Melbourne, Australia

7 - 9 December 2016

EMERGING BIG DATA TECHNOLOGIES SUMMIT

Hosted by **the International Innovative Research Network**

BOOK OF ABSTRACTS

Table of Contents

PLENARY SESSIONS

Day 1 (7th Dec)

Session 1

1. Challenges of Big Data in Scientific Discovery
   Prof. Benjamin W. Wah
   Chinese University of Hong Kong (China)
2. The Data Game: Potential, players and privacy
   Prof. Ian Oppermann
   New South Wales Data Analytics Centre (Australia)

Session 2

3. Data Engineering Redefined in the Era of Big Data
   Prof. Saman K. Halgamuge
   The Australian National University (Australia)

Day 2 (8th Dec)

Session 3

4. Tensor Decompositions and Tensor Networks for Big Data Processing
   A Review, Challenges and New Perspectives
   Prof. Andrzej Cichocki
   RIKEN (Japan)
5. Sparse machine learning methods to extract meaning from large chemical, materials, and genomic data sets
   Prof. Dave Winkler
   CSIRO (Australia)

Session 4

6. Time-Critical Social Mobilazation
   Dr. Manuel Cebrian
   DATA61 (Australia)

Day 3 (9th Dec)

Session 5

7. The Science of Big Data and the Data of Big Science
   Prof. Timos Sellis
   Swinburne University of Technology (Australia)
8. Driving data science at scale
   Dr. Amy Shi-Nash
   Commonwealth Bank (Australia)

Session 6

9. Pattern and Knowledge Discovery for Biomedical Applications and Genomics Context
   Prof. Yi-Ping Phoebe Chen
   La Trobe University (Australia)

Session 7

10. Big Data Analytics for Smart Grids: Challenges and Opportunities
    Dist. Prof. Xinghuo Yu
    RMIT University (Australia)

Challenges of Big Data in Scientific Discovery

Prof. Benjamin W. Wah[1]

[1]*Chinese University of Hong Kong, China*

Big Data is emerging as one of the hottest multi-disciplinary research fields in recent years. Big data innovations are transforming science, engineering, medicine, healthcare, finance, business, and ultimately society itself. In this presentation, we examine the key properties of big data (volume, velocity, variety, veracity, and value) and their relation to some applications in science and engineering. To truly handle big data, new paradigm shifts will be necessary. Successful applications in big data will require in situ methods to automatically extracting new knowledge from big data, without requiring the data to be centrally collected and maintained. Traditional theory on algorithmic complexity may no longer hold, since the scale of the data may be too large to be stored or accessed. To address the potential of big data in scientific discovery, challenges on data complexity, computational complexity, and system complexity will need to be solved. We propose a new approach based on identifying kernel data to harness the complexity of big data applications. Kernel data is a compact and manageable representation of the original data, with similar structure, data properties, or meta-properties. We illustrate these challenges and approaches by drawing on examples in various applications in science and engineering.

The Data Game: Potential, players and privacy

Ian Oppermann[1]

[1]CEO and Chief Data Scientist at NSW Data Analytics Centre, Sydney, Australia

There are many powerful opportunities which arise from data sharing, linking, modelling and prediction. This is true for advertisers, search engines, coffee shops and many parts of the economy. It is also true for government. Whilst the technical and legal ability to create value exists, there are many issues still to be addressed before we can realise this full potential. This presentation will address some of the potential, provide case examples from government and offer some unanticipated challenges which still need to be addressed.

Data Engineering Redefined in the Era of Big Data

Prof Saman K. Halgamuge[1]

[1]*The Australian National University, Australia*

We are listening to the universe hoping to hear a "hello" from a species living outside Earth, but we do not have a clue about 98% or more of the microbial organisms living on Earth including inside our bodies. We sequenced the human genome, but we still do not have an accurate treatment for cancer. We are unable to make sense of the evidence in front of our eyes because this evidence is hidden in complex, imperfect and large quantities of data.

In the age of information and connectivity, we receive data BIG in terms of volume, heterogeneity and complexity. This data of different types and sizes, of different levels of perfection and of different levels of integrity may contain important information that can be extracted by powerful data analysis methods.

Data Engineering promotes an engineering approach to analyze big "imperfect" data by creating and using new algorithms such as Unsupervised Deep Learning, appropriate electronic hardware platforms and mechanical/mechatronic/chemical-based approaches to interrogate and acquire missing data or information. I will present some of the work conducted by members of my group and collaborators. Their research varies from algorithm development to applications in health, energy, agriculture and business.

ACKNOWLEDGMENT

This work is partially supported by 15 University of Melbourne PhD Scholarships and Australian Research Council Grants: "Near Unsupervised computational methods for exploring omic data (DP150103512), "Discovering Patterns using Near Unsupervised Learning to Support the Quick Detection of New Animal Disease Outbreaks Caused by Viruses" (LP140100670) and YourGene Australia and AIC Tactical research project "Near Off-Grid Solutions using Renewable Energy Technologies and Demand Side Prediction". The current PhD students: D. Mendis, Y. Deerasooriya, H. Weeratunga, P. Hamead, C. Jayawardena, D. Herath, D. Senanayake, A. Khalaj, W. Wei, Y. Sun and K. Abdulla and previous students C. Wijetunga, K. Amarasinghe, Z. Li, U. Premaratne, S. Jayasekara, D. Jayasundara, D. Alahakoon and K. Chan and collaborators: S. L. Tang, S. Petrou, U. Kayande, K. Steer, A. Wirth, B. Chang, M. Premaratne, A. Hsu, I. Saeed, U. Roessner, M. Kirley, K. Verspoor, J. Browne, G. Narsilio, D. Ackland and A. Bacic are acknowledged

Tensor Decompositions and Tensor Networks for Big Data Processing: A Review, Challenges and New Perspectives

Prof. Andrzej Cichocki[1]

[1]*RIKEN (Japan)*

Sparse machine learning methods to extract meaning from large chemical, materials, and genomic data sets

Dave Winkler[1-4,*], Tu Le[1] , Frank Burden[1], Vidana Epa[1]

[1]*CSIRO Manufacturing, Clayton 3168, Australia*
[2]*Monash Institute of Pharmaceutical Sciences, Parkville 3052, Australia*
[3]*Latrobe Institute for Molecular Science, Bundoora 3084, Australia*
[4]*School of Chemical and Physical Sciences, Flinders University, Bedford Park 5042, Australia*

[*]E-mail: dave.winkler@csiro.au

As well as the familiar problem of dealing with very large data sets now emerging from a diverse range of new technologies, scientists are starting to address the challenges posed by the vastness of molecular spaces – potential drug-like molecules that could be synthesized, new materials that are accessible to current synthesis and fabrication methods, and biological and biomaterials that could be constructed from known synthetic monomers or natural and non-natural amino acids.

Powerful combinations of sparse methods of identifying important features in data, new machine learning algorithms that create quantitative links between structure/physicochemical properties and biological or macroscopic properties, evolutionary algorithms that efficiently search vast spaces, and more capable and cheaper robotics are providing a practical way to address the molecular space problem for the first time.

This paper will describe these new in silico methods, illustrate their power using case studies, and discuss their potential for solving large data set and vast molecular space problems to generate novel materials and therapies.

