Form 180: Research Contributions (2000–2006)

1. Analyzing Brain Tumors (http://www.cs.ualberta.ca/~btgp)

Gliomas are diffuse, invasive brain tumors, that are often treated by irradiating the cancerous regions. Fortunately, portions of such tumors can often be detected in Magnetic Resonance (MR) images of the brain of a cancer patient. Unfortunately, there can be other "radiographically occult" tmour cells, that need to be treated as well. This project (in collaboration with Prof J. Sander (CSD) and Dr A. Murtha (MD, Radiation Oncologist) addresses the task of using prior knowledge (eg, brain templates), and various imaging techniques, to find the complete tumor volume (both visible and invisible), given a set of MR images.

The first step is to accurately locate the visible tumor region. This involved a long pipeline of pre-processing steps. to find relevant features for each voxel. **[C9]** addressed the major challenge of determining exactly which features (allow a learner to) produce an effective classifier. Another issue is combining the labels for the individual voxels, to find a "consistent" labeling for the entire image; here we extended the now-standard Condition Random Fields by using (in essense) a support vector machine; this produced better results **[C11]**, fairly efficiently **[C2]**. The overall result is one of the best tumor segmentation systems (from T1, T2 and T1c MR images). **[C6]** uses these segmented images to address the challenge of learning a system that can predict how the glioma will grow. (This uses the assumption that "where the tumor is visible tomorrow, it is invisible today".)

This work has resulted in two MSc theses (M Morris and and M Schmidt), the latter nominated for the "Best MSc thesis" prize. We have also applied for a patent for this technology, and discussed it in an Alberta-wide radio show[1]. This work is also mentioned as one of the four research "success stories" in the "Alberta Surplus" newsletter[2] that was sent to every Albertan resident.

2. WEBIC: An All-WWW Recommendation System (http://www.web-ic.com)

There are currently a large number of recommendation systems, each designed to give useful advice to the user. Essentially all such Web recommendation systems are specific to single web site; *e.g.*, Amazon.com's recommendation system is designed to suggest Amazon.com pages to users currently visiting the Amazon.com website. Such systems can base their recommendation on where other "similar" users have gone, using notions like the "support" and "confidence" of various pages and trajectories, based on the dozens to thousands of previous user visits to each page.

Our goal, however, is a system that can locate and recommend "information content (IC) pages" — pages the current user must see to complete his/her task — from essentially anywhere on the web. As most of the billions of pages have essentially no visits, or at least none that we know, support and confidence are not meaningful here. We therefore need to use a very different technology for this class of tasks.

Our WEBIC system first extracts the "browsing properties" of each word encountered in the user's current click-stream — eg, how often each word appears in the title of a page in this sequence, or in the "anchor" of a link that was followed, etc. It then uses a user- and site-independent model, learned from a set of annotated web logs acquired in a user study, to determine which of these words is likely to appear in an IC page **[C18]**,**[C25]**,**[C26]**. We then show how to use these words to find IC-pages themeselves, and demonstrate empirically that this browsing-based approach works effectively **[C15]**,**[C17]**. This work, in collaboration with Prof G Häubl (UofAlberta Business Faculty) and postdoc B Price, is the basis of T Zhu's PhD thesis. It was also described in both an Alberta-wide radio show[3], and in several articles that have been repeated in dozens of publications around the world. This research is also the foundation of a current start-up company.

