APPLIED RESEARCH IN ACTION
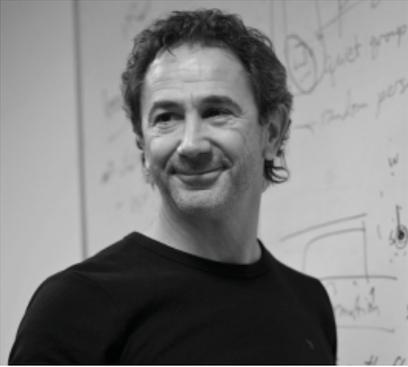
2015

MASTER OF SCIENCE APPLIED COMPUTING (MScAC)

2014-2015 PROJECTS

Computer Science
UNIVERSITY OF TORONTO

MESSAGE FROM THE PROGRAM DIRECTOR

Since its launch on September 2010, the Master of Science in Applied Computing (MScAC) program has educated the next generation of technical leaders, innovators and entrepreneurs. In addition to a wide array of academic graduate coursework, the MScAC offers a unique professional program that unites university research with industrial application through an intensive applied research internship.

We believe the MScAC offers companies a singular opportunity to engage with experienced graduate students who will develop a significant project based on the application of research, from inception to completion. Current and carefully selected prospective students will make lasting contributions to the organization and further advance their careers.

Now in its second year, the Applied Research in Action (ARIA) showcase highlights the accomplishments of the MScAC students who began their studies in September 2014 and their internships in May 2015.

I congratulate the 2014-2015 class, and I am grateful to our industry partners for their ongoing interest and support of the MScAC program.

Professor **Eugene Fiume**, FRSC
Director, Master of Science in Applied Computing program

3D OBJECT DETECTION AND POSE ESTIMATION USING DEPTH CAMERA – Epson Canada Ltd

3D object detection and pose estimation (ODPE) is an important area in computer vision, which has many real world applications, such as robotics and augmented reality. We present a system for 3D ODPE using depth camera. In the first stage of the project, we implemented an existing depth-based ODPE algorithm, which segments the scene into objects and estimates the pose of an object in each segment. However, this and similar methods are only applicable to clean scenes where objects are well separated in 3D space and thus object segmentation is simple. In the second phase of our project, our main contribution, we extend the system to handle complex scenes where objects are much closer to each other. In such conditions, standard segmentation methods would result in over - or under - segmentation. Instead, we generate an over-segmentation of the scene and group the resulting components into multiple overlapping segmentation hypotheses. These hypotheses are passed to the pose estimator to recover a 3D pose. We show that the resulting system significantly outperforms the previous ODPE system based on standard segmentation in complex scenes.

Project Team
Liwen Xu, MScAC Student
Alex Levinshtein, Epson Canada Ltd
Professor **Sven Dickinson**

A MODERATED E-FORUM – Cardiac eHealth

In the Cardiac eHealth and Behavioural Cardiology Research Unit, researchers study the effects of an internet-based behavioural counseling program on the motivation and quality of life of patients with heart failure. The current e-counseling program utilizes an interactive web-based platform, which allows users to obtain expert information and support as they manage their illness through healthy lifestyle changes, including changes in diet, exercise, and smoking behaviors. In order to further enhance the interactive experience that patients with heart failure have on the e-counseling platform, the research unit proposed to design an e-Forum. The e-Forum is web-based software, intended to be a place where patients with heart failure, who currently use the e-counseling platform, continue to be supported in their lifestyle changes, beyond the limits of the structured online program. Because the mean age of patients with heart failure currently enrolled in the study is 57 years, a tailored interface was implemented in order to increase accessibility and usage by this older, and perhaps, less technologically skilled cohort of patients. Special requirements aimed towards enhancing the user experience were established by the research team, who had prior experience with developing web-based applications for this patient population. These requirements included, but were not limited to: smooth design, large font sizes, limited clicks to access information, integration with the existing e-counseling platform, testing, and validating the e-Forum design.

Project Team
Jacqueline Bermudez, MScAC Student
Rob Nolan and **Rika Tanaka,** Cardiac eHealth
Professor **Daniel Wigdor**

A VISUALIZATION ENGINE FOR GIS DATA – Riva Modeling Systems

The objective of the project was to research different ways to explore and analyze data in a geospatial context. We employ asset management data gathered by companies from different industries (public sector, water and wastewater, electric and gas, and others).

