RESEARCH STATEMENT

Romer E. Rosales

RESEARCH OBJECTIVES

My primary research interest is applied machine learning and data mining, with special focus on the analysis and modeling of mid to large-scale, real-world data/information sources as a means to understand and in particular exploit this information in more valuable ways. I am interested in the analysis of systems and phenomena through data and in ultimately enhancing them.

We are in the midst of a dramatic growth in the diversity and amount of data collected and stored electronically at a very low cost. As a consequence, we are now able to address and pose numerous research questions (some of them completely new) that can lead to more meaningful and innovative uses of this data. Such applications have fundamental implications in areas as varied as business, society, healthcare, and science. For example, large-scale modeling of web search information already generates revenue of billions of dollars for a few innovative companies; analysis of on-line product ratings is redefining the marketing and sales landscape; automated analysis of health records is making strides (with remarkable government support) towards changing the way medicine is fundamentally practiced; modeling patterns of economic/human activity is currently of special interest for national stability.

My research is based on the development of computational and mathematical models and algorithms capable of representing and efficiently extracting valuable information from the available data, which in turn permit the successful utilization of this data in productive ways. Based on my academic and industrial research experience, this approach is suitable for addressing the problems of interest because it allows for conceiving general, abstract models that are amenable to formal analyses and can be more systematically linked to algorithmic solutions or even new problems. Highly-available processing power can then be used for the automation of information/knowledge analysis and extraction. This approach focuses on the use of data as a means of objectively evaluating competing solutions (allowing for continuous, automated experimentation), it is grounded on solid theoretical (e.g.; mathematical and computational) principles, and it has shown ample success in addressing challenging problems as described in this statement.

In order to make progress towards these goals I focus on: (1) analyzing/developing mathematical models and algorithms aimed at large-scale experimentation with data, (2) maintaining close relationships (with industry, healthcare, government, society, academia) to help identify, validate, and support our work on legitimate problems, and (3) innovate by tackling new possibilities unlocked by new forms of data and operational ideas (new tools, applications, business models, services) based on our research findings.

PRESENT CONTRIBUTIONS

Throughout my research work in industry and academia I have addressed numerous interrelated questions. Most of this work integrates applied research with more theoretical fundamentals.

1. INFORMATION RETRIEVAL FROM FREE TEXT

Natural language text as information source is everywhere and its influence is growing: user queries, web pages, email messages, electronic health records, and more common material (such as books, scientific papers, and business documents) are just some examples. During my industry work, I led the work that allowed us to exploit the text/narrative information in health records for various information retrieval products; specifically: (1) patient/document identification [P1,P5,P10], or how to identify patients/documents based on the free text content in the health records; (2) concept search [P2], or how to search for concepts rather than keywords in health records; and (3) knowledge transfer across different healthcare institutions via active learning [12]. Each of them is a piece of the puzzle to address two fundamental questions: How can we automatically identify and efficiently learn the occurrence of an event or concept in free text? For example, how can we identify that a patient stopped smoking more than a year ago (a critical element in the treatment of Congestive Heart Failure)? The variability of natural language, the differences in the documentation across medical institutions or doctors, and the almost infinite source of medical concepts of interest make these problems extremely challenging. In the context of internet applications, this translates into many key questions like: What type of information is the user requesting (air travel, shopping, news, facts) based on his typed query? Which user review/article gave positive quality/design/usability feedback for product/company X? This type of questions is of great practical and theoretical importance today. We showed how this can be done with medical data by combining user-annotated and unlabeled text [12] with medical ontologies [P1].