REFERENCES

1. Le, T.C; Winkler, D.A. Discovery and optimization of materials using evolutionary approaches. Chem Rev. 2016;116 (10), 6107.
2. Oksel, C.; Winkler, D.A.; Ma, C.Y.; Wilkins, T.; Wang, X.Z. Accurate and interpretable nanoSAR models from genetic programming-based decision tree construction approaches, Nanotoxicol, 2016; 10:1001.
3. Thornton, AW; Winkler, DA; Liu, M; Haranczyk, M; Kennedy, DF. Towards computational design of zeolite catalysts for CO_2 reduction. RSC Adv. 2015; 5, 44361.
4. Le, T.C.; Winkler, D.A. A bright future for evolutionary methods in drug design, ChemMedChem 2015; 10:1296.
5. Winkler, D.A. Adrien Albert Award: How to mine chemistry space for new drugs and biomedical therapies, Aust. J. Chem. 2015: 68: 1174.
6. Autefage, H.; Gentleman, E.; Winkler, D.A.: Burden, F.R.; Stevens, M. Novel sparse feature selection methods identify unexpected global cellular response to strontium-containing materials. Proc. Natl. Acad. Sci. USA 2015; 112:4280.
7. Celiz, A. D.; Smith, J. G. W.; Langer, R.; Anderson, D. G.; Barrett, D. A.; Young, L. E.; Winkler, D.A.; Davies, M. C.; Denning, C.; Alexander M. R. The Search for Materials for the Stem Cell Factories of the Future. Nature Mater. 2014; 13:570.
8. Le, T.C.; Epa, V.C.; Burden, F.R.; Winkler, D.A. Quantitative Structure–Property Relationship Modeling of Diverse Materials Properties. Chem. Rev. 2012; 112:2889.

Time-Critical Social Mobilization

Dr. Manuel Cebrian[1]

[1]*DATA61*

This talk explores the physical and behavioral limits of crowds-assembly for problem solving, by following a number of real-world experiments where we utilized social media to mobilize the masses in tasks of unprecedented complexity. From finding red weather balloons to locating people in distant cities, to reconstructing shredded documents, the power of crowdsourcing is real, but so are exploitation, sabotage, and hidden biases that undermine the power of crowds.

The Science of Big Data and the Data of Big Science

Prof. Timos Sellis[1*]

[1]*Centre for Big Data and Data Analytics, Swinburne University of Technology, Hawthorn, Victoria 3122, Australia.*

*E-mail: tsellis@swin.edu.au

Big Data has become the hottest research topic in data (and in general information) management nowadays. At the same time, science is heavily guided by discoveries in large data streams collected through experimental settings. In this talk we discuss the common, major issues rising in the two these modern research areas: managing Big Data and driving big Science, through modern Data Science methods.

Big Data research aims to advance the core scientific and technological means of managing, analyzing, visualizing, and extracting useful information from large, diverse, distributed and heterogeneous data sets in order to advance the development of new data analytic tools and algorithms, and facilitate scalable, accessible, and sustainable data infrastructures. Big Science on the other hand has gone through a new paradigm shift; having passed through the empirical/observational, the theoretical/experimental, and the computational paradigms, science is now conducted predominantly following the data exploration paradigm. Incredible amounts of data of great complexity is generated, which is then analyzed in an automatic or semi-automatic fashion; this results in identification of common patterns and trends or rare phenomena, which often constitute new scientific discoveries or lead to those. We focus on key relevant data-management challenges that arise in the context of scientific efforts and require significant advances in current technology.

Driving data science at scale

Dr. Amy Shi-Nash[1]

[1]*Commonwealth Bank*

Enhance the Financial wellbeing of people, business and community is at the heart of our service as a bank. Data driven decision-making is critical in achieving the goal. In this talk, Amy will outline the framework of developing data science capability at scale, the key challenges as well as the mindset change required

Pattern and Knowledge Discovery for Biomedical Applications and Genomics Context

Prof. Yi-Ping Phoebe Chen[1]

[1] *Department of Computer Science and Information Technology, La Trobe University*

Solving modern biomedical problems, especially those involving genome data, requires advanced computational and analytical methods. The huge quantities of data and escalating demands of modern biomedical research increasingly require the sophistication and power of computational techniques for their pattern discovery. In this talk, I will demonstrate recent methodologies and data structures for gathering high-quality approximations and modelling of genomic information, and will use these innovations as the basis for developing methods to cluster and visualize biomedical data in pattern discovery.

Big Data Analytics for Smart Grids: Challenges and Opportunities

Dist. Prof. Xinghuo Yu[1]

[1]RMIT University, Melbourne, Australia

*E-mail: x.yu@rmit.edu.au

Smart Grids are electric networks that employ innovative and intelligent monitoring, control, communication, and self-healing technologies to deliver better connections and operations for generators and distributors, flexible choices for prosumers, and reliability and security of electricity supply. Smart Grids are complex cyber-physical network systems in which information retrieval, processing, intelligence are critical. The exponentially increasing quantities of real-time measurements from the grids present technical challenges to the efficiency and effectiveness of grid operations in uncertain environments. Big data analytics will play a significant role in addressing the challenges to deal with complex, dynamic, hybrid big data. In this talk, we will first give a brief overview of Smart Grids and their recent developments, focusing on broad challenging issues from a big data perspective. The interplay of Big Data Analytics and Smart Grids will be discussed and mutually benefits explored.

CONCURRENT SESSIONS

DAY 1 (7th Dec)

Session 1: Real time analytics

1. RAPID: Real-time Analytics Platform for Interactive Data-mining
   A/Prof. Shanika Karunasekera
   The University of Melbourne (Australia)
2. Abnormality Detection in Co-evolving Data Streams
   Prof. Jing He
   Victoria University (Australia)
3. Detection of money laundering groups using supervised learning in networks
   A/Prof. Xiuzhen Zhang
   RMIT University (Australia)
4. Computing with the Cloud, the Crowd, Personal Data, IoT, and Drones
   A/Prof Seng W. Loke,
   La Trobe University (Australia)
5. Beyond Points and Paths: Counting Private Bodies
   Maryam Fanaeepour
   The University of Melbourne (Australia)

Session 2: Digital cities

6. Internet of Things: From Internet Scale Sensing to Smart Services
   Dimitrios Georgakopoulos,
   Swinburne University (Australia)
7. Evaluating big geodata analytics: From emergency response to environmental monitoring
   Prof. Matt Duckham
   RMIT University (Australia)
8. Digital Cities - do Startups make a difference?
   Mr. Trevor Townsend
   Startupbootcamp (Australia)
9. Bird Acoustic Data analysis
   Dr. Saurabh Garg
   University of Tasmania (Australia)
10. Registration of KALPANA-1 VHRR and INSAT-3A CCD imagery for Meteorological Studies
    Jignesh Bhatt,
    Indian Institute of Information Technology Vadodara (India)

Session 3: Scalable Models and Algorithms for Data Analytics

11. Fast Optimisation Algorithms for Distributed Data Analytics in Apache Spark
    Prof. Hans De Sterck
    Monash University (Autralia)
12. Methods for analysing partially observed functional data
    Prof. Aurore Delaigle
    The University of Melbourne (Australia)
13. Data Science and Superannuation in Australia
    Dr. Meng Wang,
    Empirics Data Solutions (Australia)

