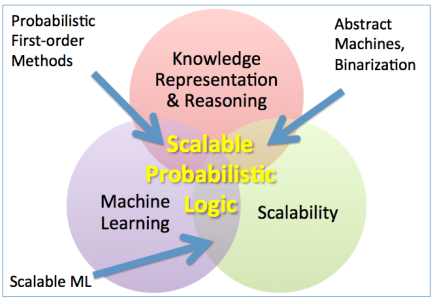

Figure 1: Scalable Probabilistic Logic is a key cross-subject research area in AI.

1.1 Scalable Inference and Parameter Learning

Many information management tasks can be formalized as inference in an appropriate probabilistic first-order logic. However, most probabilistic first-order logics are not efficient enough to be used for large-scale versions of these tasks: even a grounding of size linear in the number of facts in the database, $|DB|$, is impractically large for inference. Together with my advisor William Cohen, we propose ProPPR, a personalized PageRank based first-order probabilistic language which is well-suited to approximate "local" grounding. We present an extension to stochastic logic programs (SLP) [10] that is biased towards short derivations, and show that

this is related to personalized PageRank (PPR) [12, 3] on a linearized version of the proof space. Based on the connection to PPR, we develop a provably-correct approximate inference scheme, and an associated provably-correct approximate grounding scheme: specifically, we show that it is possible to prove a query, or to build a graph which contains the information necessary for weight-learning, in time $O(1/\alpha\epsilon)$, where α is a reset parameter associated with the bias towards short derivations, and ϵ is the worst-case approximation error across all intermediate stages of the proof. This means that **both inference and learning can be approximated in time independent of the size of the knowledge graph**—a surprising and important result. ProPPR's learning scheme is also highly scalable: by utilizing an asynchronized parallel stochastic gradient descent algorithm, one can easy scale up the parameter training with multiple threads and CPUs. A full description of the ProPPR language can be found in our award-winning CIKM paper [27].

1.2 Relational Inference with Big Data

To further investigate the scalability of our new paradigm, we consider the problem of learning and inference on a large knowledge graph that contains imperfect and incomplete knowledge. More specifically, we show that ProPPR can be used as a programmable language to recursively learn and perform inference for different target relations [30]. We situate our study in the context of the Never Ending Language Learning (NELL) research project, which is an effort to develop a never-ending learning system that operates 24 h per day, for years, to continuously improve its ability to extract structured facts from the web [2]. This task is challenging for two reasons. First, the extensional knowledge inference is not only incomplete, but also noisy, since it is extracted imperfectly from the web. Second, the sizes of inference problems are large relative to those in many other probabilistic inference tasks: even a grounding of size that is only linear in the number of facts in the database, $|DB|$, would be impractically large for inference on real-world problems. In our experiments on three NELL subsets with more than a million facts, we show that ProPPR's recursive learning theory significantly outperforms strong and state-of-the-art baselines such as Markov Logic Networks [13] and Path-Ranking Algorithm [11], varying different settings.

1.3 Structure Learning

As a part of my thesis, I also contribute a scalable solution to a core problem in statistical relational learning: structure learning to extract the inference rules from knowledge graphs. For example, we know that an athlete plays in a team, and a team plays in a league, we want to learn an inference rule that asserts that

$$\begin{aligned} &athletePlaysInLeague(A,L) \leftarrow \\ &\quad athletePlaysInTeam(A,T), teamInLeague(T,L). \end{aligned}$$

To do this, we propose a second-order abductive theory [28], whose parameters correspond to plausible first-order inference rules. To improve the performance, an iterative structural gradient approach is proposed to incrementally refine the set of learned rules. In experiments, we show that the proposed approach only needs a few minutes to learn the inference rules, and when using these learned rules, the predictive results improve traditional and statistical methods by a large margin. Furthermore, to learn a set of compact theory, we show that it is possible to apply a group Lasso based structured sparsity as soft predicate invention to this task [29], obtaining state-of-the-art results.