The initial business cases were different for every company and included challenges like visualizing different asset attributes on a map, finding spots where there is a high concentration of data points, seeing how asset attributes change over time, displaying high amounts of data without visual clutter, seeing how assets overlap spatially, giving context information about the data being analyzed, among others. The first part of the research involved gathering knowledge about how these problems are currently being approached. We found that this is a research area that is happening mainly within a business environment, and is driven by specific business requirements, so each of these problems are faced separately and on-demand. In our case the scenario is very similar, since every client has its own requirements.

Our solution was to determine which analysis and visualization tools can be used to solve to each of the previous challenges and how to build a centralized solution that can be generalized to similar problems and different industries. As a result, we built a visualization engine that leverages some analysis and mainly visualization tools to solve most of the challenges presented.

Project Team
Patricio Córdova, MScAC Student
Kiran Sachdev, Riva Modeling Systems
Professor **Ravin Balakrishnan**

BUILDING A SYNTHETIC TRAINING DATASET FOR AN INTELLIGENT FALL DETECTION SYSTEM – Toronto Rehabilitation Institute

Falls are the most common cause of injury in older adults, making fall detection systems extremely valuable. Some intelligent fall detection systems can detect falls after being trained on annotated images of people falling and not falling (performing everyday actions). In general, intelligent systems like this tend to perform better with more training data. However, it is time consuming to create large, real datasets. One way to speed up this process is to use a synthetic dataset. Our objective is to generate a large synthetic training dataset using the game engine, Unity. By training on a larger dataset, our goal is to increase the accuracy of a

fall detection system. In Unity, a 'game' was designed where an avatar is animated falling and doing everyday actions in a virtual home. Annotated screenshots of the 'game' were taken and

used to build the dataset. To validate this method of training, the fall detection system will be trained on both real and synthetic datasets of the same size, and the trained classifiers will be tested on real data. We will also investigate the effects of increasing the size of the synthetic dataset and training with a combination of real and synthetic data.
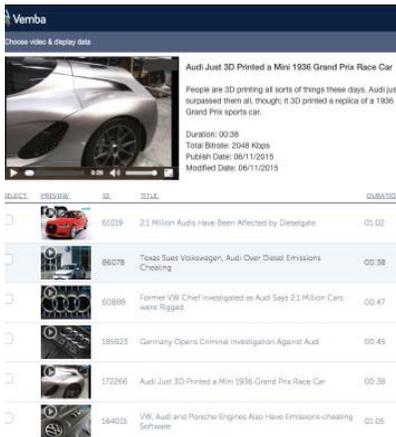
Project Team
Rachel Twiss, MScAC Student
Alex Mihailidis, Toronto Rehabilitation Institute
Professor **Frank Rudzicz**

DEEP LEARNING METHODS FOR VIDEO SIMILARITY METRICS AND VIDEO CAPTIONING –
Vemba Corp.

We are investigating a deep learning method that provides intelligent video recommendations based on the similarity of the semantic content within videos. This semantic content is determined by extracting salient objects from the various data modalities of a given video. For instance, objects can be extracted from a video's image data through fine-tuned convolutional neural networks, and from its audio data using Hidden Markov Models. Finally when text is available, some keywords can be extracted and their semantic relationships determined using a skip gram model with negative sampling. The relative importance of the detected objects is determined by how frequently they appear throughout the video, and how "interesting" the section of the video they appeared in is. Interest in a video section is quantified by an attention model, in which a section of a video that was more frequently watched by viewers is deemed more important than a section that was skimmed through or skipped by viewers. Once the semantic content of the video is extracted and weighted, a learned embedding metric maps the labels into an embedding space. Within this embedding space, similar videos are clustered, which enables the discovery of semantically similar content.

Project Team
Rohan Chandra, MScAC Student
Daniel Ivan, Vemba Corp.
Professors **Richard Zemel** and **Raquel Urtasun** (co-supervisor)

DETERMINING MERCHANT INFORMATION FROM TRANSACTION DESCRIPTION –
SmartFinance LLC

XOBI is a personal finance management mobile app being developed by SmartFinance LLC. Harnessing machine learning, the undifferentiated data tucked in the user's bank statements is transformed into helpful information for the user to make informed decisions. It is a one-stop place to easily and accurately view all of their financial information.

