INDUSTRY EVENTS

<table>
<thead>
<tr>
<th>Schedule</th>
<th>Speaker</th>
<th>Affiliation</th>
<th>Talk Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Industry Session 1</td>
<td>Chih Jen Lin</td>
<td>National Taiwan University</td>
<td>Experiences and Lessons in Developing Machine Learning and Data Mining Software</td>
</tr>
<tr>
<td>Nov 4, 2014</td>
<td>Limsoon Wong</td>
<td>National University of Singapore</td>
<td>Some Issues that are Often Overlooked in Big Data Analytics</td>
</tr>
<tr>
<td>13:30-15:00</td>
<td></td>
<td>Emerald 1</td>
<td></td>
</tr>
<tr>
<td>Industry Session 2</td>
<td>Soumen Chakrabarti</td>
<td>IIT Bombay</td>
<td>Robust Interpretation and Ranking for Telegraphic Entity-seeking Web Queries</td>
</tr>
<tr>
<td>Nov 5, 2014</td>
<td>Alex Somla</td>
<td>Google and CMU</td>
<td>Design Principles for Machine Learning at Scale</td>
</tr>
<tr>
<td>13:30-15:00</td>
<td></td>
<td>Emerald 2</td>
<td></td>
</tr>
<tr>
<td>Industry Session 3</td>
<td>Wei-Ying Ma</td>
<td>Microsoft</td>
<td>Building a Scalable System and Algorithms for Machine Comprehension of Text</td>
</tr>
<tr>
<td>Nov 6, 2014</td>
<td>Tong Zhang</td>
<td>Baidu</td>
<td>Modern Optimization Techniques for Big Data Machine Learning</td>
</tr>
<tr>
<td>13:30-15:00</td>
<td></td>
<td>Emerald 2</td>
<td></td>
</tr>
</tbody>
</table>

However, developing a good and easy-to-use machine learning package is never easy. It involves issues ranging from algorithms, implementations, and many design considerations. In this talk, we discuss the experiences in developing two machine learning packages LIBSVM and LIBLINEAR, which have been popular in both academia and industry. We demonstrate that the interaction with users leads us to develop a useful guideline for the practical use of SVM. This process also helps to identify some important research problems. For example, the decision to study and then support multi-class SVM was essential in the early stage of developing LIBSVM. We then discuss different design considerations for machine learning packages. While LIBSVM and LIBLINEAR are small-scale machine learning software, lessons learned in their development are still useful for larger packages including those for big data. In the end I will mention some challenges and concerns in developing machine learning software.

**Bio:** Chih-Jen Lin is currently a distinguished professor at the Department of Computer Science, National Taiwan University. He obtained his B.S. degree from National Taiwan University in 1993 and Ph.D. degree from University of Michigan in 1998. His major research areas include machine learning, data mining, and numerical optimization. He is best known for his work on support vector machines (SVM) for data classification. His software LIBSVM is one of the most widely used and cited SVM packages. For his research work he has received many awards, including the ACM KDD 2010 and ACM RecSys 2013 best paper awards. He is an IEEE fellow, an AAAI fellow, and an ACM distinguished scientist for his contribution to machine learning algorithms and software design.

**Talk 2: Some issues that are often overlooked in big data analytics**

**Presenter:** Limsoon Wong
National University of Singapore (http://www.comp.nus.edu.sg/~wongls/)

**Abstract:** The arrival of the "big data" era is opening up new avenues in business, healthcare, etc. Much attention has been paid to scaling challenges arising from the huge increase in volume, velocity, and variety. Not as much attention has been paid to non-scaling-related issues that affect a number of fundamental assumptions in current bioinformatics and statistical analysis approaches. Having more data is tremendously helpful in some analysis procedures. At the same time, having more data can also make the same analysis procedures fail in fundamental ways. We discuss some examples of these issues and how they might be fixed.

**Bio:** Limsoon Wong is a professor of computer science at the National University of Singapore. He currently works mostly on knowledge discovery technologies and their application to biomedicine. He is a Fellow of the ACM, named for his contributions to database theory and computational biology. He was a co-recipient of the ICDT 2014 Test of Time Award for his work...
on naturally embedded query languages. Limsoon serves on the editorial boards of Information Systems, Journal of Bioinformatics and Computational Biology, Biology Direct, and Drug Discovery Today. He co-founded Molecular Connections (India), and oversaw its steady growth over the past decade to some 900 research engineers, scientists, and curators.