[1] http://innovationalberta.com/article.php?articleid=624

[2] http://www.gov.ab.ca/home/albertasurplus/images/Surplus.pdf

[3] http://innovationalberta.com/article.php?articleid=722

3. Proteme Analyst (`http://www.cs.ualberta.ca/~bioinfo/PA`)

Proteome Analyst (PA) is a publicly available, high-throughput, web-based system for predicting various properties of each protein in an entire proteome. Using machine-learned classifiers, PA can predict, for example, the GeneQuiz general function and Gene Ontology (GO) molecular function of a protein **[J9]**,**[J7]**. In addition, PA is one of the most accurate and most comprehensive systems for predicting subcellular localization, the location within a cell where a protein performs its main function **[J6]**. These functions are organized in a hierarchy; **[C10]** investigates how a learner should exploit such hierarchical information. PA produces a Support Vector Machine classifier, which is amenable to a graphical and interactive approach to explain its predictions; transparent predictions increase the user's confidence in, and understanding of, PA **[C5]**. **[J1]** describes an extension that predicts which of an organism's proteins participate in each of a set of pathways This work is in collaboration with Profs P. Lu, D. Szafron and D. Wishart, as well as many MSc and undergrad students.

4. Budgeted Learning

Researchers often use clinical trials to collect the data needed to evaluate some hypothesis, or produce a classifier. During this "training" process, they have to pay the cost of performing each test. Many studies will run a comprehensive battery of tests on each subject, for as many subjects as their budget will allow, in a "round robin" (RR) fashion. We consider a more general model, where the researcher can sequentially decide which single test to perform on which specific individual, then use the result of this test (together with earlier information) to make the next decision; again subject to spending only the available funds. Our goal here is to use these funds most effectively, to collect the data that leads to the most accurate classifier.

We first explore the simplified "*coins version*" of this task. After observing that this is NP-hard, we consider a range of heuristic algorithms, both standard and novel, and observe that our "biased robin" approach is both efficient and much more effective than most other approaches, including the standard RR approach **[C19]**. We then apply these ideas to learning a *naïve-bayes classifier* and observed similar behavior **[C22]**. Finally, we consider the most realistic model, where *both* the researcher gathering data to build the classifier, and the user (eg, physician) applying this classifier to an instance (patient) must pay for the features used — eg, the researcher has $10,000 to acquire the feature values needed to produce an optimal $30/patient classifier. Again, we see that our novel approaches are almost always much more effective that the standard RR model **[C12]**.

This work, with postdoc O Madani, became the MSc theses of A Kapoor and D Lizotte.

5. Belief Net Algorithms

Many standard tasks inherently involve reasoning probabilistically — *e.g.*, about correlations between patient data and his disease state. (Bayesian) belief networks have become the representation of choice for many AI researchers and practioners, as they provide a succinct way to encode such probabilistic information, which allows them to reason about these situations effectively. This has led to an explosion of algorithms for both learning and reasoning with these systems.

A general Belief Net contains both a structure, which specifies what depends on what, and a set of parameters, "Conditional Probability Table values", which indicate the strength of these connections. We have investigated ways to learn the best *generative* structure **[C33]**,**[J11]**, as well as the best *discriminative* structure **[C31]**,**[C16]** for producing a good classifier. We have also provided a new, and effective, algorithm for learning the best *discriminative* parameters (for a fixed structure) **[C27]**,**[C21]**,**[J5]**.

We have also provided an effective way to compute the variance around a belief net response **[C29]** (with Statistics Professor P Hooper and students), and used this as part of a tool for combining different belief-net based classifiers **[C4]** (with PhD student Chi-Hoon Lee and postdoc Shaojun Wang). With postdoc Shaojun Wang and others, we have also used undirected probabilistic models for modeling language **[C3]**,**[C13]**.

Refereed journal papers

[J1] L. Pireddu, D. Szafron, P. Lu and R. Greiner, "The Path-A metabolic pathway prediction web server", *Nucleic Acids Research*, Volume 34 (Web Server issue), July 2006, 6 ms.

[J2] S. Damaraju, D. Murray, J. Dufour, D. Carandang, S. Myrehaug, G. Fallone, C. Field, R. Greiner, J. Hanson, C. Cass and M. Parliament, "Association of DNA Repair and Steroid Metabolism Gene Polymorphisms with Clinical Late Toxicity in Patients Treated with Conformal Radiotherapy for Prostate Cancer", *Clinical Cancer Research*, 12(8) (p2545–2554), 15 April 2006.