2. Learning from Crowds

Data can now be easily shared, organized, and processed by a large number of entities (e.g.; Wikipedia, product ratings); an effect coined *Crowdsourcing* (Howe, 2009). In the supervised learning context, this translates to having not one labeler (normally an expert or ground-truth) but many labelers. This novel scenario renders traditional supervised learning sub-optimal but also creates exciting new problems. In [1] we showed how to efficiently utilize the labels provided by several, possibly inaccurate, annotators and in particular when annotator effectiveness or accuracy varies depending on the data instance presented (e.g.; running shoes are more accurately rated by runners). To the best of our knowledge this was the first paper to address this problem when annotator effectiveness varies in this manner. This setting addresses related problems such as the case when ground-truth is by nature not available (e.g., what are the best results for this search query) or expensive to obtain (e.g.; a biopsy can provide ground-truth about cancer lesions but at a high cost). We also showed how the model is suitable for handling missing annotators, estimating ground-truth, and evaluating labelers (without ground-truth data).

3. Active Sensing

What if we have annotator(s), but we want to use their time more efficiently by advising them which of the available observations are the most useful to annotate. This is the usual active learning scenario. Now imagine that we could also decide what new observations (not just annotations) need to be made in order to learn more efficiently. For example, in diagnosing disease, we could aid the physician by advising to perform specific medical tests in order to learn how to diagnose a particular disease or condition better. From a sensor network perspective this is equivalent to choosing what sensor to observe to decrease an uncertainty function. In [3] we presented an approach to achieve this by modeling the problem as Gaussian processes. In particular, we considered the problem of efficiently collecting data to predict patient survival for lung cancer and tumor response after chemo-radiotherapy. As part of this, we also shed light on the properties of a learning concept known as co-training using this Bayesian setting [10].

4. Knowledge Transfer

Now imagine that we can still request annotators to label data; however, we do not want to improve our performance on data from the original source, but a different one! This problem arises in various setting, in particular in the IR domain where a training corpus exists for some particular domain, but the actual domain of the retrieval task varied slightly. We showed how to do this in [12], by extending the well known Query-By-Committee approach for active learning to cases where the training and test distribution are different, and also leveraging all the available unlabeled data. The resulting formulation has a natural interpretation: it proposes to choose data points that are informative but also representative of the distribution of interest. This has many practical implications since it allows for the use of a training corpus to build retrieval models that can be automatically tuned on new retrieval tasks.

5. Leveraging Expert Domain Knowledge

Obtaining knowledge by actively requesting labels from annotators is a useful scenario. However, sometimes the knowledge may be of a different kind. Consider the case, in multi-label classification where an expert knows that a sample in class A is likely also in class B, but very unlikely in class C. It can take many data annotations to learn this relationship in the usual manner, so why learn by labeling data if an expert has already done the work for us? Consider the case where we know that if a patient has disease A, he is also likely to have disease B, but not as likely to have disease C. Or, when a keyword has meaning A, it is likely to also have meaning B, but not meaning C (when the word *apple* is used to denote a fruit, it is also likely to refer to an *apple tree* or to *apple pie*, but not likely to refer to the *Apple computer company*). In [2] we introduced large margin classifier that automatically learns the underlying inter-label structure and furthermore allows the controlled incorporation of this form of knowledge. We applied this to various settings including disease diagnosis and gene function prediction.

6. Approximate Inference in Graphical Models

An effective way to represent the structure of a probability distribution is by means of a graph, where variables and their dependencies are associated to nodes and edges in the graph. The computational complexity of inference is tied to the tree-width w of the graph, $O(2^w)$. This is a limiting factor regarding the use of probabilistic models in large, real-world applications; thus, approximations are important. We introduced a general approximate inference algorithm based on an iterative process that sequentially eliminates variables and obtains an approximation of the marginal distribution of the remaining variables [21]. We showed that this basic decomposition step, followed by a projection onto the exponential family distribution space, is exact for decimatable graphs. After eliminating a variable, the *neighborhood marginal* is optimally approximated within the family of tree-structured distributions. This procedure is more accurate than mean field and tree-structured ADF. The advantage lies primarily in the inclusion of dependencies induced by marginalization but not represented in the original graph. The ability to maintain such dependencies through approximate marginalizations may underlie the superior empirical results. We also introduced a distributed strategy, placing it in the context of the sum-product algorithm; an exciting, new research direction waiting to be tackled.

