INDO-AUSTRALIAN MEETING ON:
MODELLING LARGE-SCALE LINKED DATA
(29TH NOVEMBER—1ST DECEMBER, 2014)

AIMS

The goal of the meeting is to bring together computational researchers from Australia and India–to the best of our knowledge, for the first time– working on the modelling of linked data. The outcomes will identify application areas of relevance to science, industrial and human well-being in both countries that can benefit significantly from the use of advanced techniques for data analysis. Conversely, it will contribute to a frontline area of ICT by allowing theory, implementation and application of statistical relational learning to develop in lock-step.

BACKGROUND

The existence and rapid growth of large-scale scientific and industrial databases have brought into focus the need for methods that enable the discovery of trends and predictive patterns in data, and communicating them in a manner designed to provoke insight. The immediate implication of the existence of these databases for computational tools for data analyis is that they have to have some mechanism for dealing with data volumes that range from tens to thousands of gigabytes at present, and many more in the future (the so-called "Big Data" problem). A separate, and harder, aspect concerns the nature of the data. The data are no longer simply values of known (pre-defined) features, but are in the form of observations of several inter-related variables. Here we will interchangeably call this kind of data "linked" or "relational". The combination of large data volumes and linked data that effectively do not satisfy statistical assumptions of independent, identically distributed observations now confronts data-analysts in both science and industry. Here are some examples:

Biology. Automated genome sequencing techniques have made available millions of nucleotide sequences comprising entire genomes of organisms, at a rapidly increasing rate. These are now augmented by taxonomic databases, protein sequence databases, gene expression data, metabolic pathway data and so on.

Health. The last decade has seen a huge increase in the uptake of e-health. Every day, healthcare institutions collect large amounts of clinical, administrative and social information about their patients. Usage of this information has the potential to hugely improve quality of care and patient safety while reducing costs.

Biometrics. With increasing security concerns, there is a pressing requirement of robust and efficient biometric technologies. Recognition systems based on large databases of containing fingerprint, face and iris records are currently being used in controlled conditions but performance in unconstrained environments with large scale matching remains a challenge.

Sensor Networks. With the advent of pervasive computing, our personal spaces (e.g., smartphones, homes, offices, cars and cities) are being increasingly instrumented through miniaturised sensors. These sensors collect highly personalised multi-modal data streams from our surroundings resulting in very large-scale datasets for processing.

Social Networks. Data extracted from social networks are often represented as `triples (e.g. RDF) codifying relations. These data are currently interpreted as graphs, and are giving rise to a number of new kinds of problems (link prediction is an example).

Social Sciences. Pervasive digital records of social and economic interactions at regional and national levels offer the possibility for a dramatically improved understanding of individual lives, societies and organisations. To fully realise this potential, however, the difficult questions relating to scientific validity of conclusions drawn from (incomplete) observational data must be addressed.

On the ***applications*** front, problems in these areas share a number of characteristics: (a) They are at the forefront of research and development in science, industry and social sciences; (b) They often require model-construction methods that can handle large amounts of data that do not satisfy the usual "i.i.d" assumptions; and (c) They require model-construction methods that can handle many sources of complex, but not completely relevant or accurate domain knowledge. On the ***conceptual*** front, methods of dealing with data-analysis problems of this kind employ techniques from mathematical modelling, machine learning[1] and probabilistic modelling[2]. On the ***implementation*** front, the advent of "cloud" computing has simplified the procurement of resources on demand for executing data-analysis jobs. The MapReduce programming paradigm, introduced by Google, has also made it easy to compose data-analysis workflows that can be scaled on multiple resources easily. MapReduce, however, reduces every data processing job to an embarrassingly parallel program. Scaling methods to model linked data to work elastically over cloud resources is currently an important research topic.

OUTCOMES

Specific outcomes: (a) A report on the potential of methods for modelling large-scale linked data; (b) Identification of at least one front-line data analysis problem of relevance to science, industry and human well-being that may benefit significantly from the analysis of such data; and (c) Identification of the any extensions needed to existing analysis methods to make them better suited to the front-line problems in the application areas. **General outcomes**: (a) Greater exposure for researchers in the application areas to the potential of relational learning; (b) Application-driven development of theory and implementations for researchers in the area of relational learning; (c) Increased co-operation between researchers working in the area of linked-data analysis in the two countries; and (d) Exposure to young researchers in India to an exciting new area of data-driven model construction in Computer Science.

[1] Used here broadly to denote data-driven model-construction methods that can handle linked or relational data.

