Better weather prediction through data mining

Amy McGovern

Severe weather including tornadoes, thunderstorms, hail, and wind caused $32 billion dollars of damage in 2011 and annually cause significant loss of life. Although forecasting the path and severity of hurricanes and tropical storms has improved significantly in recent years, tornadoes and other severe events on a smaller scale than hurricanes remain quite difficult to predict. While forecasters can identify conditions favorable for major tornado outbreaks several days in advance, short-term forecasting of individual storms, providing additional advanced notice, and predicting probable tornado paths remain a challenge.

The goal of much of Amy McGovern's (Ph.D. 2002, M.S. 1998) research as an associate professor in the School of Computer Science at the University of Oklahoma has been to revolutionize tornado prediction and other forms of severe weather. She does this using artificial intelligence, data mining, machine learning, and storm simulations. McGovern received a National Science Foundation (NSF) CAREER award in 2008 to jumpstart her research. She collaborates with the National Oceanic and Atmospheric Administration's (NOAA) National Severe Storm Laboratory (NSSL) and researchers in the School of Meteorology at the University of Oklahoma. She is also working on improving the prediction of aircraft turbulence in collaboration with the National Center for Atmospheric Research (NCAR).

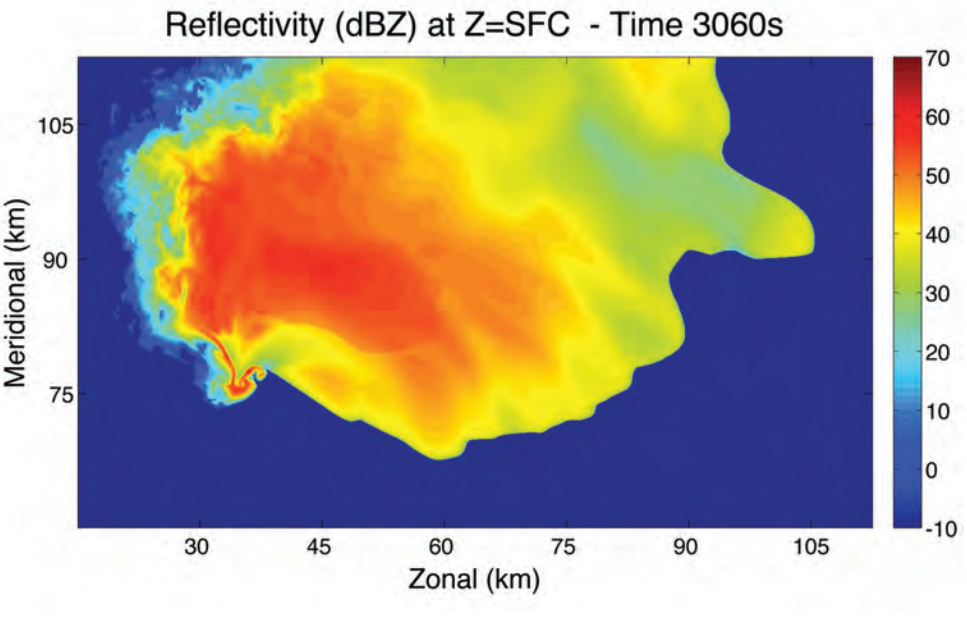
Severe weather poses a very challenging prediction and simulation problem. "Radars provide an incomplete picture of the atmosphere," says McGovern. "Although they can sense the intensity of the precipitation and a single dimension of the wind vector, there are many other important variables such as the full three-dimensional wind field, pressure, temperature, etc. that are important to prediction. Although simulations are an answer to this, fully simulating the atmosphere is not computationally feasible." McGovern is developing a unique set of high-resolution simulations of supercell thunderstorms. These are the most severe type of thunderstorms and cause the most destructive tornadoes. These simulations also provide an unprecedented view of atmospheric turbulence.

Mining the simulations is also challenging. At the resolutions McGovern is simulating, each simulation generates over 1 terabyte (TB) of data. Statistical relational learning is used to identify high-level concepts and relationships in the data that can be used to predict tornadoes. Meteorologists already study existing storm data using conceptual models. They identify high-level concepts and regions in a storm such as updrafts, a region of air blowing upward, and downdrafts, a region of air blowing downward. McGovern's models provide the ability to identify spatiotemporal relationships between these regions that can be used to predict the severe weather events. She has developed novel data mining models that make use of the spatiotemporal nature of the data because neither space nor time can be ignored for weather prediction. In addition, weather is three-dimensional and her models can identify arbitrary shapes and relationships between the shapes.