[J3] R. Greiner, R. Hayward, M. Jankowska and M. Molloy, "Finding Optimal Satisficing Strategies for And-Or-Trees", *Artificial Ingelligence*, 170: 19–58, January 2006.

[J4] G. Van Domselaar, P. Stothard, S. Shrivastava, J. Cruz, A. Guo, X. Dong, P. Lu, D. Szafron, R. Greiner and D. Wishart, "BASys: a web server for automated bacterial genome annotation", *Nucleic Acids Research*, July 2005; 33(Web Server issue): W455–W459.

[J5] R. Greiner, X. Su, B. Shen and W. Zhou, "Structural Extension to Logistic Regression: Discriminative Parameter Learning of Belief Net Classifiers", *Machine Learning*, special issue on "Probabilistic Graphical Models for Classification" 59(3), June 2005, p. 297–322.

[J6] P. Lu, D. Szafron, R. Greiner, D. Wishart, A. Fyshe, B. Pearcy, B. Poulin, R. Eisner, D. Ngo and N. Lamb, "PA-GOSUB: A Searchable Database of Model Organism Protein Sequences With Their Predicted GO Molecular Function and Subcellular Localization", *Nucleic Acids Research*, 2005, Vol. 33 (Database issue), D147–D153.

[J7] D. Szafron, P. Lu, R. Greiner, D. S. Wishart, B. Poulin, R. Eisner, Z. Lu, J. Anvik, C. Macdonell, A. Fyshe, and D. Meeuwis, "Proteome Analyst: Custom Predictions with Explanations in a Web-based Tool for High-Throughput Proteome Annotations", *Nucleic Acids Research*, Volume 32, July 2004, p. W365-W371.

[J8] J. Listgarten, S. Damaraju, B. Poulin, L. Cook, J. Dufour, A. Driga, J. Mackey, D. Wishart, R. Greiner and B. Zanke, "Predictive Models for Breast Cancer Susceptibility from Multiple, Single Nucleotide Polymorphisms", *Clinical Cancer Research*, 10(2725–2737), 15 April 2004.

[J9] Z. Lu, D. Szafron, R. Greiner, P. Lu, D. Wishart, B. Poulin, J. Anvik, C. Macdonell, and R. Eisner, "Predicting Sub-cellular Localization using Machine-Learned Classifiers in Proteome Analyst", *Bioinformatics*, 2004 20: 547–556.

[J10] R. Greiner, A. Grove and D. Roth: "Learning Cost-Sensitive Active Classifiers", *Artificial Intelligence*, 139:2, pp. 137–174, Sept 2002.

[J11] J. Cheng, R. Greiner, J. Kelly, D. Bell and W. Liu, "Learning Bayesian Networks from Data: an Information-Theory Based Approach", *Artificial Intelligence*, 137:1-2, pp. 43–90, 2002

[J12] D. Wishart, L. Querengesser, B. Lefebvre, N. Epstein, R. Greiner, and J. Newton: "Medical Resonance Diagnostics — A New Technology for High Throughput Clinical Diagnostics", *Journal of Clinical Chemistry*, 47:1918–1921, October 2001.

[J13] R. Greiner, C. Darken and N. I. Santoso, "Efficient Reasoning", *Computing Surveys*, 33:1 (March 2001), p. 1–30.

Refereed Conference Articles (Full paper refereed, under 1-in-3 acceptance rate)

[C1] J. Huang, D. Schuurmans, T. Zhu and R. Greiner, "Information Marginalization on Subgraphs" *Proc. 10th European Conference on Principals and Practices of Knowledge Discovery in Data (PKDD 2006)*, Berlin, Sept, 2006.

[C2] C. Lee, R. Greiner, O. Zaine and J. Sander, "Efficient Spatial Classification using Decoupled Conditional Random Fields", *Proc. 10th European Conference on Principals and Practices of Knowledge Discovery in Data (PKDD 2006)*, Berlin, Sept, 2006.