[2] Used here to denote methods that accommodate uncertainty in linked data or in domain knowledge, or in both.

MEETING DETAILS: 29 NOVEMBER, 2014 TO 1 DECEMBER, 2014

VENUE: BOARD ROOM, 5TH FLOOR, ACADEMIC BLOCK, IIIT-DELHI (29, 30 NOVEMBER)

GETTING THERE: HTTP://WWW.IIITD.AC.IN/HOME/CONTACT

PARTICIPANTS

AUSTRALIA	INDIA
• Arcot Sowmya (UNSW) • Mike Bain (UNSW) • Srikumar Venugopal (UNSW) • Salil Kanhere (UNSW) • Wei Wang (UNSW) • Thamali Lekamge (UNSW) • Sandeep Kaur (UNSW)	• Ashwin Srinivasan (IIITD) • Mayank Vatsa (IIITD) • Richa Singh (IIITD) • K Sriram (IIITD) • Maya Ramanath (IITD) • Srikanta Bedathur (IBM Research) • Parag Singla (IITD) • Mausam (IITD) • Sumeet Agrawal (IITD) • Madhulika Mohanty (IITD) • Lovekesh Vig (JNU) • Pankaj Malhotra (TCS Research) • Gautam Shroff (TCS Research)

Local Contact: Sunil Gupta (Mobile: +91 9711 004896)

AGENDA

Day 1	Day 2
9:00am --- 9:30am: Tea and Coffee 9:30am--10:00am: Welcome, organisational remarks **10:00am--11:00am: Session 1** 1.1 Analysis of X-Ray Crystallographic Images 1.2 Applications of ML to Some Problems in Biometrics 11:00am--11:30am: Tea and Coffee **11:30am--12:30pm: Session 2** 2.1 Detecting Anomalies Via Stacked LSTM Networks 2.2 Forecasting state failure and mass atrocities **12:30pm--1:30pm: Lunch** **1:30pm--2:30pm: Session 3** 3.1 Linked Data Queries to Natural Language Text 3.2 Keyword Search in Graphs 2:30pm--3:00pm: Tea and Coffee **3:00pm—4pm: Session 4** 4.1 Topological priors and gene expression data 4.2 Scalable Protein Similarity Search on Clouds **7:30pm: Dinner (Neyvedyam, Hauz Khas Village)**	**9:00am --- 9:30am**: Tea and Coffee 9:30am--10:00am: Organisational remarks **10:00am--11:00am: Session 5** 5.1 Classification in the Era of Big Data 5.2 Preserving Privacy in a Camera-Rich World 11:00am--11:30am: Tea and Coffee **11:30am--12:30pm: Session 6** 6.1 Large-scale semi-automatic entity resolution 6.2 OpenIE for Relation Extraction **12:30pm--1:30pm: Lunch** **1:30pm--3:00pm: Session 7** 7.1 Markov Logic Nets and their Applications 7.2 Feature-construction with ILP 7.3 Mathematical modelling for time-series data **3:00pm--3:15pm: Closing** **3:30pm: Visit to Humayun's tomb** **7:30pm: Dinner (Park Balluchi, Deer Park)**
Day 3	
Meetings and Discussion as required on Research and Future Collaboration	

ABSTRACTS OF TALKS

ANALYSIS AND CLASSIFICATION OF X-RAY CRYSTALLOGRAPHY IMAGES

B.M. THAMALI LEKAMGE, ARCOT SOWMYA AND JANET NEWMAN

X-ray crystallography is one of the main techniques used to analyse and understand the three dimensional structure of a protein at atomic resolution. The process of acquiring X-ray crystallography images demands the production of an appropriate crystalline sample of the protein, and then irradiation of the crystal with X-rays. Diffraction patterns obtained via this process are then studied to produce an illustration of the electron density associated with the protein. These crystallisation experiments usually require hundreds and thousands of individual experiments. The experiments are also time dependent and require monitoring. Automatic imagers are used to record the steps of these crystallisation experiments and they generate millions of images. Crystallographers are expected to analyse the resulting images as the image generators play no part in interpretation. Between 10,000 and 20,000 images are generated daily in the medium-throughput crystallization laboratory within CSIRO (Collaborative Crystallisation Centre, C3, http://crystal.csiro.au/). According to the estimates of crystallography experts, around 50 million images are generated via crystallisation experiments annually. As the amount of data generated is huge and there is not enough resources to analyse every single image by the naked eye, many experimental images are never inspected. Additionally, although numerous experiments are carried out, only a few produce crystals. Some experiments can lead into different non-crystalline paths such as producing "precipitation", "skin" or other results. Even though crystal production is the main goal, the crystallography experiments, can also provide important information such as conditions conducive to crystallisation, the phase behaviour of a particular sample and stability of a protein. Therefore, it is important to identify the end product in each image.The experiments are time dependent, with the time varying from a few hours to a fewweeks. The automatic image capture process records different stages of the experiments, but the time intervals between captured frames are not identical, and may vary from hours to weeks. The number of frames in the experiments also varies from 8 to 16 in each.