McGovern's three-dimensional weather modeling of spatiotemporal hazards will be valuable to aviation weather forecasting in support of the future U.S. National Airspace System, known as NextGen. The current Federal Aviation Administration (FAA) system provides guidelines about how close an aircraft can fly to a thunderstorm. Working with researchers at NCAR and using observations collected from aircraft flying over the continental United States to study convectively-induced turbulence, McGovern is improving the prediction of how far turbulence can spread from a thunderstorm. "This can be used to save money by flying more efficient routes and to prevent injuries by flying safer routes," says McGovern.

Another goal of McGovern's work as a professor is to engage, retain, and graduate more underrepresented students. She focuses on developing authentic CS and ML applications, especially those involving severe

continued on page 11



Simulated reflectivity (measure of the intensity of the precipitation) near the ground for one of McGovern's high resolution models.

McGOVERN – – – – – – – – – – *continued from page 10*

weather. Since her teaching environment is literally located in the proverbial "tornado alley" or "tornado capital of the world," it provides many real world severe weather experiences in students' actual lives. "They can more easily see the practicality of their CS and ML classes and related labs, homework, projects, and research throughout their college experience," notes McGovern. She also engages students through a variety of active learning projects and has won several teaching awards while at the University of Oklahoma.

McGovern's volunteer and community service activities are varied and many relate to the encouragement of undergraduate minorities and younger students to study CS. For the past five years, she chaired the Oklahoma EPSCoR (NSF's Experimental Program to Stimulate Competitive Research) Women in Science conference held for middle and high school students and counselors across the state of Oklahoma. She also serves as faculty advisor to OU's chapter of Alpha Sigma Kappa, a sorority for women in technical studies. As chair of the American Meteorological Society's Committee on Artificial Intelligence Applications to the Environment, McGovern both brings together researchers in the environmental sciences and artificial intelligence and is also reaching out to students in grades kindergarten through sixth. Current development includes an iPad application to demonstrate the uses of artificial intelligence for weather applications.

McGovern joined the University of Oklahoma in 2005 and is currently the Director of the Interaction, Discovery, Exploration, Adaptation (IDEA) Laboratory. While at UMass Amherst, she was advised by Professor Andrew Barto.

ACM recognizes outstanding alums

The Association for Computing Machinery (ACM), recently announced that CS alum Lixin Gao (Ph.D. '97) has been named an ACM Fellow for contributions to network protocols and internet routing. She is a professor of Electrical & Computer Engineering at UMass Amherst. The grade of Fellow recognizes the top 1percent of ACM members for their outstanding accomplishments in computing and information technology and/or outstanding service to ACM and the larger computing community.

ACM also named three CS Ph.D. alums, Antony L. Hosking (Ph.D. '95), Erich M. Nahum (Ph.D. '97), and Peri Tarr (Ph.D. '96), as 2013 Distinguished Members for their individual contributions and their singular impacts on the dynamic computing field. Nahum and Tarr are researchers at IBM Thomas J. Watson Research Center and Hosking is an associate professor at Purdue University. The ACM Distinguished Member program, initiated in 2006, recognizes those members with at least 15 years of professional experience who have achieved significant accomplishments or have made a significant impact on the computing field.

Alum Connections

Bo An (Ph.D. '11) was selected to China's 1,000 Young Talents Program. In in June 2012, he joined the Institute of Computing Technology of the Chinese Academy of Sciences as an associate professor. According to the Program's website, this program is for young researchers under age 40 who have obtained a doctorate degree in a world-famous university, and have no less than three years of overseas working experience. "The applicants should be the top-notch talents in their research fields, and have the potential to become future leaders in relevant areas. Special admissions are granted to those who have made distinguished research achievements in their doctorate studies or in other areas." As part of the program, recipients receive a living subsidy and a three year research grant.

Mark D. Smucker (Ph.D. '08), assistant professor at the University of Waterloo, and co-author Charles L.A. Clarke, received the Best Paper Award at the 35th Annual International Conference on Research and Development in Information Retrieval (SIGIR '12) for their paper "Time-Based Calibration of Effectiveness Measures."