1.4 Joint Information Extraction and Reasoning

In an effort to improve information extraction (IE), I show that ProPPR can be used as a generic framework to combine the knowledge and text based methods [19]: a partially-populated knowledge base (KB) and a set of relation mentions in context are used as input, and we jointly learn (1) how to extract new KB facts from the relation mentions, and (2) a set of logical rules that allow one to infer new KB facts. It is shown that the proposed joint model for IE and relational learning outperforms universal schemas [14], a state-of-the-art joint method, and that incorporating latent context further improves the results. In the area of Natural Language Processing (NLP), I also show that ProPPR can also be used as a programmable dependency parser [24] for theory engineering in structure prediction tasks on social media data, outperforming state-of-the-art parsers.

2 An Interdisciplinary Approach to Data Science

The rapid development of machine learning techniques and the abundantly available data are making many fundamental changes to how scientific research is conducted in many disciplines, from finance to law to behavioral science. The complex nature of these problems require interdisciplinary approaches from many areas, including machine learning, natural language processing, multimodality, social science, etc.. In addition to my thesis research on learning from relational datasets, I was also very fortunate to work with many researchers to advance methods for various complex problems in cross-disciplinary data science, and I foresee strong impacts in the future.

2.1 Machine Learning for History, Law, Finance, Marketing, and Behavioral Science

Together with economists Suresh Naidu (Columbia) and Jeremiah Dittmar (London School of Economics), we are among the first to study Bayesian statistics and latent variable models to enhance historical analysis of large corpora [26]. In particular, we study the United States property law judgements related to slaves, with a special focus on shifts in opinions on controversial topics across different regions. During my graduate school, I also made a fundamental contribution to text-based financial analytics by proposing a scalable algorithm for training and inference with a semiparametric Gaussian copula [23]: we quantitatively study how earnings calls are correlated with the financial risks, with a special focus on the financial crisis of 2009. I am an advocate of introducing statistical machine learning approaches to marketing science: in my paper on computational branding analytics with Laplacian structured sparsity [25], we collected customer reviews from Starbucks, Dunkin' Donuts, and other coffee shops across 38 major cities in the USA. We studied the brand related language use through these reviews, with focuses on the brand satisfaction and gender factors. I am also interested in enhancing data harvesting for behavioral sciences: I have introduced an embedding and data augmentation based Twitter classification method for computational behavioral analysis in my recent award-winning EMNLP paper [34].

2.2 Natural Language Processing for Computational Social Science

During my time at Columbia, I leveraged natural language processing methods to solve problems in social collaborating systems, where I proposed a Web-based automatic vandalism detection algorithm for Wikipedia [31]. At CMU, I worked with Justine Cassell on sparsity-inducing learning approaches to sociolinguistics [20], where we analyze the notion of politeness, positivity, and social dynamics in child conversations. Recently, with my colleague Wen, we have studied nonparanormal approaches to automatic generation of Internet memes [33], which is covered by media in U.S. and Spain. I am also interested in designing scalable algorithms for core NLP tasks in social media: I have worked on dependency parsing for Weibo [24].

2.3 Speech, Language, and Vision for Semantic Understanding

Within AI, I am excited about collaborating with leading researchers from different subfields to build intelligent interactive machines. My background in speech enables me to work on paralinguistics signal processing in speech: my work in predicting intoxication and levels of interests from speech [17, 1, 22] are reported in the Best Science Book of 2014 by Amazon, Wired, the Guardian, and NBC [15]. Together with researchers from University of Southern California and Columbia, we have made significant contributions for building better language understanding and generation systems, including phonetic mixture models [16], semi-supervised event detection models [32], and a unit-selection speech synthesizer [21]. During my internship at Microsoft Research, I was fortunate to work with Eric Horvitz on crowdsourcing approaches to language understanding [18]. In the past few years, I was also a core member of an ambitious project, where we exploit unsupervised approaches to automatic induction and filling of semantic slots for spoken dialogue systems using frame-semantic and distributional methods [5, 7, 9, 6, 8, 4]. I have hands-on experiences of combining computer vision and NLP techniques for generating popular descriptions of Internet memes [33].

3 Future Work

In the beginning of the first section, I have mentioned that reasoning with symbolic systems, such as first-order logics, is a primary research theme in the AI community in the 1980s. Interestingly, neural network based learning methods revived in the 1980s as well, even though researchers back then did not figure out how to scale up training on large datasets.