TOPICS IN THE ANALYSIS OF LINKED-DATA

WEI WANG, UNSW

In this talk I will be covering a range of topics related to the analysis of linked data. I intend to talk on:

- **Large-scale semi-automatic entity resolution**. To effectively construct knowledge graphs from semi-structured or unstructured data, an essential task is to recognise if two references to entities refer to the same real-world entity or not. An example is "new york" and "big apple" refers to the same city, while two mentions of "George Bush" may refer to different person. This is a recurring task under many different names, such as entity resolution, record linkage, deduplication, data cleansing, etc. We are interested in methods that combines rules, machine learning techniques, unlabelled data, and external signals (e.g., exploring external knowledge base/graph). We are also interested in the system aspects of the problem as we will need to deal with large volume of data.
- **Querying and exploring graph data.** Large graph data (e.g., knowledge graph, databases, social networks) typically do not have fixed schemas, hence making it hard to query (both in terms of coverage and efficiency). We have already built new indexing and query processing methods to support flexible, schema-agnostic, and error-tolerant queries in real-time (as Google does for Web Search). We are looking to support other interesting query/explore patterns that may arise from various data analysis, mining, and visualation applications.
- **Theory/Methods.** Approximate neares neighbor / low-distortion embedding. Our recent VLDB'15 paper substantially improves the state-of-the-art LSH-based method for approximate NN queries, enabling indexing billions of high-dimensional points on a commodity PC (while previous approaches from MIT+Intel uses 100 powerful servers). We are interested in further exploring the theoretical aspects of LSH /dimensionality reduction / embedding, as well as applications where our techniques may be beneficial.
- **Active Learning/Crowd Sourcing.** We are interested theories and methods related to active learning / crowd-sourcing, especially with application to entity resolution, where class imbalance poses additional challenges.

LINKED DATA QUERIES TO NATURAL LANGUAGE TEXT (OR: LEARNING TO SPEAK WITH SEMANTICS)

SRIKANTA BEDATHUR, IBM RESEARCH – INDIA

Linked Data, particularly the kind derived from natural language texts usinginformation extraction (IE) techniques, has grabbed the fascination of everyonein recent times. Notwithstanding the foundational work over years done in academic and research labs, it has taken the internet industries by storm. Starting from multi-billion dollar companies like Google/Microsoft/IBM to niche playersin medicine, education, entertainment and similar domains have been investingtheir money and efforts towards the goal of going from "Big Text" data to "Big Knowledge". The knowledge thus extracted from text is typically modeled as a big (hyper) graph, represented as an RDF or a property-graph. Resulting graph integrates information extracted from multiple sources, can be used to intelligently intepret, connect, understand and answer complex analytics queries. However,this is only one side of the story – it is also equally crucial to understand howto present the results effectively. Current approaches range from simply showing the resulting subgraph of the knowledge graph to developing a customized interface that utilizes the results of "canned" queries. We believe that the time is ripe to explore more general, natural forms of presentation.

One of the recent directions that researchers have started exploring is to synthesize natural language answers based on the results of computation over knowledge graphs. I would like to briefly outline these efforts, including our recent work in this direction, discuss the future directions to pursue with the participants.

KEYWORD SEARCH ON GRAPHS

MADHULIKA MOHANTY AND MAYA RAMANATH, IITD

Graph-structured data on the web is now massive as well as diverse, ranging from social networks, web graphs to knowledge-bases. Effectively querying this graph-structured data is non-trivial and has led to research in a variety of directions – from structured queries, keyword and natural language queries, automatic translation of these queries to structured queries, etc. In this talk, we will discuss a class of queries called “relationship queries”, which are usually expressed as a set of keywords (each keyword denoting a named entity). We will also look at the various challenges and shortcomings in the state of the art and ways to improve it.