In November, **Steven Sinofsky** (M.S. '89), former president of the Windows division of Microsoft Corporation, left the company after a 23 year career with the company. According to his tweet, he will teach this spring at Harvard Business School as an "Executive in Residence."

The Technical University of Denmark has appointed **Joseph Kiniry** (M.S. '95) professor. He is the head of the Software Engineering Section in the Department of Informatics and Mathematical Modeling. His inaugural lecture, "Saving Democracy from Technology," was presented on February 1, 2013.

In Memoriam

We are sad to announce that two of our CS alums passed away recently. **Doug Niehaus** (Ph.D. '93) died on August 21 at the age of 54. He was an associate professor of electrical engineering and computer science at the University of Kansas. **Gerald (Gerry) Hanam** of Framingham, MA died in October at the age of 46. He graduated cum laude with a B.S. in 1991. He leaves his wife, Catherine McCarthy ('88), and their daughter.

Dwyer named IEEE Fellow

CS Alum Matthew Dwyer (Ph.D. '95) was named a 2013 IEEE Fellow. He received the distinction "for contributions to specification, testing, analysis, and verification of concurrent software." IEEE Fellow is the highest grade of membership and is recognized by the technical community as a prestigious honor and an important career achievement. Dwyer is a Henson Professor of Software Engineering at the University of Nebraska. While at UMass Amherst, he was advised by Prof. Lori Clarke.

Recent Computer Science Ph.D. graduates (Fall 2012)

Michael Bendersky; *Information Retrieval with Query Hypergraphs;* (W. Bruce Croft, Advisor); Sept. 2012; Software Engineer, Google, Inc.

In this thesis we focus on verbose natural language search queries. To this end, we propose an expressive query representation based on query hypergraphs. Unlike the existing query representations, query hypergraphs model the dependencies between arbitrary concepts in the query, rather than dependencies between single query terms. Query hypergraphs are parameterized by importance weights, which are assigned based on their contribution to the retrieval effectiveness.

Query hypergraphs are not limited to modeling the explicit query, and we develop two methods for query expansion using query hypergraphs. In these methods, the expansion concepts may come either from the retrieval corpus or from a combination of external information sources. We empirically demonstrate that query hypergraphs are significantly more effective than many of the current state-of-the-art retrieval methods. Query hypergraphs improve the retrieval performance for all query types, and, in particular, they exhibit the highest effectiveness gains for verbose queries.

Filip Jagodzinski; *Towards Large Scale Validation of Protein Flexibility Using Rigidity Analysis;* (Ileana Streinu, Advisor); Sept. 2012; Assistant Professor, Dept. of Computer Science, Central Washington Univ.

Proteins flex and bend to perform their functions. At the atomic level, their motions cannot be observed. Rigidity analysis is a graph-based technique that infers the flexibility of molecules. Due to the lack of convenient tools for curating protein data, the usefulness of rigidity analysis in inferring biophysical properties has been demonstrated on only a handful of molecules. Conversely there is no agreed-upon choice of modeling of important stabilizing interactions.

We make progress towards large-scale validation of protein flexibility using rigidity analysis. We develop the KINARI software that permits automated curation of protein data. Rigidity analysis of protein biological assemblies generated by KINARI provides information that would be missed if only the unprocessed data were analyzed. We develop KINARI-Mutagen, which permits evaluation of the effects of mutations. Finally, we systematically vary the modeling of inter-atomic interactions and measure how rigidity parameters correlate with experimental data.

[Note: Special thanks to Filip for all of his efforts as a *Significant Bits* graduate student liaison for the past six years.]

Jin Young Kim; *Retrieval and Evaluation Techniques for Personal Information;* (W. Bruce Croft, Advisor); Sept. 2012; Applied Researcher, Microsoft Bing

Providing an effective mechanism for personal information retrieval is important for many applications, and requires different techniques than have been developed for general web search. This thesis focuses on developing retrieval models and representations for personal search, and on designing evaluation frameworks that can be used to demonstrate retrieval effectiveness in a personal environment.

From the retrieval model perspective, personal information can be viewed as a collection of multiple document types each of which has unique metadata. Based on this perspective, we propose a retrieval model that exploits document metadata and multi-type structure. Proposed retrieval models were found to be effective in other structured document collections, such as movies and job descriptions.