7. ℓ_1 Regularization for Learning Sparse Models

In machine learning and statistics it is often the case that only a small fraction of covariates (features) suffices to build a model comparable in performance with others that use a larger number of features. However, the problem of obtaining an optimal subset of features for a linear classifier is NP-hard and not practical in most scenarios. The use of the ℓ_1 penalty provides an implicit feature selection strategy. In [13,38], we presented an exhaustive comparison of optimization techniques to solve the problem of minimizing $f(x) = L(x) + \lambda\|x\|_1$, a twice-differentiable loss function L subject to ℓ_1 regularization. We also introduced generalized versions of several existing methods proposed for specific loss functions and new strategies for this general problem. The new approaches are based on a smooth convex approximation for the ℓ_1 regularizer (improving bounds on the approximation error between iterates) and on constrained optimization (including a specialized gradient projection method). Our experiments across 13 optimization strategies, in 7 different scenarios showed that the proposed approaches are more general and often outperform classical and state-of-the-art specialized methods in both accuracy and speed. This work contributed to an important practical area in machine learning and has enhanced our work in §1. To the best of our knowledge this is the most exhaustive study on the use of the ℓ_1 norm for building sparse models in machine learning.

8. Structure Learning in Graphical Models

This work ties our previous two contributions. The inference problem assumes that we know the *structure* of the distribution, represented by a graph G. Well, if G is not appropriate for the problem at hand, then even exact inference will provide useless results. Normally this structure is given *by hand*, by an expert. A few approaches had tried to address the problem of estimating G for Markov random fields (MRF). In [8] we concentrated on learning the structure of discriminative models (such as Conditional Random Fields) and non-binary MRFs; to the best of our knowledge, the first paper to do so. We developed a general method for learning (sparse) graph structures for this kind of models via block-ℓ_1 regularization. The formulation involved casting the task as a convex optimization problem and introducing, what we called, a ℓ_1 ℓ_∞ regularization. We developed a new efficient approach to finding the global minimum of the resulting objective function, in particular for cases in which the Hessian is intractable to compute/store using standard methods.

9. Other

My experience in related areas include: computer vision, modeling body motion/pose[18,33], tracking people in video, image rendering by transferring the statistical properties of one image onto another[24]; clustering data using neighborhood structure (rather than the usual similarity measures)[22]; gene expression analysis [22]; privacy preserving data mining and Cox regression [7], learning to rank [19], learning to detect heart motion abnormalities from ultrasound images[15] (obtaining a best paper award), web-based interfaces for searching/annotating and managing electronic heath records [P2], and other areas.

Future Directions

In the long-term, the goal of my research is to make a clear impact on how information is analyzed and utilized at a large scale by creating suitable machine learning/data mining models and algorithms. In the short-term, I believe that a successful approach for achieving this must include working on real, significant problems and potentially replacing/extending them with new ones; establishing purposeful collaboration in appropriate areas; and making continuous improvements while exploring novel directions of research. These elements are conducive to funding but also to keeping the research program fresh. In pursuing these goals, the key machine learning areas to explore involve models and algorithms (current and novel) directed towards the full utilization of the *knowledge of the masses* (e.g.; learning from the interaction of multiple users with a system, including user annotations), the efficient use of limited expert knowledge (e.g.; actively seeking for the important information: labels or data), the appropriate incorporation of widely available data (e.g.; unlabeled data); and the better exploitation of (currently sub-utilized) unstructured data sources (e.g.; free text). The ability to efficiently work with large datasets (in terms of storage and processing) requires especial emphasis as it is at the core my research goals.

More specifically, in the immediate term I will continue on-going research work and collaborations while also establishing new directions of research. I will build on the exposure I have gained during my five years in industry and propose new work with appropriate emphasis in the medical domain. Considerable government funding opportunities in this area, in particular based on stimulus grants, will be beneficial. In addition, my work in research and development projects, with various healthcare institutions (like the Mayo Clinic) and experience with granting agencies such as the Department Health and Human Services (HHS), and National Institutes of Health (NIH), and the Department of Defense (DoD) will also be of assistance. Due to its importance, I plan to make text analysis from various data sources a key research area.