First-order logic based methods are often deterministic, reliable, and perform well using a handful crisp inference rules. However, an obvious drawback of these traditional symbolic systems is the scalability issue: it is expensive to create and update the hand-written rules, and they often suffer from the coverage issues when new data comes in. On the other hand, neural network models, are reviving again recently, due to the availability of massive computing resources, as well as better learning and inference methods. Even though there have been many promising results reported for simple classification tasks, it is still unclear how well we can use neural models to perform reasoning tasks in statistical relational learning. Another notable issue with the pure neural network based methods is the explainability: the learned hidden layers are often hard to explain, and it is often difficult to perform error analysis, and understand the science behind neural methods.

I believe that now it is the right time to revisit the idea of integrating logical and neural systems. For example, it would be interesting to design a deep logic model that allows joint learning and inference of first-order clauses and latent representations. In one of my recent papers [19], I introduce a novel extension of the joint IE and reasoning model called Latent Context Invention (LCI), which associates latent states with context features for the IE component of the model. We show that LCI further improves performance, leading to a substantial improvement over prior state-of-the-art methods for joint relation-learning and IE.

In addition to pursuing research in neural logical models, I would also like to continue my interdisciplinary research in data science: I believe that with smart algorithms and collaborations with researchers of different background, now it is an exciting to tackle complex problems in data science for social good.

References

[1] Fadi Biadsy, **William Yang Wang**, Andrew Rosenberg, and Julia Hirschberg. Intoxication detection using phonetic, phonotactic and prosodic cues. In *Proceedings of 12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011)*, Florence, Italy, August 2011. ISCA.

[2] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka Jr, and Tom M Mitchell. Toward an architecture for never-ending language learning. In *AAAI*, volume 5, page 3, 2010.

[3] Soumen Chakrabarti. Dynamic personalized pagerank in entity-relation graphs. In *Proceedings of the 16th international conference on World Wide Web*, pages 571–580. ACM, 2007.

[4] Yun-Nung Chen, **William Yang Wang**, Anatole Gershman, and Alex I. Rudnicky. Matrix factorization with knowledge graph propagation for unsupervised spoken language understanding. In *Proceedings of The 53rd Annual Meeting of the Association for Computational Linguistics and The 7th International Joint Conference of the Asian Federation of Natural Language Processing*, Beijing, China, 2015. ACL.

[5] Yun-Nung Chen, **William Yang Wang**, and Alex Rudnicky. Unsupervised induction and filling of semantic slots for spoken dialogue systems using frame-semantic parsing. In *Proceedings of the ASRU 2013*, Olomouc, Czech Republic, December 2013. IEEE.

[6] Yun-Nung Chen, **William Yang Wang**, and Alex Rudnicky. Learning semantic hierarchy with distributional representations for unsupervised spoken language understanding. In *Proceedings of 16th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Dresden, Germany, 2015. ISCA.

[7] Yun-Nung Chen, **William Yang Wang**, and Alex I. Rudnicky. An empirical investigation of sparse log-linear models for improved dialogue act classification. In *Proceedings of the 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2013)*, Vancouver, Canada, May 2013. IEEE.

[8] Yun-Nung Chen, **William Yang Wang**, and Alex I. Rudnicky. Jointly modeling inter-slot relations by random walk on knowledge graphs for unsupervised spoken language understanding. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Denver, CO, USA, 2015. ACL.

[9] Yun-Nung Chen, **William Yang Wang**, and Alexander I Rudnicky. Leveraging frame semantics and distributional semantics for unsupervised semantic slot induction in spoken dialogue systems. In *Spoken Language Technology Workshop (SLT), 2014 IEEE*, pages 584–589. IEEE, 2014.

[10] James Cussens. Parameter estimation in stochastic logic programs. *Machine Learning*, 44(3):245–271, 2001.

[11] Ni Lao, Tom Mitchell, and William W Cohen. Random walk inference and learning in a large scale knowledge base. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 529–539. Association for Computational Linguistics, 2011.

[12] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: bringing order to the web. 1999.

[13] Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine learning*, 62(1-2):107–136, 2006.

[14] Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M Marlin. Relation extraction with matrix factorization and universal schemas. In *NAACL-HLT*, 2013.

[15] Adam Rogers. *Proof: The Science of Booze*. Houghton Mifflin Harcourt, 2014.

