Institute for Big Data Analytics

To create knowledge and expertise in
the field of Big Data Analytics by facilitating
fundamental, interdisciplinary and collaborative
research, advanced applications, advanced training
and partnerships with industry.

Research - Training - Outreach

Newsletter

Spring 2015

Message From the Director

Winter and spring were a very busy time for the Institute. We have worked on a number of projects, some of them described in this issue. Some other exciting developments include our work with the Orange dataset: the global telecommunications service provider Orange has made available to us a complete set of the use, over a period of one year, of mobile phones in a country. Working with that data we were able to identify some events we did not know about beforehand, in time and space. Moreover, we were able to use this dataset to experiment with a new data privacy technique designed specifically for mobile phone use records. Our approach, on the one hand, completely de-links user personal, identifiable information from the mobile phone usage information. On the other hand, the data is still adequate for performing certain elementary types of analytics and other, more complex, analyses of the data can then build on the results of these elementary steps.

Summer time is particularly exciting as we welcome then a number of students for internships. We host this year students from Brazil, Mexico, India and Poland working with us on a variety of projects.

We are also actively involved in the Canadian AI conference held in Halifax June 2-5. Our students present three papers related to Institute projects. And just a week later we are co-organizing the Big Data Workshop, sponsored by the Fields Institute and AARMS (Atlantic Association for Research in the Mathematical Sciences), in which several Institute researchers present their work and participate in a panel on training of Data Scientists.

We are also bringing in a new cohort of graduate students to Dal, Simon Fraser and Université de Montréal. Many of them benefit from the NSERC CREATE grant for training in Big Text Data lead by Dal, with participation of the other two universities. This is an industrially oriented program, so the students will all have to do an internship with a company in the area related to Big Data. If you are a company interested in hosting such a student, please get in touch with us at bigdata@cs.dal.ca.

Big Data Research in Seniors Homes

A lot of interesting research problems come to the Institute For Big Data Analytics from industry partners who are grappling with growing quantities of data and the general awareness that value can be derived from this data if they can find the tools and techniques to use it. One such partner is Shannex, a regional company which provides services and facilities in home care, retirement living and nursing homes. Shannex became aware that their growing collection of text data concerning patient histories contained information that might help them to improve their services if they could find a way to access the hidden value. Our project with Shannex, funded by the Province of Nova Scotia is to the use the tools of text mining to to study the correlations that might exist with incidences of patients falling. According to Masters student Lulu Huang this is not a simple task as the data contains a lot of medical terms and abbreviations which are difficult for the non-expert to understand. The United Medical Language System (UMLS) is employed to make sense of the medical terminology and then a model is built to extract from the data the risk factors associated with instances of falling. Early results indicate that among the most significant risk factors are heart and circulatory problems, family stresses, weight loss and even the type of neighbourhood surrounding the facility. These findings will help Shannex to ameliorate the risks and reduce the instance of falling amongst the people in their care.

Masters Student Lulu Huang

DALHOUSIE UNIVERSITY
INSTITUTE FOR
BIG DATA ANALYTICS

News

An Interview with Dr. Tom M. Mitchell

In May Dr. Tom M. Mitchell, chair of the Dept of Machine Learning at Carnegie Mellon was in Halifax to receive an honorary degree, Doctor of Laws (honoris causa).

In addition to being one of the world's most influential computer scientists, Dr. Mitchell is a member of the United States National Academy of Engineering, a Fellow of the American Association for the Advancement of Science and of the Association for the Advancement of Artificial Intelligence. He is or has been on the advisory boards of seven leading journals in artificial intelligence and cognitive science. In 1997, Tom Mitchell published the book Machine Learning, the standard reference on the topic for generations of graduate students.

In 2006, Dr. Mitchell founded the Machine Learning Department within the School of Computer Science – the first and only such department in the world – and was appointed Chair, a position he holds today. His key areas of research involve using brain imaging to understand human language processing, and leading the team that created and runs the Never-Ending Language Learner, or NELL, the first computer system to attempt, in analogy with human learning, to learn many different skills and knowledge over years of continuous learning.

We caught up with Dr. Mitchell for a brief interview while he was here.

Q: *Since you started the Machine Learning Department at Carnegie Mellon, how as the field developed?*

TMM: The Machine Learning Department started in 2006, growing out of an earlier Research Centre for Automated Learning and Discovery. Probably the greatest change since that time is that today many more people know the phrase "Machine Learning" and know that it is about computers using data to improve how they function. This is a confluence of the fact that there is much more data online to work with, there is great progress in the field of algorithms and society is now into the idea of evidence-based decision making.

Q: *Are other universities following your example or are you still the only Machine Learning Department?*

TMM: There are a number of other research centres but we are still the only university department. There might be others in the near future.