[C3] F. Jiao, S. Wang, C. Lee, R. Greiner and D. Schuurmans, "Semi-Supervised Conditional Random Fields for Segmenting and Labeling Sequence Data via Entropy Regularization", *Int'l Committee on Computational Linguistics and the Association for Computational Linguistics (COLING-ACL)*, July 2006, Sydney.

[C4] C.-H. Lee, R. Greiner and S. Wang, "Using Query-Specific Variance Estimates to Combine Bayesian Classifiers", *Int'l Conference on Machine Learning (ICML06)*, June 2006, Pittsburgh.

[C5] D. Szafron, B. Poulin, R. Eisner, P. Lu, R. Greiner, D. Wishart, A. Fyshe, B. Pearcy, C. MacDonell and J. Anvik, "Visual Explanation and Auditing of Evidence with Additive Classifiers" Innovative Applications of Artificial Intelligence (IAAI), July 2006, Boston.

[C6] M. Morris, R. Greiner, J. Sander, M. Schmidt and A. Murtha, "Classification-based Glioma Diffusion Modeling using MRI Data", *Canadian Conference on Artificial Intelligence (AI06)*, May 2006.

[C7] R. Isukapalli, A. Elgammal and R. Greiner, "Learning Multiclass Object Detection Using Binary Classifiers", *European Conference on Computer Vision (ECCV)*, May 2006.

[C8] B. Price, G. Häubl, R. Greiner and A. Flatt, "Automatic Construction of Personalized Customer Interfaces", *Proc. Int'l Conference on Intelligent User Interfaces (IUI06)*, January 2006, Sydney, Australia.

[C9] M. Schmidt, I. Levner R. Greiner, A. Murtha and A. Bistritz, "Segmenting Brain Tumors using Alignment-Based Features", *Proc. Fourth Int'l Conference on Machine Learning and Applications (ICMLA05)*, December 2005.

[C10] R. Eisner, B. Poulin, D. Szafron, P. Lu and R. Greiner, "Improving Protein Function Prediction using the Hierarchical Structure of the Gene Ontology", *2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, November 2005.

[C11] C-H. Lee, R. Greiner and M. Schmidt, "Support Vector Random Fields for Spatial Classification", *Proc. 9th European Conference on Principals and Practices of Knowledge Discovery in Data (PKDD 2005)*, Porto, Portugal, October 2005.

[C12] A. Kapoor and R. Greiner, "Learning and Classifying under Hard Budgets", *Proc. 16th European Conference on Machine Learning (ECML 2005)*, Porto, Portugal, October 2005, pp. 166–173.

[C13] S. Wang, S. Wang, R. Greiner, D. Schuurmans and L. Cheng, "Exploiting Syntactic, Semantic and Lexical Regularities in Language Modeling via Directed Markov Random Fields", *Proc. 22nd Int'l Conference on Machine Learning (ICML05)*, Bonn, June 2005, p. 953–960.

[C14] Y. Guo, D. Schuurmans and R. Greiner, "Learning Coordinate Classifiers", *Proc. 19th Int'l Joint Conference on Artificial Intelligence (IJCAI05)*, Edinburgh, Aug 2005.
Awarded "IJCAI05 Distinguished Paper Prize"

[C15] T. Zhu, R. Greiner, G. Häubl, K. Jewell and B. Price, "Using Learned Browsing Behavior Models to Recommend Relevant Web Pages", *Proc. 19th Int'l Joint Conference on Artificial Intelligence (IJCAI05)*, August 2005, p. 1589–1594.

[C16] Y. Guo and R. Greiner, "Discriminative Model Selection for Belief Net Structures", *Proc. 20th National Conference on Artificial Intelligence*, Pittsburgh, July 2005, p. 770–776.

[C17] T. Zhu, R. Greiner, G. Häubl, K. Jewell and B. Price, "Goal-Directed Site-Independent Recommendations from Passive Observations", *Proc. 20th National Conference on Artificial Intelligence (AAAI-05)*, Pittsburgh, July 2005, p. 549–556.

