Programme and Abstracts

8 March (THU)	
9:00 - 9:30	**Registration and Morning Tea**
9:30 - 9:40	**Opening Address** by Prof. Phee Louis
9:40 - 9:50	**Photo Session** at the first floor of NEC
9:50 - 10:20	***Open Data, Closed Government: Unpacking data.gov.sg***

	By Assoc. Prof. **Hallam Stevens**
	School of Humanities, Nanyang Technological University
	http://hallamstevens.org/
	
	Abstract: Singapore's government has signaled its intention to become a leader in data science and analytics. In 2014, for example, Singapore appointed a "chief data scientist" as part of its Infocomm Development Authority. In 2011, the government created data.gov.sg, a website for making large quantities of data available to the public in various formats. The website's stated aims include "creating value by catalysing application development" and "facilitating analysis and research." Presently, the site includes demographic data, traffic data, crime data, economic data, geographic/GIS data, health data, and a wide variety of other kinds of data.
	
	    Data.gov.sg apparently represents a commitment to openness and availability of data ("Data shall me made easily accessible"; "Data shall be released in a timely manner"; "Data shall be as raw as possible"). However, the Singapore government has also been broadly criticized for its lack of transparency and accountability. This became a significant issue in the lead up to the 2015 General Election when opposition parties pressed the government for details of financial and managerial dealings.
	
	    This talk uses a close reading of data.gov.sg to investigate the possible meanings and potential of data sharing and open data in tightly controlled society. What possibilities does data open up in an aspiring "smart city" deeply concerned with its own security? In such a context, "data sharing" and "data openness" may become intertwined with tools of political control and legitimation. Following, so far as possible, data in Singapore suggests its potential to entrench existing social, political, and economic structures.

10:20 - 10:50	***Neural Generative Model of Small Dataset for Leveraging Our Knowledge***
	By Dr. **Takashi Matsubara** and Prof. Kuniaki Uehara Department of Computational Science, Graduate School of System Informatics, Kobe University https://sites.google.com/site/takashimatsubarasprofile/profile_en Abstract: Following the great success of deep learning in computer vision, it has been applied to various practical tasks. Application to medical data potentially finds specific biomarkers of disorders, and gives an early diagnosis and an opportunity for appropriate treatments. However, such dataset is still composed of a limited number of samples compared to datasets for computer vision tasks, we have to pay special attention to it. For leveraging the flexibility of deep learning and the regularization based on our prior knowledge, we propose deep neural generative models for fMRI, stock price, and manufacturing data. The proposed models outperform conventional methods and regular deep learning.
10:50 - 11:00	**Short Tea Break**
11:00 - 11:30	***AI and Data Science for Neurological Disorders***
	By Prof. **Guan Cuntai** School of Computer Science and Engineering, Nanyang Technological University http://www.ntu.edu.sg/home/ctguan/
11:30 - 12:00	***Malicious Website Detection and Exploit-kit Identification Techniques Using Website Structure***
	By Dr. **Yasuhiro Takano**, Dr. Yoshiaki Shiraishi, Prof. Masakatsu Morii Department of Electrical and Electronic Engineering, Graduate School of Engineering, Kobe University Abstract: Cyber incidents are often caused by malicious websites using exploit-kits. The exploit-kits are ready-made attacking tools that facilitate attackers to perform the drive-by download (DBD) attack. They are typically armed with offensive, defensive, and management components. Moreover, due to their protection mechanisms, the exploit-kits obfuscate themselves to prevent from reverse-

engineering. Therefore, it has been reported that malicious websites using an exploit-kit have similarity in their website structure (WS).

This study investigates, first of all, linkmining techniques for actual website data obtained by our web-crawler, in order to verify the reported malicious website feature. According to our experiment results, a convolutional neural networks (CNN)-based algorithm achieves detection accuracy of 87%, while a support vector classification (SVC) technique obtains 74%. In this first study, however, we focus attack-sites individually and do not fully analyze a sequence of the DBD attack.