Q: *What accomplishment are you most proud of?*

TMM: There are two really. The first is the development of the underlying theory of Machine Learning through, for example, the use of co-training. This is used in NELL – by training many functions that are coupled you can make unlabelled data valuable for training. The second is at the application level. In work with Marcel Just on brain imaging with fMRI data we created an algorithm to decode what people are thinking and for studying how brains represent meaning and understand language; this is now a valuable technique for cognitive science research.

Q: *What concerns you most in your field?*

TMM: I have two main concerns. Privacy is the first one. Privacy is important to all of us. Still, we need to have a better understanding of the trade-off between the impact on privacy of making specific personal data available, and the benefits to society. For example, is it worth making personal medical records available to a third party if they can use everybody's records to discover improved medical treatments?. There are many things we could do of immense value that we are not doing due to privacy issues. The second concern is ownership. There is a lot of data owned by private companies for which they may have no incentive to share. There is currently no discussion on this topic.

Q: *What are your thoughts on some of the anxieties concerning Artificial Intelligence and its potential one day to be a threat to humans?*

TMM: We may not reach a point where this becomes an issue for another 100 years but even so we shouldn't dismiss these anxieties. We have the time to consider how to approach the issue.

Q: *Where do you think Big Data is heading in the years to come? Where will it be in five years?*

TMM: The short answer is that it will get bigger and better. One of the interesting developments that I expect to see in the next five years is a big increase in embedded continuous data systems. There may be regulatory changes in the near future to allow people access to their own data, and there may arise a data economy in which the trading and sharing of data becomes facilitated through monetizing it.

Q: *Do you have any advice for students looking towards a career in Big Data or Machine Learning?*

TMM: I would encourage students to pursue this path. It's going to be a great career as we are only at the beginning of a decades long trend. The growth of evidence-based decision making will continue to expand across all walks of life.

News

Halifax hosting International Conference on Knowledge Discovery and Data Mining

Over 1,000 of the world's leading edge researchers and practitioners in big data are coming to Halifax for the 2017 Conference on Knowledge Discovery and Data Mining.

Stan Matwin, Canada Research Chair (Tier 1) at the Faculty of Computer Science, Dalhousie University and the director of the Institute for Big Data Analytics, announced today, March 9, that Halifax was the successful bidder. The conference, with Dr. Matwin as the general chair and Evangelos Milios of Dalhousie University as the local chair, will be held in the new Halifax Convention Centre, which opens in 2017.

The bid was led by the Institute for Big Data Analytics and the Halifax Convention Centre, in collaboration with a local host committee of academic, government and industry representatives.

"This announcement is further evidence that the Institute for Big Data Analytics has established Dalhousie and Halifax as leaders in the global fields of data science and big data analytics," said Dalhousie president Richard Florizone.

"This is a great opportunity for Dalhousie -- as well as other local organizations and institutions -- to showcase the world-class research, ongoing collaboration and pool of talent we have here in the region to national and international audiences," said Dr. Matwin. Halifax now ranks amongst top cities who have previously hosted the conference, including:

-- Sydney, Australia
-- New York City, New York
-- Beijing, China
-- Paris, France

"We're proud to partner and collaborate with our local experts to host this conference and showcase Nova Scotia's strengths in big data to the world," said Scott Ferguson, president and CEO of Trade Centre Limited, the Crown Corporation that manages the convention centre. "This is an exciting opportunity for our industry, academic and research communities to highlight their work and connect with their global counterparts."

Nova Scotia has a booming information technology sector and the province is quickly establishing itself as an international hub of excellence in big data research. Hosting this conference will allow the local sector to benefit from top academic research and industrial presence aimed at promoting collaboration and growth in big data analytics.

First established in 1995, the Knowledge Discovery and Data Mining conference is the premier international forum for data mining and big data research, bringing together practitioners from academia, industry, and government to share their ideas, research results and experiences. Sponsors of note include Microsoft, Yahoo, Deloitte, Accenture, Facebook, LinkedIn, Google and IBM. The 2016 conference will take place in San Francisco.

Dr. Rob Warren presents paper at Museums on the Web 2015

In April Dr. Rob Warren presented his paper, "Data-Driven Augmented Reality for Museum Exhibits and Lost Heritage Sites" at the Museums on the Web 2015 conference in Chicago. The paper reviews the possibilities, pitfalls, and promises of recreating lost heritage sites and historical events using augmented reality and "Big Data" archival databases. It defines augmented reality as any means of adding context or content, via audio/visual means, to the current physical space of a visitor to a museum or outdoor site. Examples range from simple prerecorded audio to graphics rendered in real time and displayed using a smartphone.

Previous work has focused on complex multimedia museum guides, whose utility remains to be evaluated as enabling or distracting. The paper proposes the use of a data-driven approach where the exhibits' augmentation is not static but dynamically generated from the totality of the data known about the location, artifacts, or event. For example, at Bletchley Park, reenacted audio conversations are played within rooms as visitors walk through them. These can be called "virtual contents," as the audio recordings are manufactured. Given that a number of documentary sources, such as meeting minutes, are available concerning the events that occurred within the site, a dynamic computer-generated script could add to the exhibits.