1. UNDERSTANDING FREE TEXT The understanding of free text (not limited to natural language) is considered one of the critical technologies in the future of information services because free text is a natural communication interface, large amounts of information in many aspects of life are and continue to be stored as free text, it is at the intersection of other areas such as speech recognition and language translation, and numerous applications will benefit from better text understanding. Not surprisingly, leading companies in the information industry are spending billions in R&D to build the next generation automated text processing products and tools (Google, IBM, Microsoft, Amazon, to name a few). I believe that my current research based on learning to recognize text concepts is among the most promising approaches because while it is focused, at the same time it is at the core of addressing the most general problems. I wish to continue my work in understanding health records but also utilize other sources such as on-line data, particularly user generated communication and interaction with retrieval systems (see next).

2. LEVERAGING THE POWER OF THE CROWDS Data generated by multiple entities about the same general subject (product review, news article, search result, opinion, medical diagnosis, and question in general) is becoming common place. As a focus area in machine learning, we are just at the beginning. The data availability will be so large that this will be a critical area to allow us to process information more efficiently and usefully. Some essential problems that still need to be tackled include choosing annotators efficiently, modeling annotator biases, evaluation of annotators, and estimating unavailable ground-truth.

3. COMPARATIVE EFFECTIVENESS RESEARCH (CER) In healthcare, CER focuses on using the available data and medical expertise to compare treatments and strategies to improve health, clearly a data mining problem that also involves representing available expert medical knowledge. There is so much interest in this that the federal government allocated ~$1.1 billion in funding from the American Recovery and Reinvestment Act (ARRA) due to its potential at improving health and making healthcare more efficient. The technical problem is to discover or corroborate what (treatment) works, what does not, for which patients, for what illness, at what cost. Machine learning and statistics already have a powerful set of tools that can be the basis for approaching this problem. For example, measures of probabilistic dependency in information theory can be used to determine what treatments have worked [5]. This can be used as the basis to refine models, for example by making them sub-population specific [4]. Based on my research with electronic records, I believe that a solution for this problem will combine taping into available data sources, extracting information from unstructured text, and collaborating with government/healthcare providers.

4. PERSONALIZED MEDICINE (PM) This directly brings us to another key area in the future of medicine. Personalized medicine is based on the premise that people are different and so, in general, will respond differently to the same treatment. A medicine can help someone while produce a dangerous side effect to someone else. The antibody Herceptin is only effective for patients whose cancer tissue (over) expresses the protein called HER2 (~20% of the population). The commercialization of this test/treatment combo is a success story for PM. I wish to create many more of these successes by establishing the modeling and algorithmic foundations for large scale discovery of such patient-treatment-outcome relationships. One of the technical problems is the existence of too many confounding factors, and thus models that seek simpler explanations of the dependent variable, such as [7] may be of great interest. In establishing a successful research plan, the Personalized Medicine Coalition can be a valuable partner for collaboration.

5. MODELING AND ORGANIZING INFORMATION STREAMS Information is generated at astonishing rates. Thus, analyzing and organizing this on-line information stream can give us a glimpse into the world today. I want to understand to what extent it is possible to accurately use this information to predict possible user information needs/preferences, epidemics, economic collapse, security treats. It is very likely that we would need to tap into our research on modeling crowds in order to approach this problem.

6. APPROXIMATE INFERENCE IN GRAPHICAL MODELS Our previous work in Focused Inference [21] represents a rather unique avenue for approaching the general inference problem in statistics. Because of its practical implications and potential, I wish to study more carefully the properties and uses of the distributed algorithm we introduced.

These areas encapsulate my current thinking about this evolving field. This thinking is heavily influenced by my experience in both industry and academia where I have had the fortune to learn and contribute. I strongly believe in their complementarity in the advancement of science and technology. These experiences have given me a rather unique vantage point for choosing research directions, for collaboration, for building support, and for developing new ideas. I am very confident and at the same time excited about the effect that machine learning and data mining technology will have on the future of business and society.