Evaluating these methods is particularly challenging for personal information due to privacy issues. This thesis introduces a set of techniques that enables realistic and repeatable evaluation of techniques for personal information retrieval. In particular, we describe techniques for simulating test collections and show that game-based user studies can collect more realistic usage data with relatively small cost.

Scott Kuindersma; *Variable Risk Policy Search for Dynamic Robot Control*; (Roderic Grupen and Andrew Barto, Advisors); Sept. 2012; Postdoctoral Associate, MIT Computer Science and Artificial Intelligence Laboratory

In this thesis, I present efficient global and local risk-sensitive stochastic optimization algorithms suitable for performing policy search in variety of problems of interest to robotics researchers. These algorithms exploit new techniques in nonparametric heteroscedastic regression to directly model the policy dependent distribution of cost. For local search, learned cost models can be used as critics for performing risk-sensitive gradient descent. Alternatively, decision-theoretic criteria can be applied to globally select policies to balance exploration and exploitation in a principled way, or to perform greedy minimization with respect to risk-sensitive criteria. This separation of learning and policy selection leads to variable risk control, where risk sensitivity can be flexibly adjusted and appropriate policies can be selected at runtime without requiring additional policy executions. I describe several experiments with the uBot-5 including learning dynamic arm motions to stabilize after large impacts, lifting heavy objects while balancing, and developing safe fall bracing behaviors.

Matthew Rattigan; *Leveraging Relational Representations for Causal Discovery;* (David Jensen, Advisor); Sept. 2012; Digital Analyst, Obama For America

This thesis represents a synthesis of relational learning and causal discovery, two subjects at the frontier of machine learning research. Relational learning investigates algorithms for constructing statistical models of data drawn from multiple types of interrelated entities, and causal discovery investigates algorithms for constructing causal models from observational data.

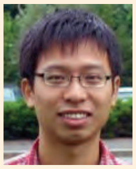
Traditionally, propositional (or "flat") data representations have dominated the statistical sciences. These representations assume that data consist of independent and identically distributed (iid) entities which can be represented by a single data table. More recently, data scientists have increasingly focused on "relational" data sets that consist of interrelated, heterogeneous entities. However, relational learning and causal discovery are rarely combined.

This unexplored topical intersection represents an opportunity for advancement, in which we can provide insight into the challenges found in each subject area. By adopting a causal viewpoint, we can clarify the mechanisms that produce previously identified pathologies in relational learning. Analogously, we can utilize relational data to establish and strengthen causal claims in ways that are impossible using only propositional representations.

Xiaobing Xue; *Modeling Reformulation as Query Distributions;* (W. Bruce Croft, Advisor); Sept. 2012; Software Developer, Twitter

Query reformulation modifies the original query with aim of providing a better representation of a user's information need and consequently improving the retrieval performance. Previous reformulation models typically generate words and phrases related to the original query, but do not consider how these words and phrases would fit together in realistic or actual queries.

A novel framework is proposed that models reformulation as a distribution of reformulated queries, where each reformulated query is associated with a probability indicating its importance. This framework considers a reformulated query as the basic unit and can capture the important query-level dependencies between words and phrases in a realistic or actual query. Since a reformulated query is the output of applying a single or multiple reformulation operations, this framework combines different operations such as query segmentation, query substitution and query deletion within the same framework. Moreover, a retrieval model is considered as an integrated part of this framework, which considers the reformulation model and the retrieval model jointly.



Tingxin Yan; *Designing Novel Mobile Systems By Exploiting Sensing, User Context, and Crowdsourcing;* (Deepak Ganesan, Advisor); Sept. 2012; Assistant Professor, Department of Computer Science & Engineering, Univ. of Arkansas

We focus first on the domain of context-aware mobile systems. We study the problem of how to incorporate user context into mobile operating system design by presenting a system named FALCON—an application-preloading engine, which infers user context from sensing data, learns associations between user context and application usage, and preloads applications to improve their responsiveness. Compared with existing caching schemes, Falcon improves the application responsiveness by two times.

The second focus is on the domain of participatory sensing. We explore the problem of improving image search accuracy by presenting a mobile service named CrowdSearch that achieves over 95 percent accuracy consistently across multiple categories of images with response time in a minute.

We then study the problem of image search under resource constraints, by presenting a mobile system named SenSearch that turns smartphones into micro image search engines, where images are collected, indexed, and transmitted using compact features that are two magnitudes smaller than their raw format. SenSearch improves the energy and bandwidth cost by five times compared with existing image search engines.