Session 4: Cloud and Internet Computing

14. Ensuring big data reliability with minimum replication for cost effectiveness in the cloud
    Prof. Yun Yang
    Swinburne University of Technology (Australia)
15. Making Big Data work across disciplines - The AURIN Solution
    Dr. Jack Barton

AURIN (Australia)

16. Mining complex data from highly streaming environment
    Zhinoos Razavi
    RMIT University (Australia)

Day 2 (8th Dec)

Session 5: Digital healthcare

17. Big Data in Healthcare Operations – a Tale of 2 Sites, 1 City
    A/Prof. Christopher Bain
    Mercy Hospitals Victoria (Australia)
18. Data usability with personalised autonomous mining in health care systems
    Prof. Hua Wang
    Victoria University (Australia)
19. Data Driven Decision Making: From Remote Health Monitoring to Venture Prediction
    Dr Sanjeev Naguleswaran
    QSpectral Systems (Australia)
20. Cognitive Computing for skin analytic and early detection of Melanoma
    Dr Mani Abedini
    IBM Research (Australia)
21. Handling Healthcare Data using Polyglotpersistence
    Karamjit Kaur,
    Thapar University (India)

Session 6: Predictive business analytics

22. Lost in Translation: On the Disconnect between Business-speak and Data Science
    A/Prof. Michael Brand,
    Monash University (Australia)
23. Big Data: an almighty concept in Chinese research and capital market
    Dr. Alex (YiFei) Dong,
    CUCPAY (China)
24. From Digital to Data-Centric
    Dr. Tim Cahill
    The Conversation Media Group (Australia)
25. A business analytics approach to injury risk management for Australian football
    A/Prof Kok-Leong Ong,
    La Trobe University (Australia)
26. Data driven innovation and innovation driven data
    Dr. Alexe Bojovschi
    AIB (Australia)

Session 7: Digital Business

27. Churn Analytics and the impact to business
    Mr. Gavin Whyte,
    KPMG (Australia)
28. Big Data - A Key Driver in the Digital Revolution
    Mr. Stuart Growse
    GCS AGILE (Australia)
29. Unlocking the commercial value of Big Data in a digital age
    Lee Anderson
    Delloite (Australia)
30. Data aggregation and security
    Dr. Andy Song

RMIT University (Australia)

Session 8: Digital Health

31. Cognitive Computing Transforming Healthcare
    David Yip,
    IBM Research (Autralia)
32. Scalable Data Management and Analysis for Bioinformatics
    A/Prof. Uwe Roehm,
    University of Sydney (Australia)
33. Genomics in the battle against antibiotic resistance
    Dr. Kelly Wyres
    University of Melbourne (Australia)
34. Building Ensemble of Models by Exploiting the Richness of Feature Variables in High-Dimensional Data and Application in Protein Homology
    Dr. Jabed H. Tomal
    University of Toronto (Canada)

Day 3 (9th Dec)

Session 9: Systems, Digital Business

35. The Next-Gen Information Management System – The Zetaris Way
    Dr John Brudenell,
    ZETARIS (Australia)
36. Caches all the way down: Infrastructure for Data Science
    Prof. David Abramson
    The University of Queensland (Australia)
37. Multiview Learning
    Dacheng Tao
    University of Technology Sydney (Australia)
38. AI for Healthcare
    Dr. Truyen Tran
    Deakin University (Australia)

Session 10: Internet of Things (IoT)

39. Enabling Intelligent Transport Systems (ITS) with cognitive analytics
    A/Prof. Quoc Bao Vo,
    Swinburne University (Australia)
40. Privacy and security in the age of big data
    A/Prof. James Thom
    RMIT University (Australia)
41. Performance Analytics Using Tracking Data and Computer Vision: A New Statistical Era for Elite Tennis
    Dr Stephanie Kovalchik,
    Game Insight Group at Tennis Australia & Institute of Sport, Exercise & Active Living (Australia)
42. Big Data Analytics for Data Streams from Sensors
    A/Prof. Tao Gu
    RMIT University (Australia)

RAPID: Real-time Analytics Platform for Interactive Data-mining

Shanika Karunasekera[1] and Aaron Harwood[1]

[1]*Department of Computing and Information Systems, University of Melbourne, Australia*

*E-mail: karus, aharwood@unimelb.edu.au

Interactive mining of streaming data is an emerging research topic with challenges in both user control and scalability. Conventional streaming data platforms do not address user interaction as a core aspect of the system and therefore user interactivity is difficult to achieve on a fine grain processing level. We present RAPID - Real-time Analytics Platform for Interactive Data-mining - a real-time analytics platform that gives users the ability perform interactive data mining on Twitter data streams using high-level mining queries.

RAPID enables topic-specific information discovery through dynamically changing tracking based on keywords, tweets and users relevant to the topic through automatic and semi-automatic query expansion techniques [1, 2]. The system supports built-in stream mining algorithms for query expansion, community detection and discussion tracking. RAPID addresses the following open research questions: *(i)* what are the most effective approaches for interactive knowledge discovery and query expansion? *(ii)* how to support high level streaming data mining queries, beyond SQL-like queries, such as clustering, community detection, and frequent item set mining? and *(iii)* what are the most efficient scalable architectures that support data acquisition, stream processing, storage and visualization?

The RAPID architecture demonstrates the use of state-of-the-art software frameworks, Apache Storm, for stream processing, Apache Kafka, for publish/subscribe messaging, and MongoDB, for high performance document storage, with a unique, event-based protocol for multi-tenant, user interaction via a graphic user interface. The salient aspects of the system for discussion are: *(i)* the use of publish/subscribe for both synchronous and asynchronous interactions between the client and system back-end, *(ii)* the distributed caching architecture that maintains consistent cache state across all of the components of the system, *(iii)* the dynamic allocation of multiple layers of data processing, and *(iv)* query expansion and autonomous tracking.

REFERENCES

1. M. Efron. Information search and retrieval in microblogs. J. Am. Soc. Inf. Sci. Technol. 62(6):996–1008, June 2011.

2. S. Karunasekera, A. Harwood, et al. Topic-specific post identification in microblog streams. In Big Data, 2014 IEEE Int. Conf. on, pages 7–13, Oct 2014.

Abnormality Detection in Co-evolving Data Streams

Prof. Jing He[1]

[1]*Victoria University, Australia*

Detecting/predicting anomalies from multiple correlated data streams is valuable to those applications where a credible real-time event prediction system will minimize economic losses (e.g. stock market crash) and save lives (e.g. medical surveillance in the operating theatre). This talk will introduce an effective and efficient method for mining the anomalies of correlated multiple and co-evolving data streams in online and real-time manners. It includes the detection/prediction of anomalies by analyzing differences, changes, and trends in correlated multiple data streams. The predicted anomalies often indicate the critical and actionable information in several application domains.