PUBLICATIONS IN TECHNICAL PROCEEDINGS AND JOURNALS (PEER REVIEWED)

1. Yan Yan, Romer Rosales, Glenn Fung, Jennifer Dy, and Gerardo Hermosillo. **Modeling Annotator Expertise: Learning when Everybody Knows a Bit of Something.** In Proc. *International Conference on Artificial Intelligence and Statistics* (AISTATS) 2010.
2. Volkan Vural, Glenn Fung, Jennifer Dy, and Rómer Rosales. **Multi-Class Classifiers and Their Underlying Shared Structure.** In Proc. *International Joint Conference on Artificial Intelligence* (IJCAI) 2009.
3. Shipeng Yu, Balaji Krishnapuram, Rómer Rosales, and R. Bharat Rao. **Active Sensing.** In Proc. *International Conference on Artificial Intelligence and Statistics* (AISTATS) 2009.
4. Marianne Müeller, Rómer Rosales, Harald Steck, Sriram Krishnan, Bharat Rao, and Stefan Kramer. **Subgroup Discovery for Test Selection: A Novel Approach and its Application to Breast Cancer Diagnosis.** *In Proc. Intelligent Data Analysis* (IDA) 2009.
5. Marianne Müeller, Rómer Rosales, Harald Steck, Sriram Krishnan, Bharat Rao, and Stefan Kramer. **Data-Efficient Information-Theoretic Test Selection.** *In Proc. Conference on Artificial Intelligence in Medicine* (AIME) 2009.
6. Rómer Rosales, Glenn Fung, and Wei Tong. **Automatic Discrimination of Mislabeled Training Points for Large Margin Classifiers**. *(Snowbird) Machine Learning Workshop* 2009.
7. Shipeng Yu, Glenn Fung, Rómer Rosales, and R. Bharat Rao. **Privacy-Preserving Cox Regression for Survival Analysis.** To Appear in Proc. *Knowledge Discovery and Data Mining* (KDD) 2008.
8. Mark Schmidt, Glenn Fung, Kevin Murphy, and Rómer Rosales. **Discriminative Structure Learning in Random Fields and an Application for Heart Motion Abnormality Detection**. To Appear in Proc. *Computer Vision and Pattern Recognition* (CVPR) 2008.
9. R. Bharat Rao, Glenn Fung, and Rómer Rosales. **On the Dangers of Cross-Validation**. An Experimental Evaluation. In Proc. *SIAM Data Mining* (SDM) 2008.
10. Shipeng Yu, Balaji Krishnapuram, Rómer Rosales, Harald Steck, and R. Bharat Rao. **Bayesian Multi-View Learning**. In Proc. *Neural Information Processing Systems* (NIPS) 2007.
11. O. Yakhnenko, L. Lita, R. Rosales, and S. Niculescu. **Principled Generative-Discriminative Hybrid Hidden Markov Model**. *In Neural Information processing Systems, Workshop on Representations and Inference on Probability Distributions* (NIPS) 2007.
12. Rómer Rosales, Praveen Krishnamurthy, and R. Bharat Rao. **Semi-supervised Active Learning for Modeling Medical Concepts from Free Text**. In Proc. *International Conference on Machine Learning Applications* (ICMLA), 2007.
13. Mark Schmidt, Glenn Fung, and Rómer Rosales. **Fast Optimization Methods for L1 Regularization: A Comparative Study and Two New Approaches**. In Proc. *European Conference on Machine Learning* (ECML) 2007.
14. Glenn Fung, Rómer Rosales, and R. Bharat Rao. **Feature Selection and Kernel Design via Linear Programming**. In Proc. *International Joint Conference on Artificial Intelligence* (IJCAI) 2007.
15. Maleeha Qazi, Glenn Fung, Sriram Krishnan, Rómer Rosales, Harald Steck, R. Bharat Rao, Don Poldermans, and Dhanalakshmi Chandrasekaran. **Automated HeartWall Motion Abnormality Detection from Ultrasound Images using Bayesian Networks**. In Proc. *International Joint Conference on Artificial Intelligence* (IJCAI) 2007. [Distinguished Paper Award]