Rattigan on Obama For America analytics team

CS Alum Matthew Rattigan (Ph.D. '12) held an important behind-the-scenes role on the Obama For America (OFA) campaign as a member of the digital group within the analytics department. His group focused on the online aspects of the campaign (email, social media, etc.). They are highlighted in a November, 2012, *TIME* article, "Inside the Secret World of the Data Crunchers Who Helped Obama Win."

Rattigan's work was primarily centered on using Facebook to reach people through their network of supporters, culminating in a massive get-out-the-vote campaign on Election Day. He worked with Ph.D. grads from statistics, political science, physics, and computer science working together to try to change the way campaigns are run.

"We found that in many cases, the messenger is just as important as the message itself," says Rattigan. "People are much more likely to act (volunteer, attend an event, or even vote) when the request comes from a trusted friend rather than someone they have never met on a mailing list."

Rattigan, advised by Associate Professor David Jensen, was a member of the Knowledge Discovery Laboratory while at UMass Amherst. His research focuses on learning causal models with relational data.

Second place in ACM competition

In the Northeast Regional Preliminary Contest of the 2012 ACM International Collegiate Programming Competition, one of the UMass Amherst Computer Science teams placed second, qualifying them for the regional finals.

The "Garbage Collector" team, consisting of Khanh Nguyen, Aibek Sarbayev, and Tung Pham *(shown l. to r.)*, competed against 17 other teams from nine schools at the event held in October at Western New England University. Associate Professor Erik Learned-Miller coached the UMass Amherst CS teams. During the Northeast North America Regional Finals held at Rochester Institute of Technology in November, the team placed sixth.

Bryan receives Aspirations in Computing Award and CS scholarship

Becky Bryan receives her award from CS Chair Lori Clarke (l.) and Assoc. Professor Yanlei Diao (r.).

Rebecca "Becky" Bryan, a freshman computer science undergraduate student at UMass Amherst, was a 2011-2012 winner of the Massachusetts Aspirations in Computing Affiliate Award (MACAA) and a runner-up of the National Center for Women & Information Technology (NCW IT) Award for Aspirations in Computing. As a recipient of the Massachusetts award, she also received a $5,000 UMass Amherst Computer Science scholarship.

The NCWIT is partnering with local technology companies and universities to honor young high school women with the MACAA for their computing-related achievements and interests.

Never exposed to the subject before, Bryan, of Westborough, MA, took a computer science course during her junior year at Westborough High School. She expressed an interest in CS to her father, so he took her on a tour of MathWorks in Natick, MA, where he works. After that tour, Bryan decided to pursue a career in CS. She has already done some software engineering work and hopes to obtain a Research Experience for Undergraduates (REU) position this summer in a field other than software engineering so that she can explore as much of computer science as she can, stating, "There are just so many options with computer science!"

Bryan was on the high school ski team for two years and participated in cross-country running and indoor track. A member of the National Honor Society, she was also involved in the start-up of the Christian Student Alliance with two close friends at her high school. Through her junior and senior years, the alliance grew from the initial three students to fifteen. "It is so rewarding," says Bryan, "to watch something grow for which you have worked so hard." At UMass Amherst, Bryan is part of Navigators, a Christian group on campus, and is involved with MercyHouse, a local church in Amherst, MA.

UMass Amherst CS will again sponsor a scholarship for 2013 MACAA winners who meet the eligibility criteria.

CS alum social in Cambridge

On September 27, 2012, CS hosted a social gathering for UMass Amherst CS alums at Google's Kendall Square, Cambridge, MA facility. CS Chair Lori Clarke and CS Alum Steve Vinter welcomed guests to the event. With over 125 people registered, the night was a great opportunity for our guests to talk with UMass Amherst Chancellor Kumble R. Subbaswamy, College of Natural Sciences Dean Steve Goodwin, fellow alums, CS faculty, and some of our CS students approaching graduation. CS Associate Professor David Jensen, Director of the Knowledge Discovery Laboratory, gave a presentation, "From Big Data to Effective Action." The CS alum social organizing committee consisted of Carla Brodley ('94), Carol Broverman ('91), David Miller ('06), Marisa Pacifico ('10), Irene Ros ('06), Steve Vinter ('85), Steve Willis ('78), and John Woods ('80). More photos at **www.cs.umass.edu/alumsocial2012**.