Detection of money laundering groups using supervised learning in networks*

Xiuzhen (Jenny) Zhang[1]

[1]*RMIT University, Australia*

Money laundering is a major global problem, enabling criminal organisations to hide their ill-gotten gains and to finance further operations. Prevention of money laundering is seen as a high priority by many governments, however detection of money laundering without prior knowledge of predicate crimes remains a significant challenge. Previous detection systems have tended to focus on individuals, considering transaction histories and applying anomaly deteciton to identify suspicious behaviour. However, money laundering involves groups of collaborating individuals, and evidence of money laundering may only be apparent when the collective behaviour of these groups is considered. In this paper we describe a detection system that is capable of analysing group behaviour, using a combination of network analysis and supervised learning. This system is designed for real-world application and operates on networks consisting of millions of interacting parties. Evaluation of the system using real-world data indicates that suspicious activity is successfully detected. Importantly, the system exhibits a low rate of false positives, and is therefore suitable for use in a live intelligence environment.

* Joint work with David Savage, Qingmai Wang, Pauline Chou and Xinghuo Yu

Computing with the Cloud, the Crowd, Personal Data, IoT, and Drones

Dr. Seng W. Loke[1]

[1] *Department of Computer Science and Information Technology, School of Engineering and Mathematical Sciences, College of Science, Health and Engineering, La Trobe University, Australia*

There have been considerable developments in
Cloud Computing, Crowd Computing, Mobile Data Analytics, Internet-of-Things, and Drones, in terms of frameworks, algorithms, scenarios, and new forms of services and applications.
This talk will propose and discuss five ideas (or five directions for thought) at the intersection of these developments:
(a) systems using crowd-power supported by the cloud (so-called crowd+cloud machines),
(b) personal and group data analytics, (c) extreme cooperation in IoT,
(d) drone services, and (e) scalable context-awareness.

Beyond Points and Paths: Counting Private Bodies

Maryam Fanaeepour[1]

[1]*The University of Melbourne, Australia*

Mining of spatial data is an enabling technology for mobile services, Internet-connected cars, and the Internet of Things. But the very distinctiveness of spatial data that drives utility, comes at the cost of user privacy. In this work, we continue the tradition of privacy-preserving spatial analytics, focusing not on point or path data, but on planar spatial regions. Such data represents the area of a user's most frequent visitation—such as "around home and nearby shops". Specifically, we consider the differentially-private release of data structures that support range queries for counting users' spatial regions. Counting planar regions leads to unique challenges not faced in existing work. A user's spatial region that straddles multiple data structure cells can lead to duplicate counting at query time. We provably avoid this pitfall by leveraging the Euler characteristic. To address the increased sensitivity of range queries to spatial region data, we calibrate privacy-preserving noise using bounded user region size and a constrained inference that uses robust least absolute deviations. Our novel constrained inference reduces noise and introduces covertness by (privately) imposing consistency. We provide a full end-to-end theoretical analysis of both differential privacy and high-probability utility for our approach using concentration bounds. A comprehensive experimental study on several real-world datasets establishes practical validity.

Internet of Things: From Internet Scale Sensing to Smart Services

Prof. Dimitrios Georgakopoulos[1]

[1]Swinburne University of Technology, Australia

The Internet of Things (IoT) is the latest Internet evolution that incorporates billions of Internet-connected devices that range from cameras, sensors, RFIDs, smart phones, and wearables, to smart meters, vehicles, medication pills, signs and industrial machines. Such IoT things are often owned by different organizations and people who are deploying and using them for their own purposes. Federations of such IoT devices (often referred to as IoT things) can also deliver timely and accurate information that is needed to solve internet-scale problems that have been too difficult to tackle before.

To realize its enormous potential, IoT must provide IoT solutions for discovering needed IoT devices, collecting and integrating their data, and distilling the high value information each application needs. Such IoT solutions must be capable of filtering, aggregating, correlating, and contextualising IoT information in real-time, on the move, in the cloud, and securely and must be capable of introducing data-driven changes to the physical world. In this talk we present an overview of IoT solutions we have developed (which we refer to them collectively as IoT platform) to address these technical challenges and help springboard IoT to its potential. We mainly focus on open source components and standards we have developed jointly with other prominent international collaborators. We also describe a variety of IoT applications that have utilized the proposed IoT platform to provide smart IoT services in the areas of smart farming, grids, manufacturing, and mining. Finally, we discuss future research and a vision of the next generation IoT infrastructure.

Evaluating big geodata analytics: From emergency response to environmental monitoring

Matt Duckham[1]

[1]*Mathematical and Geospatial Sciences, School of Science ,RMIT University, Victoria 3000, Australia.*

*E-mail: matt.duckham@rmit.edu.au

Geospatial data has always been "big." But today's big data sources present new challenges for spatial computing and spatial data mining, including the difficultly of evaluating the outputs of analytics based on big geospatial data. Using three case studies---in geospatial web search, emergency response, and environmental monitoring---this invited lecture explores some of the problems and solutions for evaluating spatial data mining and analytics procedures both for crowdsourced and automated big geodata sources.

Digital Cities - Do Startups make a difference?

Mr. Trevor Townsend[1]

[1]*Startupbootcamp, Australia*

Examining a global perspective on Digital Cities and how startups are impacting SmartCity programs and making a difference. Looking at case studies from Europe and Asia and discussing how we can encourage startup activity in our Digital City programs in Australia.

Bird Acoustic Data Analysis

Dr. Saurabh Garg[1]

[1]*University of Tasmania, Australia*

Standard approach for environmental health monitoring and assessment requires on-the-ground surveys which are costly, labour-intensive, and cannot cover all areas of interest with equal frequency. Consequently, there has been growing interest in and use of acoustic sensors to monitor biodiversity among species. Even though acoustic sensing has the potential to increase the spatial and temporal scales of biodiversity monitoring for environmental scientists, yet the large volume of data collected presents its own challenges: acoustic data can be complex to analyse and detect different species due to environmental factors such as wind and rain. This talk will discuss the research challenges involved in fully automatic environment monitoring and current research work done at the University of Tasmania.

Registration of KALPANA-1 VHRR and INSAT-3A CCD imagery for Meteorological Studies

Jignesh Bhatt[1]*, and Narayan Padmanabhan[2]

[1]Indian Institute of Information Technology Vadodara (IIITV), Gandhinagar Campus, Building #9, Government Engineering College, Sector 28, Gandhinagar 382028, India.
[2]Institute of Engineering & Technology, Ahmedabad University, Ahmedabad Education Society FP4, Navrangpura, Ahmedabad 380009, India.

*E-mail: jignesh.bhatt@iiitvadodara.ac.in

The launch of KALPANA-1 satellite in the year 2002, and INSAT-3A satellite in 2003 helped India improve their meteorological forecasts. Registration is one of the fundamental operations to generate all geophysical data products from remotely sensed data. Automatic image registration is a challenging task due to the presence of radiometric and geometric distortions during the data acquisition process. Besides the presence of clouds makes the problem more complicated. In this work, we present our two contributions: (1) a fast adaptive algorithm for automatic multiband and multitemporal image registration for both the payloads, and (2) generation of reference boundaries for INSAT-3A CCD imagery. The complete implementation consists of the following steps: 1) match-points: automatic identification of the ground control points in the sensed images by extracting statistical and geometrical features, 2) warping: estimation of the model parameters for bivariate polynomial least-squares surface fitting based on the identified match-points, 3) wild-point removal: discarding the outlier(s) based on mean and standard deviation error calculated at each warped point, and 4) resampling: interpolation of the warped image to the reference coordinates. Importantly, we have generated reference boundaries for India in the CCD imagery from INSAT-3A data using our proposed warping algorithm. The software is coded in C language on Linux platform at Data Product Software Laboratory, Space Applications Centre (SAC), Indian Space Research Organization (ISRO), India. The proposed algorithm has been successfully tested and validated using the KALPANA-1 and INSAT-3A datasets acquired over various seasons and/or years.