16. T. Wilmes, K. Bohy, A. Gilson, R. Rosales, S. Krishnan, S. Niculescu, M. Qazi, F. Rahmanian, W. Landi, B. Rao. **Automated Chart Abstraction Can Provide Highly Accurate Data Extraction For Clinical Quality Measures: Assessment of REMIND for CMS Heart Failure Measures.** *Circulation 114: II_864-a. American Heart Association* 2006.
17. Rómer Rosales and Glenn Fung. **Learning 'Sparse' Metrics via Linear Programming**. In Proc. *Knowledge Discovery and Data Mining* (KDD) 2006.
18. Rómer Rosales and Stan Sclaroff. **Combining generative and discriminative models in a framework for articulated pose estimation**. *International Journal of Computer Vision* (IJCV). Volume 67 (3) p. 251-276, 2006.
19. Glenn Fung, Rómer Rosales, and Balaji Krishnapuram. **Learning Rankings via Convex Hull Separation**. In Proc. *Neural Information Processing Systems* (NIPS) 2005.
20. Venk Gottipaty, Rómer Rosales, Prasad Aloni, John Beard, Paul Zimmermann, Linda Adams, R. Bharat Rao. **Automated Identification of MADIT-II Eligible Patients Using REMIND Artificial Intelligence Software.** *Circulation 111:e310-e359. American Heart Association* 2005.
21. Rómer Rosales and Tommi Jaakkola. **Focused Inference**. In Proc. *International Workshop on Artificial Intelligence and Statistics* (AISTATS) 2005.
22. Rómer Rosales, Kannan Achan, and Brendan Frey. **Learning to Cluster using Local Data Topology**. In Proc. *International Conference in Machine Learning* (ICML), 2004.
23. Rómer Rosales, Kannan Achan, and Brendan Frey. **Unsupervised Image Translation**. In Proc. *International Conference on Computer Vision* (ICCV), 2003.
24. Rómer Rosales, Kannan Achan, and Brendan Frey. **Translating Images by Unsupervised Estimation of Switching Filters**. Invited paper. In Proc. *IEEE Statistical Signal Processing* (SSP), 2003.
25. Rómer Rosales and Brendan Frey. **Learning Generative Models of Affinity Matrices**. In Proc. *19th Conference on Uncertainty in Artificial Intelligence* (UAI), 2003.
26. Rómer Rosales and Stan Sclaroff. **A Framework for Heading-Guided Recognition of Human Activity**. *Computer Vision and Image Understanding* (CVIU), 2003.
27. Rómer Rosales and Stan Sclaroff. **Algorithms for Inference in the Specialized Mappings Architecture** 2002. In *Proc. IEEE International Conference on Automatic Face and Gesture Recognition* (FG2002), 2002.
28. Rómer Rosales and Stan Sclaroff. **Learning Body Pose via Specialized Maps**. In *Proc. Neural Information Processing Systems 14*, 2001.
29. Rómer Rosales, Matheen Sidiqqui, Joni Alon, and Stan Sclaroff. **3D Body Pose through Virtual Cameras**. In *Proc. IEEE Computer Vision and Pattern Recognition* (CVPR), 2001.
30. Stan Sclaroff, George Kollios, Margrit Betke, and Rómer Rosales. **Motion Mining.** Proc. 2nd International Workshop on Multimedia Databases and Image Communication, 2001.
31. Rómer Rosales, Vassilis Athitsos, Leonid Sigal, and Stan Sclaroff. **3D Hand Pose Reconstruction using Specialized Mappings**. In *Proc. IEEE International Conference on Computer Vision* (ICCV), 2001.
32. Rómer Rosales and Stan Sclaroff. **Specialized Mappings and the Estimation of Human Body Pose from a Single Image**. In Proc. *IEEE Workshop on Human Motion*, 2000.