We investigate, then, to identify exploit-kits by leveraging HTTP-redirection events often observed in DBD attacks. Notice that exploit-kits can block inspection access by their defensive components. We are, hence, difficult to capture a WS-tree of exploit-kit, even if an entire HTTP traffic data of DBD attack was available. As a solution to the problem, we propose a new WS-tree construction technique by noticing that a redirection event happens in a certain duration. Our experiment results verify that a tree edit distance (TED)-based identification technique using the proposed method can achieve reasonable performance in the D3M dataset.

12:00 - 14:00	**Lunch** and **Poster Presentation** (see the appendix)
14:00 - 14:30	***Weekly and Semi-Supervised Learning for Detection of Texts in Scenes***
	By Asst. Prof. **Lu Shijian** School of Computer Science and Engineering, Nanyang Technological University http://research.ntu.edu.sg/expertise/academicprofile/Pages/StaffProfile.aspx?ST_EMAILID=SHIJIAN.LU&print=1
14:30 - 15:00	***Quantitative Approach to Aspects of L2 English Use by Learners: A Study Based on Learner Corpora***
	By Prof. **Shin'ichiro (Shin) Ishikawa** School of Languages & Communication, Kobe University http://web.cla.kobe-u.ac.jp/teacher-en2/ishikawa-shinichiro Abstract: This research project aims to quantitatively investigate aspects of L2 English use by EFL and ESL learners in Asia with a special focus on Japanese and Singaporean learners, using the ICNALE, a newly compiled learner corpus.

15:00 - 15:30	***Named Entity Recognition and Linking***
	By Assoc. Prof. **Sun Aixin**
	School of Computer Science and Engineering, Nanyang Technological University
	https://www.ntu.edu.sg/home/axsun/
	Abstract: Named entity recognition aims to identify mentions of real-world objects like persons, locations, and products, from text. With the popularity of social media platforms, we now have the access to huge volume of unstructured data in text form. Due to the informal writing and self-defined abbreviations, it is challenging to recognize mentions of objects in social media. In our proposed solutions, we aim to interpret user generated content through a dictionary in user language. The named entity recognition is then conducted by using the dictionary. The task of entity linking, also known as named entity disambiguation, is to determine the identity of a mention, by referencing a knowledge base like Wikipedia. Many studies show that collaborative linking is effective which performs linking not only based on the local context of a mention, but also the relatedness of candidate identities of other mentions in the same document. Through data analysis, we re-visit the assumption of collaborative linking and propose a much simpler version of collaborative linking, named Pair-Linking.
15:30 - 15:50	**Afternoon Tea Break**
15:50 - 16:20	***Optimization of Tensor Network Representation for Probability Distribution from the View Point of Entanglement***
	By Prof. **Tomotoshi Nishino**
	Department of Physics, Graduate School of Science, Kobe University
	http://www.phys.sci.kobe-u.ac.jp/faculty/nishino_e.html
	Abstract: Probability distribution is one of the key concepts in modern physics, in particular in the field of quantum physics and statistical mechanics, where they have many aspects in common. The tensor network formulation enables us to represent probability distribution in nature by means of the contraction among local tensors, in the manner that expectation values can be calculated with the use of a realistic resource for numerical calculation. Our aim is to develop algorithms for adjusting the tensor element each by each, so that the tensor network precisely approximate