Data about the day-to-day transactions of a user is available as transaction descriptions which are often obscure, inconsistent and difficult for customers to analyze and understand. The first step involves extracting the merchant name and location from the description, which often yields truncated or garbled names and locations. The correct merchant name and other attributes such as location and website are then compiled using resources from the web. The categorization of the merchant and identification of the correct logo is done using various machine learning techniques.

The app also lets the user ask interesting questions like - "When did I last eat sushi?" To answer such specific questions, each merchant is represented using a vector that encodes semantic information about the merchant category.

Project Team
Joms Zacharia, MScAC Student
Professor **Richard Zemel**, SmartFinance LLC
Professor **Raquel Urtasun**

ENHANCED RECOMMENDATIONS – Rakuten Kobo Inc.

Automated recommendations, whether of the item-to-item ("people who bought or liked this product also bought these products") or the user-to-item ("recommended for you") kind, are an important driver of product discovery and sales on most modern e-commerce websites. At Kobo, we are continually optimizing our in-house recommendation system, which is mainly based on co-purchase patterns of items bought on Kobo's website and through Kobo eReaders and free mobile apps. In this project, we explored several distinct ways of improving the system.

To address the cold-start challenge for both books and readers, we incorporated additional data sources other than purchases. For books with no co-purchase information we utilized the

browsing information from our website. For users with no recommendations we now include previews browsed and downloaded by the user as a source for generating recommendations. We explored the use of non-book, higher-level concepts such as similarity between authors rather than books, to enrich and improve our recommender system. Finally, we modified our core recommendation algorithm to better account for individual user temporal purchasing patterns. Since time is a very influential factor in most recommendations, this modification yielded improved results.

Project Team
Hanyang Li, MScAC Student
Inmar Givoni, Rakuten Kobo Inc.
Professor **Brendan Frey**

EXTENDING FPGA CAD SOFTWARE TO SUPPORT EXISTING IP CORES ON NEW FPGA DEVICES – Altera

Altera is a manufacturer of field-programmable gate arrays (FPGAs), which are integrated circuits that can be programmed after manufacturing. Altera also develops computer-aided design (CAD) software that allows users to program the FPGAs. A key part of this software is a catalogue of intellectual property (IP) cores that allow users to easily implement pre-packaged functionalities – for example DDR3 memory interface.

This project aims to extend Altera's CAD software to support existing IP cores on new FPGA devices. Solving this problem involves a variety of unique challenges. These include modeling hardware components in software, developing software that will correctly infer FPGA bit settings from placement data, automatically generating simulation models, and automatically verifying correctness of IP cores.

By solving these problems we were able to extend Altera's Quartus II software to support existing IP cores on a new device. This was verified by running the IP cores through each stage of the CAD flow, and simulating the behaviour of the device. An infrastructure was also put in place to easily track the bring-up of other IP cores, and a tool was developed to automatically generate simulation models.

Project Team
Michael Coles, MScAC Student
Lyndon Carvalho and **Elias El Ferezli**, Altera
Professor **Paul Chow**

FUNDING VALUATION ADJUSTMENT CALCULATION AT MULTIPLE LEVELS – IBM

Funding Valuation Adjustment (FVA), also known as Liquidity Value Adjustment (LVA), is the funding cost/benefit resulting from borrowing or lending the collaterals arising from everyday transactions. For example, when a trader borrows cash from a counterparty or money market, they need to pay the interest, namely, the funding cost. The occurrence of FVA after the last financial crisis in 2007 was because the interest paid on collateral no longer offset the increasing funding cost.