OpenIE FOR RELATION EXTRACTION

MAUSAM, IITD

Traditional information extraction systems, systems that extract entities, relations, and events from text, suffer three major limitations: 1. They require that all types be known in advance, a requirement that severely limits their coverage of available information. An analyst charged with analyzing the contents of new data must discover important facts in a timely fashion, not merely review instances of a few pre-specified fact types. 2. Targeting a new type of fact with these systems, or adapting them to new genres and domains, requires a substantial annotated development corpus. Management of the tedious annotation process multiples the development overhead. 3. The running time of these systems scales with the number of relations (|R|). Each sentence needs to be processed |R| times, which is inefficient, since most sentences contain only a few relations.

These limitations can be particularly costly for real-world applications, where the assumption that all important types of information can be specified in advance is highly incorrect. In response, my previous group at University of Washington, Seattle pioneered the Open Information Extraction (Open IE) paradigm for information extraction. Open IE extracts an unbounded number of relations, requires no training data, and has running time insensitive to the number of relations involved. Open IE systems learn a general model of how relations are expressed that is based on domain-independent linguistic features, capturing detailed relational information in general fashion. The simplest form of an Open IE extraction is a triple involving a relation string and two arguments. An extended version of Open IE (the system we use in this work) extracts more than 2 arguments per relation. Moreover, it also provides temporal and location annotation to the relevant arguments. The simplicity of this format supports efficient indexing and querying, even against Web-scale volumes of information involving multiple textual sources and millions of relations.

FORECASTING THE ONSET OF GENOCIDE AND POLITICIDE

BENJAMIN E GOLDSMITH, CHARLES R BUTCHER, DIMITRI SEMENOVICH, ARCOT SOWMYA)

Prevention of genocide is one of the most important challenges before the international community. We present what is, to the best of our knowledge, the first published set of out-of-sample forecasts of genocide and politicide based on a global dataset. Our goal is to produce a prototype for a real- time model capable of forecasting one year into the future. While building on the current literature, we take several important steps towards better forecasting. We implement a two-stage modelling approach that considers both the likelihood of instability and the likelihood of genocide in a single estimate. Our sample is restricted only by the available data, rather than through selection of controlled cases, since such an approach would not be available for forecasting into the future. We explore factors exhibiting variance over time to improve year-on-year forecasting performance. And we produce annual lists of at-risk states in a format that should be of use to policy makers seeking to prevent such mass atrocities. We employ sparse additive models which are both flexible and maintain interpretability of the results. Overall our out-of-sample forecasts for 1988-2003 predict 81.8% of genocide onsets correctly while also predicting 78.7% of non-onset years correctly, an improvement over a previous study using a case-control in-sample approach. We produce 16 annual forecasts based only on previous years' data, which identify seven of eleven cases of genocide/politicide onset within the handful of the top 5% of at-risk countries per year. This represents a considerable step toward useful real-time forecasting of such rare events. We conclude by suggesting ways to further enhance predictive performance.

DETECTING TIME SERIES BASED ANOMALIES VIA STACKED LSTM NETWORKS

LOVEKESH VIG (JNU)
PANKAJ MALHOTRA, PUNEET AGARWAL, GAUTAM SHROFF (TCS RESEARCH)

Long Short Term Memory (LSTM) networks have been demonstrated to be a powerful technique for sequence prediction, particularly when sequences contain long term dependencies. Due to their ability to remember correlations across a large number of time steps, LSTM networks have found prominent applications in fields such as speech recognition, handwriting generation, and sequence classification. Recently, it has been shown that stacking recurrent hidden layers enables the learning of higher level temporal features, which in turn enables faster learning with sparser representations. In this paper, we apply a stacked LSTM network to the problem of anomaly/fault detection. Traditional approaches to anomaly detection utilize a stochastic modeling based approach wherein an abrupt change in the statistics of the sequence over a time window is used to indicate an anomaly. However, such approaches frequently miss contextual anomalies that may not affect the statistics of the time series. We show how anomaly detection performed with stacked LSTM networks can better capture anomalies in ECG signals, power consumption patterns, and shuttle valve sensor readings. Additionally, we utilize LSTM networks to model normal engine behavior by training on an anonymized engine sensor dataset. Deviations from normal engine behavior are subsequently detected via prediction error residuals. The results obtained are encouraging and indicate that stacked LSTM's may be a viable technique for anomaly/fault detection. To the best of our knowledge, this is the first instance of stacked LSTM networks being used for anomaly detection.