**Industry Session 2 (Nov 5, 2014, 13:30-15:00, Emerald 2)**

**Session Chair:** Rajeev Rastogi

**Talk 1: Robust Interpretation and Ranking for Telegraphic Entity-seeking Web Queries**

**Presenter:** Soumen Chakrabarti

IIT Bombay (http://www.cse.iitb.ac.in/~soumen/)

**Abstract:** Over half of Web queries mention an entity (e.g., violin) or a type (e.g., scientist) and hint at relations between them (e.g., played). During the last several years, search engines have made great strides in "understanding" what the "telegraphic" query "scientist played violin" means, directly responding with a list of scientists rather than the usual "ten blue links". To enable this, three challenging feats must be accomplished. First, a knowledge graph (KG) must be built and maintained, that captures large numbers of (scientist) and instances/entities (Einstein). Freebase has over 45 million entities and types and over 2.4 billion relations connecting them, but is still tiny compared to all entities and relations expressed on the Web. Second, spans of Web text, such as "young Albert started taking fiddle lessons" need to be linked to suitable nodes in the KG. Typical Web crawls today exceed many billion pages, and a typical page may merit hundreds of annotations.

Third (and this is where we will focus), given ungrammatical queries without clues from syntax, case, quoted phrases, and punctuation, we need robust methods to find segments that can be mapped to diverse purposes: naming one or more entities, hinting at one or more desired target types, and expressing relations. (This is just one example template; entity-oriented queries can encode a variety of "query schema" using this "telegraphic query language" people use with search engines.) Finally, candidate answers (Einstein) need to be aggregated and ranked. Because any segmentation may be incorrect, answers need to be aggregated over multiple segmentations. And because the KG is always work in progress, signals from the KG and corpus need to be integrated smoothly.

We will describe techniques that learn how to segment queries from distant supervision, and achieve significant accuracy gains compared to more NLP-centric and corpus-agnostic techniques. We show that jointly exploiting the KG and corpus has synergistic effects that supersede using any one in isolation.

(The entire talk is based on public domain material. Any opinion is my own and not endorsed by IIT Bombay or Google.)

**Bio:** Soumen Chakrabarti joined in 1999 the Department of Computer Science and Engineering at the Indian Institute of Technology, Bombay, where he has been an Associate professor since 2003. In Spring 2004 he was Visiting Associate professor at Carnegie-Mellon University. He was a Research Staff Member at IBM Almaden Research Center from 1996 to 1999, where he worked on the Clever Web search project and led the Focused Crawling project. He has published widely. He won the best paper award at WWW 1999 and was coauthor on the best student paper at ECML 2008. His work on keyword search in databases got the 10-year influential paper award at ICDE 2012. He is a fellow of the Indian National Academy of Engineering, holds eight US patents on Web-related inventions, and is also author of one of the earliest books on Web search and mining. He has served as technical advisor to search companies and vice-chair or program committee member for WWW, SIGIR, SIGKDD, VLDB, ICDE, SODA and other conferences, and guest editor or editorial board member for DML and TKDE journals. He has served as program chair for WSDM 2008 and WWW 2010. His current research interests include integrating, searching, and mining text and graph data models, exploiting types and relations in search, and dynamic personalization in graph-based retrieval and ranking models.

**Talk 2: Design Principles for Machine Learning at Scale**

**Presenter:** Alex J. Smola

Google and CMU (http://alex.smola.org)

**Abstract:** Scaling machine learning poses a number of challenges ranging from the issue of statistical modeling to designing distributed algorithms and dealing with a large amount of possibly unreliable machines. In this talk I will address how to design fast function classes and algorithms based on randomization, I will illustrate the interplay between randomization and problem distribution, and I will discuss how these goals can be accomplished in a large scale open source machine learning tool, the parameter server.