We develop a FVA function which can be computed at the trade, agreement, legal entity and funding set levels. This works as a plugin in our existing system, which evaluates both baseline and real-time deals. The main framework is implemented in C++ and, with hundreds gigabytes of trades to be evaluated per day, the speed of calculation is one of the most important goals. The function takes Mark-to-Market and collateral balance of all corresponding trades as input, and then produces the final FVA numbers by averaging results of all timesteps and scenarios.
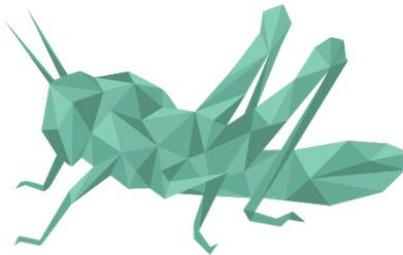
Project Team
Ling Chen, MScAC Student
James Sedgwick, IBM
Professor **Ken Jackson**

IMPROVING DEBUGGING SUPPORT FOR CLOUD PLATFORM – IBM

The project is designed to investigate user experience for debugging on cloud Platform as a Service (PaaS) and provide extension to cloud platform for advanced debugging support. Under the current setting, log tracing and live monitoring are supported for a deployed application on Cloud Foundry based platform. However, all diagnostic data are lost immediately upon a crash since Cloud Foundry redeploys the crashed application to a new container automatically. Many developers are being forced to implement unconventional and highly questionable methods when conducting a post-mortem analysis of their code. The project aims to service the developers' needs and facilitate the treacherous post-mortem analysis process. For further extension, a solution recommended service might be provided based on the discovered solutions for particular issues.

Project Team
Shuying Lin, MScAC Student
Ronald Servant, IBM
Professor **Angela Demke Brown**

LIGHTWEIGHT BUSINESS RULES ENGINE – IBM

A lightweight framework, named Pomelo, serves as an engine of numerous user-defined (mainly targeted on users from financial industry) business rules. Pomelo is brought to free rule developers out of coding-intensive and error-prone jobs, e.g., resource adapting, distributed and parallel programming, fault-tolerance, resource pooling, security management, transaction management, and integration with other enterprise systems etc. Pomelo provides comprehensive infrastructure supports for users to develop and deploy their business rules. With Pomelo, rule developers can focus on the "real" business needs, and translate these business needs into simple Java beans with minimum coding efforts. These simple Java beans could be eventually wired up to form into a Direct Acyclic Graph (DAG) by utilizing Pomelo's Inversion of Control (IoC) container. Once the DAG is instantiated by Pomelo, the remaining amount of work will be taken over by Pomelo framework, where tasks like resource adapting, multi-threading and fault-tolerance will completed by Pomelo automatically. The rules are able to exist by themselves, which are ready to be called from any application or framework.

Project Team
**Marc Huang**, MScAC Student
Tony Dozal, IBM
Professor **Ken Jackson**

MACHINE LEARNING AND NATURAL LANGUAGE PROCESSING IN MOBILE-APP ECOSYSTEM – Addictive Mobility

Addictive Mobility is a leading Canadian mobile in-app advertising technology company. It provides a technology platform that delivers brand engagement, lift and awareness along with the deep insights required for clients to run successful mobile ad campaigns. User engagement and performance of ads are the selling point of the company's products and optimizing these metrics is the goal of the Data Science team. My research internship focuses on using state-of-the-art machine learning algorithms and statistical inference techniques to build (a) a robust User-Profile system that extracts relevant user-interest profiles from more than 4TB of data received every day, (b) a Fraud Detection system to detect fraudulent activities in the mobile ecosystem, (c) an App-Categorizer that uses Natural Language Processing techniques to automatically extract enriched categories of apps, and (d) a Look-Alike Profiling system that, given a profile, can predict a user-base having similar interest profiles for a better and wider audience targeting. All of these systems help us to provide an enriched customer experience through customized recommendations and to build robust user-profiles to achieve a better market segmentation and an improved interest-based audience targeting.
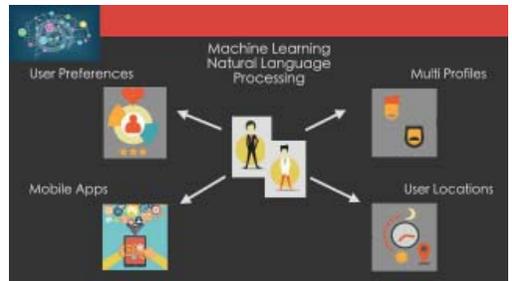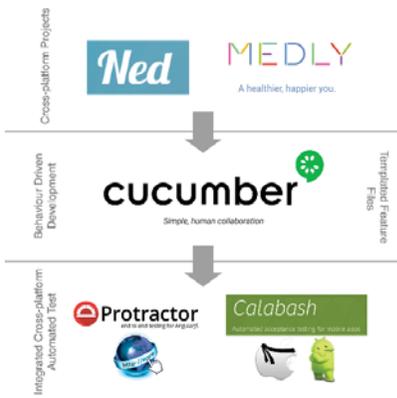
Project Team
Kanika Madan, MScAC Student
Naveed Ahmed and **Elie Mazzawi**, Addictive Mobility
Professor **Raquel Urtasun**