Fast Optimisation Algorithms for Distributed Data Analytics in Apache Spark

Hans De Sterck[1]

[1]*Monash University, Australia*

*E-mail: hans.desterck@monash.edu

Many data sets are now so large that distributed computing has become a crucial tool to analyse them. In the distributed setting, fast parallel optimisation algorithms are required to train machine learning and recommendation models. This talk will discuss some recent advances in parallel optimisation algorithms for data analytics, and their efficient implementation in the machine learning library of the Apache Spark distributed data processing environment.

Methods for analysing partially observed functional data

Prof. Aurore Delaigle[1]

[1]*The University of Melbourne, Australia*

We consider analysis of functional data which are only partially observed. Often in such cases, the observed fragments of curves are supported on quite di fferent intervals, in which case standard methods of analysis cannot be used. We propose new approaches to analysing fragments of curves observed on diff erent intervals. The techniques we suggest involve discretising the observed fragments, and then extending them outside the interval where they were observed. Using the same approach, we can construct estimators of the mean and covariance functions, and, for example, deal with functional linear regression.

Data Science and Superannuation in Australia

Meng Wang[1]

[1]*Empirics Data, Melbourne, Victoria, Australia*

*E-mail: mwang@empirics.com.au

Australia's superannuation industry has seen remarkable growth in recent times, and has become an extremely important sector and drives the funds management industry in Australia. The main objective of a superannuation fund is to create wealth to enable its members to enjoy a reasonable standard of living in retirement. In this seminar, we first give an overview of the superannuation industry in Australia, and then focus on integrating data science and super. In particular, we discuss two fundamental data science challenges in Australia's superannuation industry. The first problem is the member de-duplication. The average number of super account per employee is above 3 in 2008 and increasing [1], and we have to identify these duplicated members so we can perform data science on a member level. To this end, we developed a fast de-duplication algorithm using Bayesian inference with attribute clustering specifically for the super industry. The second problem is member journey analytics. Superannuation is a life-long journey and members usually goes through a path/sequence of different touch points. This can be modelled as a Multi-channel attribution problem [2]. We use Markov chain models where we represent every member journey as a chain in a directed Markov graph, and control the cohort, year and time effects. Results can be used to identify and recommend the next best action to keep member on a good pathway or move members to a good pathway. Finally, we give a summary of other data science works for the superannuation industry, including predictive models (propensity to defect, job change suspects, member engagement), and campaign evaluation methods.

REFERENCES

[1] Fear, J, and Pace, G. Choosing Not to Choose Making Superannuation Work by Default, The Australia Institute Industry Super Network, Nov. 2008, Discussion Paper 103.

[2] Anderl E, Becker, I, Wangenheim, Schumann, J. Mapping the Customer Journey: A Graph Based Framework for Online Attribution Modelling, 2014.

Ensuring big data reliability with minimum replication for cost effectiveness in the cloud

Prof. Yun Yang[1]

[1]*Swinburne University of Technology, Australia*

In current cloud computing environments, management of big data reliability has become a challenge. For data-intensive scientific applications, storing data in the cloud with the typical 3-replica replication strategy for managing the data reliability would incur huge storage cost. To address this issue, in this talk we present a novel cost-effective big data reliability management mechanism named PRCR, which proactively checks the availability of replicas for maintaining data reliability. Our simulation indicates that, comparing with the typical 3-replica replication strategy, PRCR can reduce the storage space consumption by one-third to two-thirds, hence reduce the storage cost significantly in the cloud for big data.

Making Big Data work across disciplines – The AURIN Solution

Dr. Jack Barton[1]

[1]*AURIN, Australia*

Big Data transcends all disciplines, across many institutions, organisations and government bodies. It comes with its own big issues around access, licensing, integration, interrogation, interpretation and visualisation.
AURIN, a National eResearch Infrastructure, is providing a Big Solution. AURIN Workbench and its flagship application, the AURIN Portal, as delivering access to diverse data from multiple sources, and facilitating data integration and data interrogation using open source e-research tools. Through this, facility researchers are able to gain meaningful knowledge that provides the evidence base researchers need to make informed decisions for the smart growth and the sustainable development of Australia's cities and towns.
This presentation led by Dr Jack Barton will describe the AURIN solution for breaking down data barriers and opening access for researchers, and providing tools for analysis and interpretation

Mining complex data from highly streaming environment

Zhinoos Razavi Hesabi[1]*, Timos Sellis[2] and Jenny Zhang[1]

[1]School of Computer Science, University of RMIT, Melbourne, Australia
[2]Department of Computer Science and Software Engineering, Swinburne University, Melbourne, Australia.

*E-mail: zhinoos.razavi@rmit.edu.au

In the current digital era the massive progress and development of internet and online world technologies (e.g. filmless imaging, online Global Positioning System and so on) we counter huge volumes of information and different data types day by day from many different resources and services which was not available to human kind just a few decades ago. This data comes from available different online resources and services which are established to serve the customers. Services and resources like Sensor Networks, Cloud Storage, Social Networks etc., produce big volumes of data and also need to manage and reuse that data or some results stemming from the analysis of the data. Although this massive volume of data can be really useful for people and corporates it could be problematic as well; therefore, big volume data, or big data, has its own deficiencies as well. They need big storage capacities, high bandwidth for transmission over internet and also this volume makes operations such as analytical, process, and retrieval operations difficult and hugely time consuming. One resolution to overcome these difficult problems is to have big data summarized so they would need less storage and much less time to get transmitted, processed and retrieved. The summarized data will be then in "compact format" and still serve as an informative version of the entire data. Although summarization can be performed in various ways the aim of this talk is to introduce new big data summarization algorithms and frameworks through clustering and compression techniques to lessen the limitations of the traditional big data summarization approaches, especially in the context of streaming data. The focus will be the development of new summarization algorithms compatible with much more complex data such as positional data (e.g. GPS coordinates), or big 2D arrays (e.g. large medical images) or distributed streaming data (e.g. sensor data). In this talk we will try to cover the aforementioned aspects and new concerns of big data by introducing new summarization algorithms.

Big Data in Healthcare Operations - a Tale of 2 Sites, 1 City

A/Prof. Christopher Bain[1]

[1]*Mercy Hospitals Victoria, Australia*

Data usability with personalised autonomous mining in health care systems

Prof. Hua Wang[1]

[1]*Victoria University, Australia*

Data Driven Decision Making: From Remote Health Monitoring to Venture Prediction

Dr Sanjeev Naguleswaran[1]

[1]*QSpectral Systems, Australia*

Health systems face an increasing demand on resources, as the world's population ages. Technology in the form of sensor-based remote monitoring systems has emerged as a viable option to mitigate this asymmetry between supply and demand enabling vulnerable individuals to live at home safely and independently. We will present novel data driven systems for sensor based human activity recognition, chronic disease monitoring as well as health risk assessment. The advance of analytics methods provides a means to contextualise and aggregate data from disparate devices in a meaningful manner. In this presentation, we will also discuss applications across a number of domains including the exciting area of data driven evaluation of startups.

