*i*WML Invited Talks

List of Abstracts

1. **Prateek Jain** (Microsoft Research)
   Compressive Sensing
   Abstract: The area of compressive sensing is concerned with encoding and decoding signals with certain special structures such as sparsity, low-rank etc. Recently, compressive sensing area has seen a lot of progress and several novel algorithms and tools have been developed for both encoding as well as decoding problem. In this talk, I will define and motivate compressive sensing of sparse signals. I will then talk about some of the successful approaches for this problem and will try to give a proof sketch of these methods. I will then discuss some of the other related problems in this area such as low-rank matrix sensing, multi-signal sensing with block sparsity etc. I will conclude the talk with some open problems and challenges in the area.

2. **Vikas Raykar** (IBM Research)
   Learning from Crowds
   Abstract: Crowdsourcing has recently become popular among scientists as an effective way to collect large-scale experimental data from distributed workers. With the advent of crowdsourcing services (Amazon's Mechanical Turk being a prime example) it has become quite easy and inexpensive to acquire labels from a large number of annotators in a short amount of time. This setting introduces a number of interesting learning problems that are currently attracting great interest in the general machine learning community. In this talk I will summarize some of my research in this area, which is at the interface of machine learning and crowdsourcing.

3. **Sunita Sarawagi** (IIT Bombay)
   Graphical Models for Structure Extraction and Annotation
   Abstract: Ever since the successful deployment of Conditional Random Fields for Named Entity Recognition (NER) and other sequential extraction tasks, graphical models have been lauded as a powerful tool for solving various structure extraction tasks. In this talk I will cover how they have been used for three non-conventional extraction and annotation tasks. First, we will see how to annotate raw tables on the Web with links to manually created ontologies. We propose a joint graphical models to annotate table cells with entities that they likely mention, table columns with types from which entities are drawn for cells in the column, and relations that pairs of table columns are seeking to express. Second, we show how to answer multi-attribute record extraction queries by jointly mapping columns of several raw tables via a graphical model. Third, we show how to jointly segment several lists on the Web while exploiting overlap in their content. We will observe that in all cases, graphical models provide an elagant solution to the problem of combining diverse clues in a single unified framework.

4. **Sudeshna Sarkar** (IIT Kharagpur)
   Machine Learning for Recommender Systems
   Abstract: Recommender systems recommend items from an inventory to a user to support users in their decision-making. This may be done by predicting the rating or preference that user would give to an item in a given context. The prediction is often done by exploiting past history of users, user similarity or item characteristics, and is personalized for a given user. User feedback is either implicit or explicit via click, ratings or preference relations. Researchers have worked on offline modelling and rating prediction, as well as for online modelling in a live system. The objective of a recommender algorithm is to learn user preferences through continuous feedbacks in order to optimize certain utility functions. The utility functions may be click rates, advertising revenue, etc. The challenging machine learning problems in this domain are dealing with sparse data, high dimensions, and doing online recommendation. We will discuss the various machine

learning frameworks that are used for offline systems. The two broad approaches are content based approaches and collaborative filtering (CF). CF approaches using user-user and item-item matrix, as well as matrix factorization methods will be discussed. We will also discuss the use of preference relations in our work for this task. For online systems the explore/exploit paradigm will be presented. We will also discuss ways to deal with temporal pattern of item popularity, and shift in user interests. We discuss various evaluation metrics and methods, and challenges for evaluation of recommender systems. RMSE, MAE and ROC, Hit rate are certain standard metrics. But one may also look for other factors like freshness, novelty and diversity. We will also discuss ranking which is important for top-k recommendation tasks. Finally we discuss the current challenges in this area.

5. **Pushpak Bhattacharyya** (IIT Bombay)
   Natural Language Processing and Machine Learning: Points of Synergy and Divergence
   Abstract: We present in this talk insights into the simultaneous synergistic and divergent relationship between ML and NLP. These insights are borne out of our long standing work on problems at the interface of NLP and ML, specifically word sense disambiguation and sentiment analysis. We will present our findings in these two areas, to show how BOTH linguistics and language phenomena on one hand and learning on the other are required for effective solution. Neither is dispensable. The talk will conclude with the description of some of our experiences in generation of parallel corpus through crowd sourcing for statistical machine translation, where NLP helps in the specification and setting up of the task and ML helps incremental automatisation for efficiency.

6. **B. Ravindran** (IIT Madras)
   A very brief introduction to Multi-arm Bandit Problems
   Abstract: With online commerce becoming a multi-billion dollar business there has been interest in "online" learning of user behavior and preferences. Typically these are cast as reinforcement learning problems, specifically the "explore-exploit" distillation - multi-arm bandits (MABs). In this talk I will motivate the need to study bandits and talk briefly about the various solution concepts that exist for MABs. I will also introduce the notion of PAC analysis of MABs and discuss the analysis of a couple of MAB algorithms - a naive one and also the current champion. We will also look at regret analysis for bandit problems and a very quick look at contextual bandits. This talk assumes no prior knowledge of reinforcement learning.