MEDLY AND NED: THE PATH TO INTEGRATED CROSS-PLATFORM AUTOMATED TESTING –
Centre for Global eHealth Innovation

Medly is an Android platform that enables self-care to empower patients with complex chronic diseases. Ned is a platform that supports individuals with prostate cancer and their clinicians, enabling patients and providers to collect patient-reported outcomes, which can be analyzed to prompt behavioural or treatment changes, and inform trends in health outcomes. Arvind and Anson worked on these projects as part of our respective internships. We used a behaviour driven development process using Cucumber feature files. These files are used to automate testing, by linking feature steps to test step definitions. While working on our projects, we identified a redundancy of test scripts being created, as there were no identified templates, and thus, each project defined their own suite of test scripts. As part of our research, under the guidance of our industry supervisor, Melanie Yeung, and our academic supervisor, Dr. Marsha Chechik, we identified commonalities in feature files across multiple projects and technology platforms. We used this to create an integrated automated test framework that can be reused across projects.

Project Team
Jyotheeswar Arvind Manickavasagar and **Anson Liang**, MScAC Students
**Melanie Yeung**, Centre for Global eHealth Innovation
Professor **Marsha Chechik**

MOBILE PATIENT BILLING – McGill University Health Centre

The research project sought to determine the viability of creating a mobile service that would allow physicians to keep a digital database of the patients they treat. A database has tremendous advantages for the physician. For example, a surgeon needs to keep a record of each patient they see for billing purpose. A database also has benefits for medical research such as the observation of changing frequencies for the necessity of a particular procedure.

The industry side of this project saw value in creating a way for physicians to get reimbursed for their services. The traditional method of billing requires an agent and maintaining a physical record of each session with a patient.

The major hurdle of the project was to create a solution that would make adding new records into the database as quickly as possible. It was not feasible to integrate with existing medical software because those services would not allow it. The final solution used a combination of a well-tested user interface, image pre-processing,
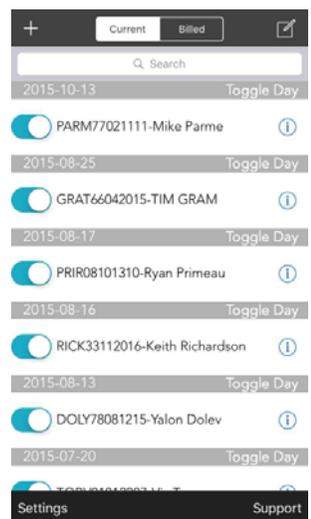
novel algorithms on existing optical character recognition libraries, and a post-processing system that attempts to select the correct information from an image based on previous selected correct option and heuristics.

Project Team
Ryan Primeau, MScAC Student
Keith Richardson, McGill University Health Centre
Professor **Khai Truong**

NATURAL LANGUAGE PROCESSING RESEARCH ON ONLINE TUTORING SYSTEM – NeuRecall Inc.

This project has two phases. Phase I contains three parts: (I. a) extract premises from corpora and chunk passages into memorable phrases. (I. b) concept mapping questions creation. (I. c) short-answer responses and essay evaluation. Phase II contains integration of (II. a) speech synthesis, (II. b) recognition and (II. c) training. For I. a, NLP algorithms would be evaluated on their ability to extract meaningful premises from bodies of online text corpora. In addition to that, the NLP algorithms would be evaluated on their ability to parse text-based materials such as poems, lyrics, scripts, or speeches. For I. b, this part focuses on question creation such as fill-in-the-blank, true-or-false, multiple-choice and matching. Also, this part of research investigates the process of concept mapping question creation. For I. c, there may be wide variation in the range of valid responses that the algorithm must consider. The focus is to extract syntactic and even semantic content from novel sentences. Automated essay scoring (AES) algorithms have now advanced to the point where they can provide scores well-correlated with those of human raters. For II. a, Speech synthesis (Text-to-speech or TTS) should allow for the aural presentation of text to NeuRecall's users. For II. b, for the recording of spoken utterances by NeuRecall users, it would be preferable to store their responses in the form of text rather than memory-intensive audio. However, for audio recordings to be converted to text would require accurate speech recognition. For II. c, it deals with the training of better speech in non-native speakers.

