

## **A Short Course**

### **Data Mining Techniques and Applications**

**organized by Pattern Recognition and Machine Intelligence Association (PREMIA)**

**10 & 11 May 2011 (Tue & Wed), 9.00 am – 5.30 pm**

**Seminar Room 3/Programming Lab 2, School of Computing, Computing Drive, NUS,  
Singapore 117417**

This course comprises a tutorial on data mining techniques, three invited talks by data mining practitioners on challenging applications of data mining, and a hands-on session on a data mining software.

#### Registration Fees

Members of PREMIA: S\$450.00

Non-members: S\$550.00

Student members of PREMIA: S\$200.00 (Limited seats)

Student non-members: S\$250.00 (Limited seats)

The registration fee includes course notes, refreshments, and a one-year free PREMIA membership subscription. For non-members, entrance fee to PREMIA membership is waived.

#### Registration Procedure

Please register online at [www.premia-sg.org](http://www.premia-sg.org). Make your cheque payable to PREMIA and send it to PREMIA's treasurer Dr. Lu Shijian as follows: Dr. Lu Shijian, Institute for Infocomm Research, 1 Fusionopolis Way, #21-01 Connexis, South Tower, Singapore 138632.

If you wish to do online fund transfer, please email Dr. Lu Shijian ([slu@i2r.a-star.edu.sg](mailto:slu@i2r.a-star.edu.sg)) for PREMIA account information.

*Registration will close on 6 May 2011. However, due to the lab space limitation for hands-on, the registration may close before the deadline if the class limit is reached. Please register early to avoid disappointment.* If your organization needs an invoice for the course fee registration, or if you encounter any problem during registration, please contact Miss Zhang Xiaoyan at [ZHAN0292@e.ntu.edu.sg](mailto:ZHAN0292@e.ntu.edu.sg)  
PREMIA reserves its right to cancel the course due to circumstances beyond its control.

#### **Day 1 (10 May 2011, Tuesday): Tutorial on Data Mining Techniques**

##### **Venue: Seminar Room 3, Level 2, Com 1, School of Computing**

The tutorial will cover an introduction to analytics, 'what is it?' 'what is the motivation and the need for such a domain?'. Illustrating from real-world examples a high level mapping as to how analytics can provide in depth insight and enhance the decision making process. We will further proceed to introduce the concept of data mining, going into detailed description of various methodologies (such as decision trees and regression methods) as well as advanced elements such as feature selection etc, making the connection with the previously highlighted problems.

9:00 am – 10:30 am	Intro to Analytics (What is analytics? What is the motivation from a real-world perspective?)
10:30 am – 11:00 am	Tea Break
11:00 am – 1:00 pm	Intro to Machine Learning (Linear Discriminant Learning, Perceptron Algorithm, SVM and Kernel Methods and non linearity)
1:00 pm – 2:00 pm	Lunch (on your own)
2:00 pm – 3:30 pm	Data Mining methodologies (Goals of data mining, classification, association analysis, cluster analysis, anomaly detection)

3:30 pm – 4:00 pm	Tea Break
4.00 pm – 5:30 pm	Features (feature selection including hyperspectral data analysis, and how they relate to real-world problems, dimensionality reduction, subspace representation); Moving from research into practice.

Biodata: Dr. David R. Hardoon is Principal, Analytics at SAS Singapore. His areas of expertise include, but are not limited to, data mining, information retrieval, knowledge discovery, pattern recognition and machine learning. These have been applied across a wide cross-disciplinary scope including problems/applications in music, medical analysis, retail, time sequence analysis, aerospace, taxonomy, content based information retrieval, vision and finance. He received a B.Sc. in Computer Science and Artificial Intelligence with first class honors at Royal Holloway, University of London within the Department of Computer Science in 2002 and a PhD in Computer Science in the field of Machine Learning from the University of Southampton in the Information: Signals, Images, Systems research group in 2006. He has also received the PhD PASCAL label award for his active participation in the PASCL Network of Excellence. He is currently an Adjunct Assistant Professor at the School of Computing, National University of Singapore, a Honorary Senior Research Associate at the Centre for Computational Statistics & Machine Learning, University College London and is also a visiting Research Fellow at Institute of Psychiatry, King’s College London. Dr. Hardoon is also a member of the Pattern Analysis, Statistical Modeling and Computational Learning (PASCAL) Network of Excellence, and a board member of the Pattern Recognition and Machine Intelligence Association (PREMIA) Singapore.

**Day 2 (11 May 2011, Wednesday): Invited Talks on Data Mining Applications and Hands-on Session**

**Venue: Programming Lab 2, Basement, COM1, School of Computing**

9:00 am – 10:10 am	Semantic Technologies (by Dr. Kanagasabai Rajarman)
10:10 am – 10:40 am	Tea Break
10:40 am – 11:50 am	Privacy Data Mining (by Dr. Han Shuguo)
11:50 am – 1:00 pm	Statistical Tools for Data Mining (by Dr. Feng Mengling)
1:00 pm – 2:00 pm	Lunch (on your own)
2:00 pm – 3:30 pm	Data Mining Hands-on Workshop, Part I (by Mr. Jason Loh)
3:30 pm – 4:00 pm	Tea Break
4:00 pm – 5:30 pm	Data Mining Hands-on Workshop, Part II (by Mr. Jason Loh)

Talks’ Synopses and Speakers’ Biodata

1. Semantic Technologies (by Dr. Kanagasabai Rajarman)

Semantic technology, a key part of Web 3.0, is already changing the way we organize, manage and structure information and data, revolutionizing traditional IT practices and solutions. This talk will introduce semantic technologies and discuss foundations such as ontologies and semantic modeling & querying using W3C recommended standards. A technical introduction to RDF, OWL and SPARQL will be provided together with illustrative examples. State-of-the-art case studies from variety of domains such as Advertising, Marketing, and Healthcare will be presented. Also popular tools will be discussed to get the participants started on exploring real life applications of semantic technologies.