	a group of given probability distribution. Among a variety of tensor network structures, we focus on the matrix product state (MPS), where the distribution function is represented as a partial trace of 3-leg tensors, which are aligned one-dimensionally. The MPS representation is characteristic in the point that the entanglement structure is contained in an explicit manner, the property which is convenient for a class of strongly correlated physical systems. We conjecture that a portion of non-physical probability distributions share similar real-space entanglement structure in common. In our presentation, we explain the concept of tensor network formulation from the basic notion to modern applications.
16:20 - 16:50	***Introduction to Data Stream Algorithms***
	By Asst. Prof. **Li Yi** School of Physical and Mathematical Sciences, Nanyang Technological University http://www.ntu.edu.sg/home/yili/ Abstract: This talk aims to give a brief introduction to data stream algorithms studied in the field of theoretical computer science, starting with the definition of the models. Two typical problems, the sparse recovery problem and the subspace embedding problem will be discussed in greater detail as examples.
16:50 - 17:20	***Making Big Data of Motor Control Ability with Low-cost Devices Toward Developing In-home Diagnosis and Rehabilitation System***
	By Dr. **Tomoya Tamei** Center for Mathematical and Data Sciences, Kobe University Abstract: Maintaining and improving quality of life (QOL) of elderly people and patients is social needs of aging societies. Accumulating quantitative data on motor control ability in everyday environments will be useful for constructing systems for evidence-based diagnosis and rehabilitation. We developed a low-cost and compact system monitoring motion and electromyography (EMG) signals using Microsoft Kinect and Myo Gesture Control Armband.
17:20 - 17:30	**Closing Remarks** by Prof. Chee Yeow Meng
18:00 - 20:00	**Dinner**

Organized by DSAIR and SPMS at NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE and CMDS at KOBE UNIVERSITY

Appendix: Details of Poster Presentations

1. ***Towards the Optimal Structure of Deep Neural Networks from the Point of View of Entanglement***
    By Mr. **Yoshinori Sasagawa** and Prof. Tomotoshi Nishino
    Department of Physics, Graduate School of Science, Kobe University

Abstract: Machine learning techniques have state-of-the-art performance in the various fields. The core algorithm is deep neural networks. Although there are a lot of studies to clarify why the techniques of deep neural networks are so successful, the design principle of network structure is still intriguing. We introduce the concept of entanglement into the restricted Boltzmann machine, which is well known as a building block of deep neural networks. We show that the entanglement entropy can express the progress of the learning, and quantify the contribution of the hidden units to the performance of the model. It is implied that there is the possibility of exploring the optimal structure of deep neural networks based on the concept of entanglement instead of the correlation function.

2. ***Modelling Interdependence between East Asian Stock Markets and the Prices of Oil and Gold***
    By Dr. **Cai Xiao Jing** and Prof. Shigeyuki Hamori
    Graduate School of Economics, Kobe University

Abstract: Many previous studies analyzed the dependence between two assets by using the copula functions because the dependence show many natures such as the nonlinear dynamic, asymmetric, skewed and kurtosis distribution. However, the limitation of these studies is that they are restricted to one or, at most, two time scales – the short and long term. In fact, international investors should be heterogeneous with respect to their different investment horizons. Therefore, to resolve this issue, in this project, we combine the wavelet transform analysis and the conditional copula functions to examine the dependence between two assets across different horizons. Specifically, we first utilize the wavelet transform to decompose the standardized shocks obtained from marginal models into time series at different frequency (including short-term, midterm, and long-term), and then employ the conditional static and dynamic copula functions to capture their constant and dynamic interdependence and tail dependence across different time horizons respectively.

In this project, we investigate the dynamic dependence between East Asian stock markets and the prices of crude oil and gold. Many literatures have examined the oil-stock relationship for main developed countries while few studies focus on how they work for the developing region such as East Asia. In fact, this is an interesting and important subject. Over the previous decades, East Asia's miraculous economic growth and dynamism has become a popular topic for academic research. Furthermore, three of the world's top ten oil-importing nations – China (China represents Chinese mainland in our project), Japan, and South Korea are in East Asia. East Asia shows an increasing demand for oil. Additionally, the majority of East Asian oil imports

are from the volatile Middle East, and there has been no regional mechanism in East Asia to stockpile emergency petroleum supplies, which makes East Asia highly susceptible to oil shocks such as the 2003 Iraq invasion or the 2006 OPEC cut agreement. Therefore, crude oil price fluctuations had a greater impact on East Asian economies than the developed countries. Moreover, the risk reduction benefit from diversification has been a major subject in the financial literature for decades. Most East Asian stock markets that are emerging economy markets are vulnerable to the global environment and the volatility of developed stock markets. For example, East Asia suffered huge losses in the periods of 1997 Asian and 2008 global turmoil. Therefore, investors should look to a defensive diversification strategic asset such as crude oil or gold to minimize the risk and to improve the risk-return tradeoff of asset portfolios in times of financial crises.