Cognitive Computing for skin analytic and early detection of Melanoma

Mani Abedini[1]

[1]*IBM Research, Australia*

Australia and New Zealand have the highest rate of Melanoma incident in the world. Melanoma is the deadliest type of skin cancers. Excision of thin melanoma at early stage has been shown to be an effective prognosis. However, clinical diagnosis of melanoma at early stage is subjective and highly reliant on clinicians' experience. Therefore, automated analysis of skin images is extremely useful to provide more objective diagnosis and collect evidence according to well-established clinical guidelines such as ABCD (Asymmetry, Border irregularity, Color and Dermoscopic pattern). In this talk, we will take a look the advanced visual analytic developed in IBM Research Australia to target accurate diagnosis of Melanoma.

Handling Health-care Data using Polyglot-persistence

Karamjit Kaur[1]

[1]*Computer Science and Engineering Department, Thapar University, Patiala, Pb., India*

Healthcare Information Systems (HIS) are multifarious in nature and thus can be best implemented by using multiple data-stores because one database definitely doesn't fit all the storage requirements of such complex applications. Amalgamation of different databases within an application is known as Polyglot-persistence. Pre-requisite for achieving polyglot-persistent solution is the availability of different types of databases. As late as 2005, relational databases ruled as the de-facto databases but now their reign is challenged by the advent of non-relational databases known as NoSQL data-stores, making Polyglot Persistence possible.

Acknowledgement: This research has been supported by University Grants Commission, New Delhi, India under Major Research Project Grant F. No. 42-135/2013 (SR).

REFERENCES

[1] J. Sadalage and M. Fowler, NoSQL distilled: a brief guide to the emerging world of polyglot persistence. Pearson Education, 2012

[2] M. R. Genesereth and S. P. Ketchpel. Software agents. Commun. ACM, 37(7):48–53, 1994.

[3] S. S. Huang, T. J. Green, and B. T. Loo. Datalog and emerging applications: an interactive tutorial. Proceedings of the 2011 ACM SIGMOD, pages 1213–1216.

Lost in Translation: On the Disconnect between Business-speak and Data Science

A/Prof. Michael Brand[1]

[1]*Monash University, Australia*

Data Science is on the intersection between business-driven analytics and data-driven scientific enquiry. From difficulties in operationalisation, through the immaturity of tools for automation, to result unrepeatability, many of the key, recurring problems of modern data science can be better understood and better tackled when viewed through this dual-lens of science needs vs business needs. We will demonstrate this through a sequence of real-world examples, and will explore the tools with which businesses are overcoming these teething pains of the modern data-driven digital era.

Big Data: an almighty concept in Chinese research and capital market

Dr. Alex (YiFei) Dong[1]

[1]CUCPAY, China

Big data has been widely used in various Chinese industries. Most research projects are named after “Big data” words to obtain the grant easier because this is one the hottest topics in current Chinese research area. Moreover, this concept is appreciated by capital market as well. Hence most startup companies are pursuing the application of “Big Data” in different virgin territories to attract the attention of investors. In addition, the combination of big data and Fintect makes it finally as an almighty concept in China. Particularly, big data has become the main profit source of payment and credit rating industry.

From Digital to Data-Centric

Tim Cahill[1]*

[1]*The Conversation Media Group, Australia*

*E-mail: tim.cahill@theconversation.edu.au

The Conversation is an independent source of news and views, sourced from the academic and research community and delivered direct to the public. Our newsrooms are located in Melbourne, Boston, London, Paris, Johannesburg.

Since our launch in March 2011, we've grown to become one of Australia's largest independent news and commentary sites, with over 3.5m readers per month visiting our site, and over 35m readers accessing our content via republication under our Creative Commons license.

In the process, we have left behind as massive data footprint that allows us to profile every aspect of our operational and business processes – from our core editorial product, to our day-to-day financial processes, from our internal communications to our stakeholder engagement.

We are now using these data to derive insights, deliver optimisaiton, and to drive innovation. As a not-for-profit, and with a very small data science team, we are developing innovative workflows and tools to unlock the inner data analyst that lives inside everybody across our organisation.

A business analytics approach to injury risk management for Australian football

A/Prof Kok-Leong Ong[1]

[1]*La Trobe University, Australia*

While much of data science focuses on the statistical and modelling precision of the data, business analytics draw upon the methodologies inspired by business consulting and project management to connect a business problem to a data driven solution that reflect the real-world constraints in which one has to operate in. Consequently, the focus is not only on finding the suitable model but also the fit of the model within the broader business solution context. In this presentation, I will address the injury risk management problem in Australian football as an exemplar. The presentation will conclude with a big data angle to the current solution.

Data driven innovation and innovation driven data

Dr. Alexe Bojovschi[1]

[1]*AIB, Australia*

Churn Analytics and the impact to business

Mr. Gavin Whyte[1]

[1]KPMG, Australia

The customer churn analysis feature helps you identify and focus on higher value customers, determine what actions typically precede a lost customer or sale, and better understand what factors influence customer spending. When you improve customer retention, you substantially improve the bottom line.
This talk is about how to achieve the optimal churn solution to make a substantial impact to the bottom line, going through basic and advanced feature engineering and covering topics that allows the data scientist to make accurate predictions. A look into how to formulate a customer engagement methodology that really makes an impact, will be discussed in detail.

Big Data - A Key Driver in the Digital Revolution

Mr. Stuart Growse[1]

[1]*GCS AGILE, Australia*

Unlocking the commercial value of Big Data in a digital age

Lee Anderson[1]

[1]*Delloite, Australia*

The topic will discuss the various strategies and approaches being used by the leading Australian businesses to drive value from their investments in big data – and the implications for data, technology, people and process.

Data Aggregation and Security

Dr. Andy Song[1]

[1] *RMIT University, Australia*

Cognitive Computing Transforming Healthcare

David Yip[1]

[1]*IBM Research, Australia*

Scalable data management and analysis for bioinformatics

A/Prof. Uwe Roehm[1]

[1]*University of Sydney, Australia*

We are in the middle of a paradigm shift to data-driven sciences where automated scientific experiments allow us to collect data of unprecedented volume and at increasing speed and lower costs. For decades, database systems have been the "go-to" technology for large-scale data management. This has changed with the recent interest in “Big Data” analysis.

This talk explores the potential and the current limitations of using database technology for data management and analysis of bio-data. We will look at specific use cases in genomics and immunology. The talk will also give an overview of the BioSeqDB project, in which we explored the applicability of extensible databases and SQL for declarative processing of bio-data. One interesting result was a new efficient algorithm for error-correcting raw sequence data, called Blue, that combines statistical methods and scalable data processing algorithms based on k-mer consensus. Blue outperforms existing error-correction algorithms by up-to two orders of magnitude in throughput while achieving higher accuracy on both Illumina and 454 data.

Genomics in the battle against antibiotic resistance

Dr. Kelly Wyres[1]

[1]*University of Melbourne, Australia*

Following the discovery of penicillin in the late 1920s, antibiotics have become a cornerstone of medicine, providing highly effective treatment for a raft of bacterial diseases. However, despite an enormous reduction of disease burden, bacterial infections remain a considerable cause of morbidity and mortality worldwide. Furthermore, the emergence of antibiotic resistant strains now threatens our ability to control disease: organisms that were once fully susceptible to these drugs can rapidly develop or acquire resistance, creating so-called 'superbugs' that are extremely difficult to treat. As a consequence, antibiotic resistance is considered one of the greatest public health threats of our time.