Biodata: Dr. Rajaraman Kanagasabai is currently a Principal Investigator at the Data Mining Department, Institute for Infocomm Research (I2R), Singapore, and leads the Semantic Technology Group. He has widely published in top peer-reviewed journals and conferences, and served in the

Programme Committees of many international conferences. He has also chaired or co-chaired several international events related to Semantic technologies and Analytics. He was part of the core research team behind the multiple-award winning iAgent - the first multilingual search engine, WebWatch - the key technology behind the successful startup BuzzCity ([www.buzzcity.com](http://www.buzzcity.com)), and the KnowleSuite technology that has been spunoff as Knorex ([www.knorex.com](http://www.knorex.com)). He was also the leader of the team that won the Tan Kah Kee Young Inventor's

2. Privacy Data Mining (by Dr. Han Shuguo)

Personal patient health records are one of the most sensitive types of private data. To prevent the misuse of data, some laws and regulations have been proposed to prohibit companies or groups from sharing their data, such as the U.S. healthcare laws and the 1996 administrative simplification provisions in HIPAA. On the other hand, data mining over health records is vital for medical, pharmaceutical, and environmental research. For instance, one may wish to study the effect of a certain gene on an adverse reaction to a certain drug. However, due to privacy concerns, the DNA sequences and the medical histories may be stored at different data repositories and cannot be collected together. So, how to enable the hospitals/researchers to conduct the desired data mining algorithms on those data without even "seeing" the original data becomes an interesting but challenging research topic. Privacy-preserving data mining/publishing were proposed to address the problem. This talk will present the introduction and various common techniques of privacy-preserving data mining/publishing, and discuss possible applications in the healthcare and other domains.

Biodata: Dr. Han Shuguo has been with the Data Mining Department of Institute for Infocomm Research (I2R), A\*STAR after his graduation. He received his Bachelors degree (with honors) and Ph.D. degree from School of Computer Engineering, Nanyang Technological University, Singapore, in year 2005 and year 2010 respectively. His research interests include privacy/security issues in data mining/publishing/sharing, machine learning and cryptography technologies. He has published various conference/journal papers, including ACM SIGKDD, SIAM SDM, IEEE TKDE, PAKDD, and IEEE ICDE. Currently, he is working as a key member on privacy/security issues of several industry projects collaborated with companies (e.g., NBCUniversal). He is a member of IEEE Computer Society, ACM, and SIAM.

3. Statistical Tools for Data Mining (by Dr. Feng Mengling)

The objective of data mining is to discover useful and meaningful patterns hidden in the data. Statistical tools (measurements and tests) are often necessary to assess the "usefulness" and "meaningfulness" of patterns. In the literature, the commonly used statistical tools include t-test, Chi-2 test, fisher's exact test, etc. All the statistical tools are developed and can be applied under certain assumptions and constrains. However, many researchers in the data mining community tend to apply the statistical tools without careful examination of the validness of the underlying assumption. This may often lead to false discoveries and invalid conclusions. Therefore, in this talk, we will introduce the commonly used statistical tools for data mining along with their underlying assumptions and constrains. We will also review the common mistakes while applying these tools. Moreover, alternatives and necessary corrections will also be suggested.

Biodata: Dr. Feng Mengling is currently working in the Data Mining Department of Institute for Infocomm Research (I2R), A\*STAR. He is involved in a wide spectrum of research projects across the fields of bio-imaging, bioinformatics, medical data analysis, continuous time-series analysis, data

mining for business strategies and fundamental data mining. Dr. Feng has studied his bachelor in Nanyang Technological University majoring in Electrical & Electronic Engineering. He then found his real interest in knowledge discovery and data mining during his PhD study under the supervision of Prof. Wong Limsoon and Prof. Tan Yap-Peng. Dr. Feng's current research focus is on biostatistics, data mining for business intelligence and medical time-series analysis.

#### 4. Data Mining Hands-on Workshop (by Mr. Jason Loh)

This is a three hour hands-on workshop. Participants will work through a practical session of a data mining problem, using a Data Mining Software.

The following is the outline of the workshop.

- Overview of Data mining
  - Basic concept of data mining
  - Applications in various industries (government/ banking/ telco/ commercial)
- Hands on workshop – Example: “who should the bank offer loans to?”
  - Data access
  - Data exploration/ visualization
  - Data preparation/ partitioning
  - Modeling
    - Decision trees – what? / how? / why?
    - Regression – what? / how? / why?
  - Comparing models
  - Scoring new cases

Biodata: Mr. Jason Loh has more than 9 years of experience in the area of data mining & analytics solutions from leading vendors SAS and SPSS (IBM). He is currently working with SAS AP as a Product Manager for analytical products with a focus in text analytics, supporting banks/ governments/ telcos/ and other organizations across 14 countries in the Asia pacific. He specializes in solving a wide range of business challenges by employing the use of analytics - visioning, design, execution and management of a good number of successful projects for organizations in various industries working in both consulting & user environments, Jason is effectively translating enterprise or research data into actionable insights with to solve a range of objectives from CRM, reducing churn, targeting high propensity up sell/ cross-sell opportunities and sentiments analysis, to promoting voluntary policy compliance and detecting fraud. Jason graduated from Monash University, AU with a double degree – Bach. Business/ Commerce, Bach. Computing. During his work in SPSS/ SAS – he has provided numerous workshops, training, talks, etc every year to analysts and management audience across industries.