[C18] T. Zhu, R. Greiner, G. Häubl, K. Jewell and B. Price, "Off-line Evaluation of Web User Model", *Proc. Tenth Int'l Conference on User Modeling (UM'2005)*, August 2005, p. 337–341.

[C19] O. Madani, D. Lizotte and R. Greiner, "Active Model Selection", *Twentieth Conference on Uncertainty in Artificial Intelligence (UAI04)* pp 357–365. Banff, 2004.

[C20] I. Levner, V. Bulitko, L. Li, G. Lee and R. Greiner, "Towards Automated Creation of Image Interpretation Systems", Australian Joint Conference on Artificial Intelligence. 2003, pp 653–665.

[C21] B. Shen, X. Su, R. Greiner, P. Musilek and C. Cheng, "Discriminative parameter learning of General Bayesian Network Classifiers", 15th IEEE Int'l Conference on Tools with Artificial Intelligence (ICTAI03), Sacramento, 2003.

[C22] D. Lizotte, O. Madani and R. Greiner, "Budgeted Learning of Naive-Bayes Models", *Proc. 19th Conference on Uncertainty in Artificial Intelligence (UAI03)* Acapulco, August 2003.

[C23] R. Isukapalli and R. Greiner, "Use of Off-line Dynamic Programming for Efficient Image Interpretation", *Proc. 18th Int'l Joint Conference on Artificial Intelligence (IJCAI03)* Acapulco, August 2003.

[C24] V. Bulitko, L. Li, R. Greiner and I. Levner, "Lookahead Pathologies for Single Agent Search", *Proc. 18th Int'l Joint Conference on Artificial Intelligence (IJCAI03)* (Refereed Poster) Acapulco, August 2003.

[C25] T. Zhu, R. Greiner and G. Häubl, "An Effective Complete-Web Recommender System" *Twelfth Int'l World Wide Web Conference*, Budapest, May, 2003.

[C26] T. Zhu, R. Greiner and G. Häubl, "Learning a Model of a Web User's Interests", *Ninth Int'l Conference on User Modeling (UM03)*, Pittsburgh, June, p. 65–75, 2003.
"Best Student Paper Prize"

[C27] R. Greiner and W. Zhou, "Structural extension to logistic regression", *Proc. Eighteenth Annual National Conference on Artificial Intelligence (AAAI02)*, Edmonton, August 2002.

[C28] R. Greiner, R. Hayward and M. Malloy, "Optimal Depth-First Strategies for And-Or Trees", *Proc. Eighteenth Annual National Conference on Artificial Intelligence (AAAI02)*, Edmonton, August 2002.

[C29] T. Van Allen, R. Greiner and P. Hooper, "Bayesian Error-Bars for Belief Net Inference", *Proc. 17th Conference on Uncertainty in Artificial Intelligence (UAI-01)*, Seattle, p. 522–529, Aug 2001.

[C30] R. Isukapalli and R. Greiner, "Efficient Interpretation Policies", *Proc. 17th Int'l Joint Conference on Artificial Intelligence (IJCAI01)* Seattle, p. 1381–1387, August 2001.

[C31] J. Cheng and R. Greiner, "Learning Bayesian Belief Network Classifiers: Algorithms and System", *Proc. 14th Canadian Conference on Artificial Intelligence (CSCSI01)*, p. 141–151, Ottawa, June 2001.
RunnerUp, "Best Paper Prize"

[C32] B. Korvemaker and R. Greiner, "Predicting Unix Command Lines: Adjusting to User Patterns", *Proc. 17th Annual National Conference on Artificial Intelligence (AAAI00)*, p. 230–235, Austin, July 2000.

[C33] T. Van Allen and R. Greiner, "Model Selection Criteria for Learning Belief Nets: An Empirical Comparison", *Proc. 17th Int'l Conference on Machine Learning (ICML00)*, p. 1047–1054 Stanford, June 2000.

Other Publications I have also co-edited 1 conference proceedings (for the "International Conference on Machine Learning" ICML), published 28 other "lightly" refereed papers, and 20 posters and invited (but not refereed) publications.