Visitors' experiences can therefore react to their movements, provide a different experience each time, and be factually correct without requiring any expensive redesign. Furthermore, the use of a data-driven approach allows for the updating of exhibits on the fly as researchers create or curate new data sources within the museum. If artifacts need to be removed from an exhibit, pictures, descriptions, or three-dimensional printed copies can be substituted, and the augmented reality of visitor experience can adapt accordingly.

DS 2015 - A Call for Papers

The 18th International Conference on Discovery Science (DS 2015) will be held in Banff (Canada), on 4-6 October 2015, and provides an open forum for intensive discussions and exchange of new ideas among researchers working in the area of Discovery Science. The scope of the conference includes the development and analysis of methods for discovering scientific knowledge, coming from machine learning, data mining, and intelligent data analysis, as well as their application in various scientific domains. See the website for details:
https://ds2015.cs.dal.ca/

Big Data in Fisheries

The Institute for Big Data Analytics was contacted by Professor Boris Worm and PhD student Kristina Boerder from Dalhousie's Biology Department with an idea to investigate the effects of the Marine Protected Areas on fishing stocks. An interesting research opportunity was presented to us: how to use ship tracking data to model global fishing activities to detect areas of over-fishing or fishing in the protected areas. If such a model could be created it could play a part in predicting the scale of exploitation of world fish stocks. Postdoctoral fellow Erico Neves De Souza has played a leading role in researching the application of big data techniques to this problem.

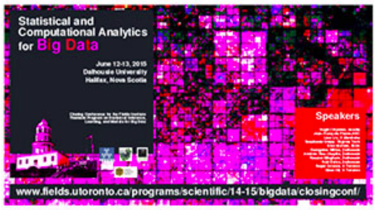
Although at first glance the problem seemed straightforward, early models were not very accurate. There are six or seven main types of fishing vessel and each behaves very differently; a separate model would be required for each type. Addressing these types is beginning to pay off: the model for North Atlantic trawlers has been seen to work with 87% accuracy. Results for modelling longliners, however have not been so good yet.

Other problems arise from inadequate information, incomplete databases and the fact the transponders which supply the positional data can be swapped between vessels. This would lead to completely false information if a signal is recorded as coming from one type of vessel whereas it is actually located a complete different type. Transponders can also be shut off if operators wish their location to be a secret. Such is the temptation in the highly competitive fishing industry where highly productive locations are jealously guarded. And finally, satellites do not cover all areas.

The technology for this project does not already exist, but has to be built from the bottom up. This is a challenge for which it would be helpful to have some funding partners. Although it is early days there is interest from SkyTruth, an organization interested in the application of remote sensing to environmental issues; and Google, already trying to build a model like ours, has also shown some interest. We plan to submit a paper to a Big Data Conference by the end of July.

Statistical and Computational Analytics for Big Data

Presented jointly with the Fields Institute and sponsored by the Canadian Statistical Sciences Institute and the Atlantic Association for Research in the Mathematical Sciences, this workshop will give an overview of highlights of the thematic program on Statistical Inference, Models and Learning for Big Data, will showcase current IBDA student projects and will present research at IBDA on text mining, high-performance computing, visualization, bioinformatics, and privacy.


Statistical and Computational Analytics for Big Data
June 12-13, 2015
Dalhousie University
Halifax, Nova Scotia
Speakers
www.fields.utoronto.ca/programs/scientific/14-15/bigdata/closingconf/

Friday June 12 & Saturday June 13, 2015
Goldberg Computer Science Building
Dalhousie University
Contact: hugh.chipman@acadiau.ca

Director

Dr. Stan Matwin PhD, CRC

Executive Committee

Dr. Michael Bliemel
Faculty of Management (representing the Dean of Management)

Dr. Tom Marrie
Dean of Medicine

Dr. Evangelos Milios
Associate Dean, Research, Faculty of Computer Science

Dr. Andrew Rau-Chaplin
Faculty of Computer Science

Dr. Nur Zincir-Heywood
Faculty of Computer Science

Dr. Stan Matwin
Director, (ex officio), Faculty of Computer Science

Advisory Board

Sue Abu-Hakima Amika
President and CEO,
AmikaMobile, Ottawa, ON

Martin Davis
Vice-President of IT, JD Irving,
Moncton, NB

Peter Hickey
Co-founder, Oris4, Halifax, NS

David Kasik
Senior Technical Fellow,
The Boeing Company, Seattle,
Washington

Stephen Perelgut
IBM Canada University Relations
Manager, Markham, ON

Newsletter Editor:
David Langstroth
dll@cs.dal.ca

"Information is the oil of the 21st century and analytics is the combustion engine."

- Peter Sondergaard, Gartner Research