The markers of antibiotic resistance can be detected in bacterial DNA. High-throughput analysis of the full complement of DNA in a bacterial cell, known as the 'genome,' can provide a wealth of information about the evolution and spread of antibiotic resistance. As the cost and ease of DNA sequencing has improved, it has become possible to implement these technologies in hospitals and public health labs, where they are poised to transform the way in which we diagnose, treat and monitor bacterial disease. Here we will explore the use of genomic sequencing for; 1) better targeting of antibiotic treatment; 2) investigation of hospital 'superbug' outbreaks; and 3) large-scale epidemiological surveillance. The preservation of antibiotics is essential for the future of modern medicine, and these genomic analyses can play an important role.

Building Ensemble of Models by Exploiting the Richness of Feature Variables in High-Dimensional Data and Application in Protein Homology

Jabed H. Tomal[1], William J. Welch[2] & Ruben H. Zamar[2]

[1]Department of Computer and Mathematical Sciences, University of Toronto, Toronto, Ontario, Canada
[2]Department of Statistics, The University of British Columbia, Vancouver, British Columbia, Canada

High-dimensional data contain a large number of observations and feature variables. In this work, we have developed a model which uses the richness of information presents in the large number of feature variables in high-dimensional data to predict a response variable. The proposed model - which is an aggregated collection of logistic regression models (LRM) - is called an ensemble, where each constituent LRM is fitted to a subset of feature variables. An algorithm is developed to cluster the feature variables into subsets in a way that the variables in a subset appear to be good to put together in an LRM, and the variables in different subsets appear to be good in separate LRMs. The strength of the ensemble depends on the algorithm's ability to identify strong and diverse subsets of useful feature variables present in high-dimensional data. We named each subset of variables a “phalanx”, and the resulting ensemble an “ensemble of phalanxes”.

Homologous proteins are considered to have common evolutionary origins. To produce an evolutionary sequence of proteins, a scientist needs to predict their biological homogeneity. The proposed ensemble of phalanxes has been applied to predict biological homogeneity of proteins using feature variables obtained from the similarity search between a native protein and a candidate protein. The feature variables represent structural similarity of proteins, and the underlying assumption is that the structural similarity of proteins is predictive to their biological homogeneity. Considering scarcity of homologous proteins, the prediction performances of a model are evaluated by checking its ability to rank rare homologous proteins ahead of the non-homologous proteins. The protein homology data are obtained from the 2004 knowledge discovery and data mining (KDD) cup website. While the prediction performance of an ensemble of phalanxes is competitive to contemporary state-of-the-art ensembles and the winning procedures of the 2004 KDD cup competition, a further improvement in prediction performances is achieved by aggregating two diverse ensembles of phalanxes obtained from optimizing two complementary evaluation metrics. Through parallel computing, the proposed ensemble is shown computationally efficient as well.

The Next-Gen Information Management System – The Zetaris Way

Dr John Brudenell[1]

[1]ZETARIS, Australia

Today's big data technologies need to support changes in the qualitative use of information generated from data, adding to it, curating it and exploiting it effectively. At Zetaris, we deliver a suite of services to our clients covering data strategy, including digital and analytics strategy, data design and data quality assessment, business data modelling, design and implementation, data-centric consulting services, data product optimisation and digital analytics infrastructure design and implementation. Capturing and making sense of masses of broad, unstructured data such as voice, email, social media content and so on, is driving many new technology offerings. Solutions such as Hadoop and Spark have emerged along with business intelligence packages that sit on top of them. However, these are only part of the solution. The ultimate value for organisations comes from marrying this unstructured and semi-structured data with the structured data that already exists within the organisation and doing it a timely and focused way, driven by business needs. Only then can you personalise customer-facing applications across the customer life-cycle, draw connections between customer behaviour and business events such as the billing cycle or new product launches, and generate targeted analytics that reveal profit opportunities.

Caches all the way down: Infrastructure for Data Science

Prof. David Abramson[1]

[1]*Research Computing Centre, University of Queensland, Australia*

The rise of big data science has created new demands for modern computer systems. While floating performance has driven computer architecture and system design for the past few decades, there is renewed interest in the speed at which data can be ingested and processed. Early exemplars such as Gordon, the NSF funded system at the San Diego Supercomputing Centre, shifted the focus from pure floating point performance to memory and IO rates. At the University of Queensland we have continued this trend with the design of FlashLite, a parallel cluster equiped with large amounts of main memory, Flash disk, and a distributed shared memory system (ScaleMP's vSMP). This allows applications to place data "close" to the processor, enhancing processing speeds. Further, we have built a geographically distributed multi-tier hierarchical data fabric called MeDiCI, which provides an abstraction very large data stores cross the metropolitan area. MeDiCI leverages industry solutions such as IBM's Spectrum Scale and SGI's DMF platforms.
Caching underpins both FlashLite and MeDiCI. In this talk I will describe the design decisions and illustrate some early application studies that benefit from the approach.

Multiview Learning

Dacheng Tao[1]

[1] *Faculty of Engineering and Information Technologies, School of Information Technologies University of Technology Sydney, Australia*

In recent years, many algorithms for learning from multi-view data by considering the diversity of different views have been proposed. These views may be obtained from multiple sources or different feature subsets. For example, a person can be identified by face, fingerprint, signature or iris with information obtained from multiple sources, while an image can be represented by its color or texture features, which can be seen as different feature subsets of the image. In this talk, we will organize the similarities and differences between a wide variety of multi-view learning approaches, highlight their limitations, and then demonstrate the basic fundamentals for the success of multi-view learning. The thorough investigation on the view insufficiency problem and the in-depth analysis on the influence of view properties (consistence and complementarity) will be beneficial for the continuous development of multi-view learning.

AI for Healthcare

Dr. Truyen Tran[1*]

[1] *Centre for Pattern Recognition and Data Analytics, Deakin University Geelong, Australia*

*E-mail: truyen.tran@deakin.edu.au

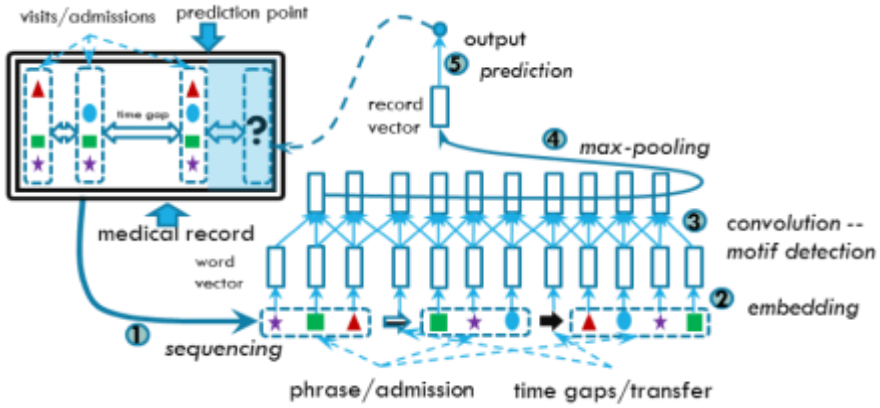

Figure 1: Deepr: A deep neural network for medical records [7].