COMBINING TOPOLOGICAL PRIORS WITH GENE EXPRESSION AND INTERACTION DATA FOR THE INFERENCE OF GENE REGULATORY NETWORKS

SUMEET AGARWAL, IIT DELHI

Inference of gene regulatory networks from microarray expression data has been an extremely active area of research in the last decade or so. A wide variety of models have been proposed for representing the dynamical relationships of genes, ranging from full-fledged coupled differential equations to Boolean networks which treat genes as on/off switches. One significant issue such efforts have faced is the problem of degeneracy: the quantity of data is insufficient to specify a single solution, meaning that a number of different solutions fit the data equally well. In a few cases people have attempted to include prior information in the form of known interactions between genes or proteins in order to address this issue: but the results have not been definitive, and such binary interaction data is also known to be quite noisy and unreliable.

Here we will explore the possibility of going beyond just binary edgewise priors, to include information on higher-level network topology. For instance, it is known that certain patterns of connections between groups of 3 or 4 nodes (known as motifs) are over-represented in gene regulatory networks. We also know of the existence of various kinds of interesting structure at even larger scales, such as modularity (densely connected subgroups) and 'scale-freeness' (a small number of nodes with very high connectivity). Given such structural priors gleaned from known gene regulatory networks, can we make use of them to restrict the search space when inferring new networks for less-studied organisms? What kind of learning framework can be used to appropriately combine such prior knowledge with the training data?

We will present some of our preliminary work towards this end, and suggest some directions for further study.

SCALABLE PROTEIN SIMILARITY SEARCH ON CLOUD RESOURCES

SRIKUMAR VENUGOPAL, UNSW

Improved efficiencies in generating and storing data have created challenges around storage, management and analysis of massive datasets. Distributed computing technologies such as cloud computing and MapReduce programming framework have provided means to address these challenges. Cloud computing provides a technology model where resources can be provisioned on demand, thereby enabling scaling up of storage and processing to match increasing data volumes. MapReduce enables programming data analysis workflows that can run across distributed resources without having to deal with issues such as node failures. However, many of the algorithms and tools used to process data in many domains still have a single machine abstraction at the core of their assumptions. Therefore, they may be inefficient when applied in a distributed system. Furthermore, these were not designed to take advantage of infrastructures that could scale on demand. Hence, there is a case for reimagining and redesigning traditional data analysis methods to be scalable across large data volumes and dynamic number of resources.

In this talk, we will outline an effort that was taken in the bioinformatics domain to scale up protein sequence similarity searching methods to meet the demands of metagenomics. Metagenomics is the study of uncultured microorganisms from their habitats and involves the analysis of genetic material gathered directly from environmental samples. It involves the computationally challenging task of interpreting extremely large, complex datasets containing fragments of sequences from different organisms. A typical metagenomics workflow starts with taking genetic samples and metadata from a particular environment. A sequencer extracts genetic sequences from the samples. These are then processed in order to perform gene prediction and annotation as well as classification. In the last step, protein sequences obtained from the sample are compared with sequences stored in known databases in order to obtain meaningful information. This is performed by comparing the residues (amino acids) at each position in the sequence strings and is a form of similarity searching. Tools such as BLAST (Basic Local Alignment Search Tool) are widely employed for this purpose. In recent years, so-called Next-Generation Sequencing Machines have boosted the speed at which genomes can be sequenced from environmental samples. This in turn has led to a deluge in the amount of available metagenome data that is estimated to double every 14 months. Similarity searching is a key bottleneck in the workflow with requirements for clusters with hundreds or thousands of computational cores to execute BLAST on terabytes of data. BLAST workloads are computationally expensive and can take weeks on single machines when large query sets of the order of a million sequences and above are involved.

This talk will describe a method to significantly improve the scalability of sequence similarity search by using multiple hash functions to generate sketches, or signatures, of each sequence from the input as well as from the reference databases. The distance between the signatures approximates the similarity between the sequences. This is the basic premise of Locality Sensitive hashing (LSH), a method that has been used effectively to scale searching over petabyte-sized collections of Web documents in search engines. We have implemented this method in MapReduce to produce a toolkit called ScalLoPS (Scalable Locality Sensitive Protein Similarity Search). This talk will report our experience with ScalLoPS and the results of evaluating it on Amazon cloud resources with metagenomic datasets. We will also use it as an example of both the advantages and pitfalls of redesigning single machine algorithms for a distributed infrastructure.