[16] **William Yang Wang**, Ron Artstein, Anton Leuski, and David Traum. Improving spoken dialogue understanding using phonetic mixture models. In *Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference (FLAIRS-24)*, Palm Beach, USA, May 2011. AAAI Press.

[17] **William Yang Wang**, Fadi Biadsy, Andrew Rosenberg, and Julia Hirschberg. Automatic detection of speaker state: Lexical, prosodic, and phonetic approaches to level-of-interest and intoxication classification. *Computer Speech amp; Language*, (0):–, 2012.

[18] **William Yang Wang**, Dan Bohus, Ece Kamar, and Eric Horvitz. Crowdsourcing the acquisition of natural language corpora: Methods and observations. In *Proceedings of the IEEE SLT 2012*, Miami, Florida, Dec. 2012. IEEE.

[19] **William Yang Wang** and William W. Cohen. Joint information extraction and reasoning: A scalable statistical relational learning approach. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and The 7th International Joint Conference of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2015)*, Beijing, China, July 2015. ACL.

[20] **William Yang Wang**, Samantha Finkelstein, Amy Ogan, Alan W Black, and Justine Cassell. "love ya, jerkface": using sparse log-linear models to build positive (and impolite) relationships with teens. In *Proceedings of the 13th annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL 2012)*, Seoul, South Korea, July 2012. ACL.

[21] **William Yang Wang** and Kallirroi Georgila. Automatic detection of unnatural word-level segments in unit-selection speech synthesis. In *Proceedings of the ASRU 2011*, Big Island, HI, December 2011. IEEE.

[22] **William Yang Wang** and Julia Hirschberg. Detecting levels of interest from spoken dialog with multistream prediction feedback and similarity based hierarchical fusion learning. In *Proceedings of the 12th annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL 2011)*, Portland, OR., USA, June 2011. ACL.

[23] **William Yang Wang** and Zhenhao Hua. A semiparametric gaussian copula regression model for predicting financial risks from earnings calls. In *Proceedings of the 52th Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, Baltimore, MD, USA, June 2014. ACL.

[24] **William Yang Wang**, Lingpeng Kong, Kathryn Mazaitis, and William W. Cohen. Dependency parsing for weibo: An efficient probabilistic logic programming approach. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, Doha, Qatar, October 2014. ACL.

[25] **William Yang Wang**, Edward Lin, and John Kominek. This text has the scent of starbucks: A laplacian structured sparsity model for computational branding analytics. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, Seattle, WA, USA, October 2013. ACL.

[26] **William Yang Wang**, Elijah Mayfield, Suresh Naidu, and Jeremiah Dittmar. Historical analysis of legal opinions with a sparse mixed-effects latent variable model. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, Jeju Island, Korea, July 2012. ACL.

[27] **William Yang Wang**, Kathryn Mazaitis, and William W Cohen. Programming with personalized pagerank: A locally groundable first-order probabilistic logic. *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management (CIKM 2013)*, 2013.

[28] **William Yang Wang**, Kathryn Mazaitis, and William W Cohen. Structure learning via parameter learning. *CIKM*, 2014.

[29] **William Yang Wang**, Kathryn Mazaitis, and William W. Cohen. A soft version of predicate invention based on structured sparsity. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI 2015)*, Buenos Aires, Argentina, July 2015. AAAI.

[30] **William Yang Wang**, Kathryn Mazaitis, Ni Lao, and William W Cohen. Efficient inference and learning in a large knowledge base. *Machine Learning*, pages 1–26, 2015.

[31] **William Yang Wang** and Kathleen McKeown. "got you!": Automatic vandalism detection in wikipedia with web-based shallow syntactic-semantic modeling. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1146–1154, Beijing, China, August 2010. Coling 2010 Organizing Committee.

[32] **William Yang Wang**, Kapil Thadani, and Kathleen McKeown. Identifying event descriptions using co-training with online news summaries. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, Chiang Mai, Thailand, December 2011. AFNLP and ACL.

[33] **William Yang Wang** and Miaomiao Wen. I can has cheezburger? a nonparanormal approach to combining textual and visual information for predicting and generating popular meme descriptions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2015)*, Denver, CO, USA, 2015. ACL.

[34] **William Yang Wang** and Diyi Yang. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, Lisbon, Portugal, September 2015. ACL.