Recent advances in Artificial Intelligence, collectively known as Deep Learning [5], have revolutionised data-intensive domains such as vision [4], speech recognition [3] and NLP [6]. The main power of Deep Learning is the ability to deliver end-to-end systems that learn from raw data, extract useful features, and predict outcomes with little human intervention [1]. Most of the successes thus far are in cognitive domains where an average human does well within less than a second. Little is known for healthcare, in which humans need specialised training to perform reasonably. This lecture will explain the theory and practice of Deep Learning with emphasis on healthcare.

The lecture consists of two parts. Part I: Introduction to Deep Learning, where I briefly discuss the key concepts in machine learning, feature learning and deep learning. I then then cover three main neural architectures: feedforward, recurrent and convolutional networks. Part II: Applications in healthcare, where I explain how Deep Learning can be applied for healthcare [7, 9–12] – especially when the goal of analytics is not only high predictive accuracy [8] but also transparency [7, 12] and interpretability [2].

REFERENCES

1. Y. Bengio. Learning deep architectures for AI. Foundations and trends® in Machine Learning, 2009, 2(1):1-127.
2. S. Gopakumar, T. Tran, D. Phung, and S. Venkatesh. Stabilizing linear prediction models using autoencoder. International Conference on Advanced Data Mining and Applications (ADMA 2016), 2016.
3. G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. Signal Processing Magazine, IEEE, 2012, 29(6):82-97.
4. A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems 25, 2012, pages 1106-1114.
5. Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. Nature, 2015 521(7553):436-444.
6. C. D. Manning. Computational linguistics and deep learning. Computational Linguistics, 2015.
7. P. Nguyen, T. Tran, N. Wickramasinghe, and S. Venkatesh. Deepr: A Convolutional Net for Medical Records. Journal of Biomedical and Health Informatics (arXiv preprint arXiv:1607.07519),2016.

8. T. Nguyen, T. Tran, S. Gopakumar, D. Phung, and S. Venkatesh. An evaluation of randomized machine learning methods for redundant data: Predicting short and medium-term suicide risk from administrative records and risk assessments. arXiv preprint arXiv:1605.01116, 2016.
9. T. Nguyen, T. Tran, D. Phung, and S. Venkatesh. Latent patient profile modelling and applications with mixed-variate restricted Boltzmann machine. In Proc. of Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Gold Coast, Queensland, Australia, April 2013.
10. T. D. Nguyen, T. Tran, D. Phung, and S. Venkatesh. Tensor-variate restricted Boltzmann machines. In Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.
11. T. Pham, T. Tran, D. Phung, and S. Venkatesh. DeepCare: A Deep Dynamic Memory Model for Predictive Medicine. Proc. of Pacific- Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Auckland, NZ, arXiv preprint arXiv:1602.00357, 2016.
12. T. Tran, T. D. Nguyen, D. Phung, and S. Venkatesh. Learning vector representation of medical objects via EMR-driven nonnegative restricted Boltzmann machines (eNRBM). Journal of biomedical informatics, 2015, 54:96-105.

Enabling Intelligent Transport Systems (ITS) with cognitive analytics

A/Prof. Quoc Bao Vo[1]

[1]*Swinburne University, Australia*

Intelligent Transport Systems (ITS) has the potential to greatly improve quality of life by using information and communication technology (ICT) to enhance safety, performance and interactivity of transport systems while reducing costs. In the context of digital cities, the key drivers behind ITS are the Internet of Things (IoT) with numerous sensors, big data technology and mobile devices coupled with innovative services such as transport and traffic management and traveller information services. Our research program aims to develop new innovative services for ITS based on the concept of analytics-enabled cognitive agents. Cognitive analytics work seamlessly with the user who can be traffic manager or driver of a vehicle and leverage advanced data analytics and machine learning algorithms to provide context-dependent and goal-driven advices to its user. The long-term objective of this research program is to have the cognitive analytics system evolving from a shared decision-making mode as part of a human-agent team to the fully autonomous mode where the agent can make decision and act on behalf of its human user.

Privacy and security in the age of big data

James A. Thom[1]*

[1]*School of Science, RMIT University, Victoria 3001, Australia*

*E-mail: james.thom@rmit.edu.au

Big data raises many security and privacy challenges [1] that need to be addressed by organisations storing and analysing data in the cloud. To protect the privacy of personal and sensitive information, it is essential to adopt appropriate security mechanisms and policies.

The Australian Information Commissioner [2] has recently identified some of the privacy risks associated with big data, as well as best practice for organisations to avoid these risks. Nevertheless, there are many recent examples of privacy issues in Australia relating to big data. These issues include policy around retention of telecommunications metadata, the effectiveness of security measures protecting the privacy of medical records, and personal data being stored offshore.

Social networks contain large volumes of personal information. When individuals share personal information on social media they may intend through their privacy settings to control the visibility of the information they share. However, default settings, changes in these settings, and the complexity of the implementation of settings can make effective control difficult. Indeed, there is a complex relationship between personal privacy and security of data.

REFERENCES

1. Mashima, D. and Rajan, S.P. (editors). Big Data Security and Privacy Handbook, August 2016, Cloud Security Alliance, Big Data Working Group. https://downloads.cloudsecurityalliance.org/assets/research/big-data/BigData_Security_and_Privacy_Handbook.pdf

2. Office of the Australian Information Commissioner, Guide to big data and the Australian Privacy Principles, Consultation draft, May 2016, Australian Government. https://www.oaic.gov.au/resources/engage-with-us/consultations/guide-to-big-data-and-the-australian-privacy- principles/consultation-draft-guide-to-big-data-and-the-australian- privacy-principles.pdf

Performance Analytics Using Tracking Data and Computer Vision: A New Statistical Era for Elite Tennis

Dr Stephanie Kovalchik[1]

[1]Game Insight Group at Tennis Australia & Institute of Sport, Exercise & Active Living, Australia

Professional tennis, like many of the most popular sports around the world, has seen a massive growth in tracking data and computer vision in recent years. Although there is more data than ever to address a wide array of questions about tennis athlete development and performance, analyses of these data have been rare. In this talk, I will introduce three areas in which we at the Game Insight Group of Tennis Australia are analysing tracking data and vision to advance how we measure and understand the physical and mental side of performance in elite tennis. The presented work includes a metric for player consistency in their preparatory routines, a measure for the amount of work performed and changes of direction during rallies, and the identification of facial expressions during matchplay. I will discuss what these new developments are teaching us about the sport and how they are contributing to a new wave of statistical innovation in tennis.

Big Data Analytics for Data Streams from Sensors

A/Prof. Tao Gu[1]

[1] *Computer Science, School of Science, RMIT University, Australia*

With an increasing number of sensors and mobile devices we have seen today, analysis of sensor data streams has become a key research area in the big data era. There are many challenges arise when dealing with the evolution over time of such data streams. This talk is to introduce some typical techniques in analysing sensor data streams. Such techniques range from classification, clustering, to pattern mining. I will also report some of the work we did in mining sensor data streams in real time.