CLASSIFICATION IN THE ERA OF BIG DATA

SYED ARSHAD AND ARCOT SOWMYA, UNSW

Data Classification is a well-known problem that has extensively been esearched. However, as we move towards the age of even larger scale data, many of the current assumptions about the problem environment become void, thereby rendering several traditional algorithms ineffective in the face of new challenges posed, specially by ubiquitous data streams. This presentation reviews some the these challenges in the context of adaptive learning algorithms. Real and virtual concept drifts are explained to help understand the workings of the basic adaptive windowing approach (ADWIN). Issues regarding error analysis and evaluation techniques for these approaches are also addressed.

CHALLENGES TO PRESERVING PRIVACY IN A CAMERA-RICH WORLD

SALIL KANHERE, CSE, UNSW

Cameras are now pervasive on consumer devices, including smartphones, tablets, and new wearable devices like Google Glass. The ubiquity of these cameras will soon create a new era of visual sensing applications, for example, devices that collect photos and videos of our daily lives, augmented reality applications that help us navigate the world around us, and enable safety and health applications like documenting law enforcement's interactions with the public and helping dementia patients to recall memories. Control of sensors and sensor data is challenging enough with smartphones but becomes even more difficult with wearable devices. These devices have the ability to collect large volumes of information, often in an opportunistic manner. Moreover, visual imagery is extremely rich in context, such that leaking image data could be particularly damaging Current solutions for managing sensors and sensor data are not sufficient, and the magnitude of this problem will only grow as these devices become more popular and more powerful. This research couples a sociological understanding of privacy with an investigation of technical mechanisms to address these needs. Issues such as context (e.g., capturing images for public use may be okay at a public event, but not in the home) and content (are individuals recognizable?) will be explored both on technical and sociological fronts: What can we determine about images, what does this mean in terms of privacy risk, and how can systems protect against risk to privacy? Specific research challenges to be addressed include formulating technical means through image and context analysis to improve the privacy of people captured in images; exploring the unique privacy needs of camera owners and how image and contextual analysis can improve privacy; and developing image transformations to afford privacy as well as enable novel applications using the cloud and crowdsourcing.

As a starting point, we will explore the use of semantic content within an image for enforcing privacy policies: some key attributes of the content and context surrounding the person can be estimated from the visual information seen by a wearable camera, and then these attributes can be used to apply user-defined policies that could react to changing environments. This could range from applying an appropriate sharing policy to an image without requiring user intervention (e.g. "Do not share any pictures of my children with the public"), to modifying device behavior based on visual surroundings (e.g. disabling recording whenever a bystander who does not wish to be recorded is in the field of view), to configuring automatic actions whenever specific visual content is detected (e.g. to automatically notify authorities if a missing child is observed). We will also explore the possibility of reactive security actions based on visual stimuli from a wearable camera, freeing users from having to actively manage their privacy and instead triggering security actions as needed. As we enter an era of pervasive cameras, we hope to spur further research in the area of security and privacy that leverages visual sensing.

RELATIONAL FEATURE CONSTRUCTION WITH ILP

ASHWIN SRINIVASAN, IIITD

ILP-based feature construction has been a well-studied area. Briefly, a feature *f* ---more correctly, a feature function---is a Boolean function of a relational instance *x*, defined relative to background knowledge *B*. This usually includes syntactic constraints in the form of a language specification *L* and may also include semantic constraints *I*.

In this talk, we will be concerned with discriminatory tasks (classification problems). For these, we assume an ILP system is provided with examples of the form *Class(a,b)* where *a* is a relational data instance (like a graph), and *b* is some class value. For simplicity, we will assume data instances are drawn from some set *X* and the classes are drawn from a set *Y*. Then, the ILP engine constructs first-order clauses each of the form *Class(x,c)*$\leftarrow$ *Cp(x)* where *x* is a variable denoting data instances and c is some class value from the set of classes *Y* (in general *x* is not restricted to a single object and can consist of arbitrary tuples of objects). For example, *Class(<m1, m2>,true)* might denote that the toxicity of molecule *m1* was higher than that of molecule *m2*. Here, we have adopted terminology from Ratnaparkhi and *Cp: X* $\rightarrow$*{0,1}* denotes a ``context predicate''. A context predicate corresponds to a conjunction of literals that evaluates to TRUE (1) or FALSE(0) for any element of *X* (for meaningful features we will usually require that a *Cp* definition contains at least one literal). Clauses found in this manner by the ILP engine are converted to a Boolean feature using a one-to-one mapping as follows: f_j(x) = 1 if Cp_j(x) = 1 (and 0 otherwise). There is now a growing body of research that suggests that augmenting any existing features with ILP-constructed relational ones can substantially improve the predictive power of a statistical model. In this talk, I will describe some of this work, including work on graph-like data and large-scale streams of relational data.