Therefore, in the next step, we want to collaborate on these issues that whether the crude oil or gold are useful in diversifying the risk of East Asian stock markets across different investment horizons and how to allocate the portfolio to minimize the risk and benefit most. For example, we consider 3 portfolios, the stock only portfolio (portfolio 1), the oil-stock portfolio (portfolio 2) and the gold-stock portfolio (portfolio 3). On the basis of the above empirical results of wavelet based copula model, we can calculate and compare the values of Portfolio Variance (PV), Value at Risk (VaR), Expected Shortfall (ES), Sharpe Ratio (SR) among 3 portfolios across different time scales. Thus, we can find which portfolio is useful in reducing the risk of stock markets and we can also find out at which investment horizons can obtain more benefits. These results have implications for heterogeneous investors and market participants. We need people who have some understanding of portfolio management and are skilled in using Matlab.

3. *Full-time Wearable Sensing and Psychological Effects on Seeing Sensor Data*
   By Dr. **Tsutomu Terada**
   Department of Electrical and Electronic Engineering, Graduate School of Engineering, Kobe University
   http://ubi.eedept.kobe-u.ac.jp/profile_e.html

   Abstract: We are developing new sensing systems equipped with glasses, such as Transepidermal Water Loss, Facial expression, and Swallowing. Full-time data can create new value for human QoL and healthcare. We can provide devices or data acquired using the devices. Also, we are working on psychological effects on presenting information using wearable devices such as head mounted displays, vibrators, and LEDs. We are measuring various types of effects on biological state by seeing sensor data.

4. *Finding Key Movements to distinguish Expert and Novice in Dance and/or Runway Model Walking*
   By Prof. **Hajime Murao**
   Graduate School of Intercultural Studies, Kobe University

http://i.cla.kobe-u.ac.jp/murao/eng/profile/index.html

Abstract: What is an essential difference between the expert behavior and the novice one at where the small differences of movements make great differences in results, such as dancing, playing darts, preparing tea at Japanese tea ceremony or runway walking? We would like to figure it out from the movement of the human joints at an activity. Which will give us a hint how to improve our behavior. In our study, Structure Sensor, a Kinect-like 3D sensor, is used to track and record 3-dimensional positions of the human joints during an action. The device is portable and does not require any sensors nor markers attached to observed persons. We then investigate the effectiveness both of a statistical approach like PCA and a deep learning approach. Results of preliminary experiments established for dancing and runway model walking will be reported.

5. ***Collecting Cybersecurity-related Contents in Dark Web***
   By Prof. **Seiichi Ozawa**
   Center for Mathematical and Data Sciences, Kobe University
   http://www2.kobe-u.ac.jp/~ozawasei/index.html

Abstract: It is well known that products for cyber-attacks such as exploits and malware codes are illegally traded on hidden web services called "Dark Web" that are not indexed by conventional search engines. In general, it is not easy to capture the whole picture of trade activities on Dark Web because special browsers and tools are needed to visit such dark markets and forums. And they usually require us to make a registration and/or to pass a qualification test. However, to understand the trends of cyber-attacks, there is no doubt that Dark Web is one of useful information sources.

   In this project, we want to efficiently find and store information on cyberattacks from Dark Web such as sales trends of illegal products for cyber-attacks at marketplace and useful information on cyberattack products at forums. Recently, we have developed an AI Web Crawler that consists of a Tor crawler and a topic analyzer to categorize contents of cyberattacks. For collected web contents, we want to select and keep only contents on cyberattacks. For this purpose, we recently use Doc2Vec and Sparse Composite Document Vectors (SCDV) as document vectors to select related documents. The classification is going well so far, but there still remains room to be improved. We hope to have good collaboration with NTU researchers to develop more efficient AI web crawler and to establish high-performance text classification/mining methods.

6. **NTU Poster 1**
7. **NTU Poster 2**
8. **NTU Poster 3**
9. **NTU Poster 4**

Organized by DSAIR and SPMS at   and CMDS at