33. Rómer Rosales and Stan Sclaroff. **Inferring Body Pose without Tracking Body Parts**. In *Proc. IEEE Computer Vision and Pattern Recognition* (CVPR), 2000.
34. Rómer Rosales and Stan Sclaroff. **Learning and Synthesizing Human Body Pose and Motion**. In *Proc. IEEE International Conference on Automatic Face and Gesture Recognition* (FG2000), 2000.
35. Rómer Rosales and Stan Sclaroff. **Trajectory Guided Recognition of Actions**. In Proc. *SPIE.,* 1999.
36. Rómer Rosales and Stan Sclaroff. **3D Trajectory Recovery for Tracking Multiple Objects and Trajectory Guided Recognition of Actions**. In *Proc. Computer Vision and Pattern Recognition* (CVPR), 1999.
37. Rómer Rosales and Stan Sclaroff. **Improved Tracking of Multiple Humans with Trajectory Prediction and Occlusion Modeling**. In *Proc. IEEE Workshop on the Interpretation of Visual Motion* (CPVR), 1998.

Papers under Review

38. M. Schmidt, G. Fung, and R. Rosales. **Fast ℓ_1 Regularization. Current and New Optimization Algorithms**. *Journal of Machine Learning Research* (JMLR).
39. S. Yu, B. Krishnapuram, Romer Rosales, H. Steck, and R. B. Rao. **Bayesian Multi-view Learning**. *Transactions Pattern Analysis and Machine Intelligence* (PAMI).

PATENTS (GRANTED/APPLICATIONS)

P1. A. Pandya, R. Rosales, B. Rao, H. Steck. **Medical Ontologies for Computer Assisted Clinical Decision Support** [Granted 2009]
P2. R. Rosales, S. Chung, F. Farooq, G. Fung, B. Krishnapuram, B. Rao, J. Weis, S. Yu. **Automated Patient/Document Identification and Categorization For Medical Data** [Provisional 2009]
P3. R. Rosales, J. Dy, Y. Yan, and G. Fung. **Automatic Labeler Assignment**. [Provisional 2009]
P4. R. Rosales, M. Bundschus, B. Krishnapuram, F. Rahmanian, B. Rao, S. Yu. **System and Method for Scoring and Ranking Quality Improvement Factors in Patient Care** [Pending 2009]
P5. R. Rosales. **Method for Automatic Labeling Of Unstructured Data Fragments from Electronic Medical Records** [Pending 2009]
P6. S. Yu, B. Krishnapuram, R. Rosales, B. Rao. **Active Multi-Modality Sensing for Training Classifiers** [Provisional 2009]
P7. O. Yakhnenko, R. Rosales, S. Niculescu, L. Lita. **System and Method for Text Tagging and Segmentation Using a Generative/Discriminative Hybrid Hidden Markov Model** [Pending 2008]
P8. R. Rosales, S. Niculescu, W. Landi, B. Rao, P. Giang, S. Krishnan. **Patient Data De-Identification by Obfuscation** [Pending 2008]
P9. R. Rosales, G. Fung, M. Schmidt, S. Krishnan, B. Rao. **Learning Classifiers for Computer-Aided Diagnosis using Multiple-Labeler Data Analysis [**Pending 2008]
P10. R. Rosales, P. Krishnamurthy, B. Rao, H. Steck. **Learning or Inferring Medical Concepts from Medical Transcripts** [Pending 2008]
P11. M. Yetisgen-Yildiz, S. Niculescu, R. Rosales, B. Rao, S. Krishnan**. Automated Interpretation and Replacement of Date References in Unstructured Text** [Pending 2007]
P12. B. Rao, S. Krishnan, W. Landi, R. Rosales, S. Niculescu, F. Rahmanian, H. Steck. **Quality Metric Extraction and Editing for Medical Data** [Pending 2007]
P13. R. Rosales, M. Müeller, S. Krishnan, B. Rao. **Guiding Differential Diagnosis through Information Maximization** [Pending 2006]