7. **Rahul Garg** (Opera Solutions)
   Anonymization of High Dimensional Data
   Abstract: In this talk I will describe some new anonymization techniques for high dimensional data sets with large number of records. These methods combine the advantages of k-anonymity and perturbation methods of anonymization. The dataset is first clustered using a nearest neighbor approach such that there are at least k points in each cluster. The nearest neighbors for every point are obtained using cover tree data structure. These clusters are then perturbed to obtain anonymized data, which is expected to retain most of the statistical properties of the original data.

8. **Srujana Merugu** (Amazon)
   Predictive Latent Factor Models for Large Scale Dyadic Data
   Abstract: Predictive modeling of response variables that are functions of a dyad, i.e., a pair of interacting entities possibly drawn from two different sets, is an important problem encountered in several application domains such as social networks, expertise modeling, recommendation systems, advertising, etc. Examples of dyadic response variables include the relationship between two users in a social network, expertise of a user for a task, preference of a user for a product, and click-through rate of an ad placed on a page. In this talk, we will consider some of the real-world problem characteristics and challenges associated with dyadic data prediction as well as limitations of traditional approaches. I'll present the key ideas in predictive latent factor modeling approaches that combine the benefits of supervised and unsupervised techniques as well as specific instantiations using co-clustering and generalized linear models.

9. **C V Jawahar** (IIIT Hyderabad)
   Visual Categorization
   Abstract: In this talk, we focus on a class of recognition problems in computer vision. The focus will be on the challenges, motivation for using machine learning, and some of the interesting problems that are emerging in this space. The talk will also connect the tasks in CV to the developments in machine learning with focus on SVMs and Kernel Methods.

10. **Dhruv Mahajan** (Microsoft Research)
    Analysis and Evaluation of Distributed (Linear) ML Algorithms
    Abstract: In this talk, we focus on the problem of learning linear classifiers when data is distributed across several machines. There have been several distributed learning methods proposed in the literature. With the advent of cloud computing and distributed machine platforms becoming part of cloud computing, it is important and useful for such platform users to make certain choices along three key dimensions, namely, system, application and machine learning (ML). In this work, we carry out systematic evaluation of various methods considering various aspects along these dimensions. In machine learning, there are a variety of methods proposed in the literature, and we categorize these methods under (1) ensemble and parameter mixing, (2) iterative parameter mixing, (3) statistical query model and (4) parallel block optimization. And, these methods are analyzed and compared along the different aspects in each dimension.

11. **Kaushik Sinha** (University of Wichita)
    Randomized partition trees for exact nearest neighbor search
    Abstract: The k-d tree was one of the first spatial data structures proposed for nearest neighbor search. Its efficacy is diminished in high-dimensional spaces, but several variants, with randomization and overlapping cells, have proved to be successful in practice. In this talk I will discuss three such schemes. Main focus of the talk will be our recent result that demonstrates that the probability that these schemes fail to find the nearest neighbor, for any data set and any query point, is directly related to a simple potential function that captures the difficulty of the point configuration. I will then discuss how we can bound this potential function in two situations of interest: the first, when data come from a doubling measure, and the second, when the data are documents from a topic model.

12. **Sumeet Agarwal** (IIT Delhi)
    Machine Learning for Systems Biology
    Abstract: In recent years, biology has increasingly moved into a high-throughput, big data era, with large numbers of experiments being conducted to probe the functioning (and malfunctioning) of biological systems. One particular formal approach that has been widely employed to deal with the complex nature of such systems is the abstraction of modelling them as networks or graphs, where the nodes may represent genes/proteins/cells/tissues/organs, as the case may be. In particular, such approaches have been used at the subcellular level to model control and regulation relationships between different genes or proteins. Some of key questions of interest in such systems have focused on the relationship between structure and dynamics: do certain sorts of structural properties of these networks enable or facilitate certain kinds of dynamical patterns of, say, protein concentration levels inside a cell? Given experimental measurements of these dynamics, can we learn or reverse engineer the cellular circuitry? Here we will survey some of the recent work in this area and also present some of our own results and ideas for how machine learning can be useful in unpacking the complexity of life.

13. **V. Vijaya Saradhi** (IIT Guwahati)
    Learning from Cricket Text Commentary
    Abstract: Strategies are important in any game to improve the chances of winning. When a game involves teams of individuals, areas of strength and improvement must be understood for each player. Cricket, a team game, is one such example where the game plans are formulated in the team by considering the observations of the team coach and technical analysts. In this talk, I will focus on arriving at team strategies by identifying the weaknesses of individual players, thereby suggesting specific areas of improvement for each individual. In particular, cricket text commentary is used for mining weaknesses of individual players. Domain specific features are extracted from the

text commentary and apply supervised learning techniques, relational classifiers and unsupervised learning techniques to arrive at rules which help in building strategies. The obtained features are validated through local experts with satisfactory performance.