**Bio:** Alex J. Smola is now a researcher at Google and a professor in the Machine Learning Department at Carnegie Mellon University. He studied physics in Munich at the University of Technology, Munich, at the Universita degli Studi di Pavia, and at AT&T Research in Holmdel. During this time he was at the Maximilianium Munchen and the Collegio Ghislieri in Pavia. In 1996 he received the Master's degree at the University of Technology, Munich and in 1998 the Doctoral Degree in computer science at the University of Technology Berlin. Until 1999 he was a researcher at the IDA Group of the GMD Institute for Software Engineering and Computer Architecture in Berlin (now part of the Fraunhofer Gesellschaft). After that, he worked as a Researcher and Group Leader at the Research School for Information Sciences and Engineering of the Australian National University. From 2004 he worked as a Senior Principal
Researcher and Program Leader at the Statistical Machine Learning Program at NICTA. From 2008 to 2012 he worked at Yahoo Research. Dr. Smola’s primary research interests include scalability of algorithms, kernels methods, statistical modeling, and applications in user modeling, document analysis, temporal models, and modeling data at scale.

**Talk 2: Modern Optimization Techniques for Big Data Machine Learning**

**Presenter:** Tong Zhang

Baidu (http://stat.rutgers.edu/home/tzhang/)

**Abstract:** Many modern big-data machine learning problems encountered in the internet industry involve optimization problems so large that traditional methods are difficult to handle. The complex issues in these large scale applications have stimulated fast development of novel optimization techniques in recent years. I will present an overview of progresses made by the machine learning community to handle these large scale optimization problems, as well as challenges and directions.

**Bio:** Dr. Tong Zhang is currently directing Baidu’s Big Data Lab, and is a professor at Rutgers University. Previously he has worked at IBM T.J. Watson Research Center in Yorktown Heights, New York, and Yahoo Research in New York city. Tong Zhang is a leading researcher in machine learning and statistics. He has published over 100 papers with more than ten thousand citations, and has an H-index of 55. He is a fellow of American Statistical Association, and has been in the editorial boards of leading machine learning journals and program committees of top machine learning conferences. Tong Zhang received a B.A. in mathematics and computer science from Cornell University and a Ph.D. in Computer Science from Stanford University.

**Talk 1: Building a Scalable System and Algorithms for Machine Comprehension of Text**

**Presenter:** Wei-Ying Ma

Microsoft (http://research.microsoft.com/en-us/people/wyma/)

**Abstract:** In recent years, we have seen dramatic improvements in deep learning, knowledge (entity) mining, and computing infrastructure that are providing powerful capabilities to process and understand text data at an unprecedented scale. We now have the ability to learn big statistical models from large amounts of data and build comprehensive symbolic knowledge graphs from the Web. We have technologies to learn different types of representations for text, including continuous vector space models based on semantic embedding, graph representations based on entity and relationship extraction, and discrete representations using information retrieval based approaches. We have distributed graph engines capable of serving a large-scale knowledge graph on which natural language understanding and generation can be performed in real time. In the industry, the rise of intelligent software such as Microsoft Cortana gives us opportunities to close the human feedback loop and create never-ending knowledge mining and machine learning to monotonically improve the precision and coverage of the various text representations. By building a scalable system and algorithms to leverage and integrate all these capabilities, we could unify many fundamental building blocks in natural language understanding, such as semantic parsing and linking of text string onto knowledge graph and reasoning and inference with common sense to decipher its meaning. In this talk, I will introduce some of our work in this area, including word and knowledge embedding, automatic construction of an entity graph for the enterprise, and real time knowledge serving and knowledge-based question answering.

**Bio:** Dr. Wei-Ying Ma is an Assistant Managing Director at Microsoft Research Asia, where he oversees multiple research groups, including Web Search and Data Management, Natural Language Computing, Knowledge Mining, Machine Learning, and Internet Economics and Computational Advertising. He and his team of researchers have developed many key technologies that have been transferred to Microsoft’s Applications and Services Group, including Bing Search Engine and Microsoft Advertising. He has published more than 250 papers at international conferences and in journals. He is a Fellow of the IEEE and a Distinguished Scientist of the ACM. He served on the editorial boards of ACM Transactions on Information System (TOIS) and is a member of the International World Wide Web (WWW) Conferences Steering Committee. In recent years, he has served as program co-chair of WWW 2008 and as general co-chair of ACM SIGIR 2011.