Project Team
Xiaorui Dong, MScAC Student
Simon Alexander Overduin, NeuRecall Inc.
Professor **Suzanne Stevenson**

ONLINE FURNITURE CUSTOMIZATION SYSTEM DESIGN – Massuni

Massuni was developed to address the frustration so many people experience while trying to find a piece of furniture that has the right mix of style, functionality and price. We believe in the power of customization and have developed a quick and simple way for anyone to design furniture that fits seamlessly into their home. Each piece of furniture can be restyled by choosing materials, finishes, trim, door designs, and hardware, and the product can be entirely reconfigured by adding and subtracting components. At any time during the design process, users are able to

generate a high quality rendering that realistically depicts the finished item. An updated price is also shown in real time as changes are made.

As an intern at Massuni, Yue was involved in solving two of the most important problems facing us. First, how to improve our user's experience on 3D rendering with limited calculation power (number of rendering servers). He built the back-end rendering queue to solve this resource allocation problem. And second, how to connect with manufacturers. Basically manufacturers expected some kind of input (files) from us to drive their production process. This is a very hard problem. Fortunately, within the very limited range of our needs, the variation among those files is small. This important discovery led us to apply the idea of "regression." Yue was responsible for designing and building this system into our backend.
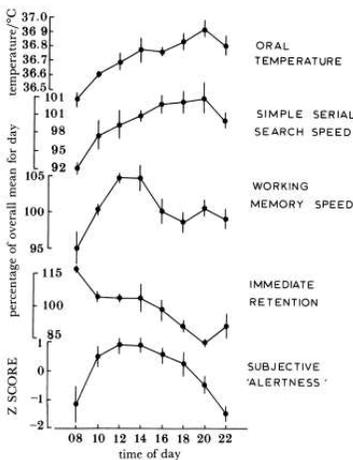
Project Team
Yue Yan, MScAC Student
Jeff Wilson, Massuni
Professor **Christopher Beck**

OPTIMIZING RESPONSE PREDICTION BY AN INTERACTIVE TUTORING SYSTEM – NeuRecall

The aim of the project is to develop an explicit computational framework for the NeuRecall tutoring system. This will be accomplished through incorporation of additional data about each learner as well as the broader population of learners. We also adopt a model-based, sequential decision-making approach to predicting learner behaviour. Using these parameters, it should be able to predict the answers that learners will give to the questions more accurately. Getting better answer predictions would help the software effectively tweak the learning path of the users according to their skill-sets and memories.

Project Team
Vishal Raheja, MScAC Student
Simon Overduin, NeuRecall
Professor **Sheila McIlraith**

PERSONA-TAILORED USER EXPERIENCE FOR ENTERPRISE SOFTWARE – Riva Modeling Systems Inc.

The objective of this project was to research a personas-tailored user experience for enterprise software. The main approach in the project was to study the personas that use such software and analyze the user experience challenges. The study showed that different personas use different sets of functionalities and features to accomplish their tasks. Also, it revealed that each persona looks at the data

from their individual perspectives. Those findings led the project to have two main directions. One direction focused on the UI navigation methods. Different prototypes were developed based on a new persona focused system structure. The other direction was to implement a dashboard strategy that presented different personas with their respective information in the most intuitive and meaningful form. The dashboard introduced new features such as resizable and rearrangeable widgets, multiple board views, and real time widget additions. Furthermore, the dashboard incorporates new responsive, interactive, and customizable charts, visualizing user's data. Both directions' prototypes were tested by end users and results showed that a personas-tailored user experience elevated the overall user experience in enterprise software.

Project Team
Alaa Abdulaal, MScAC Student
Kiran Sachdev, Riva Modeling Systems Inc.
Professors **Eyal de Lara** and **Daniel Wigdor**

## RE-ARCHITECTURE AND DATA MIGRATION OF SIDEFX.COM – Side Effects Software Inc.
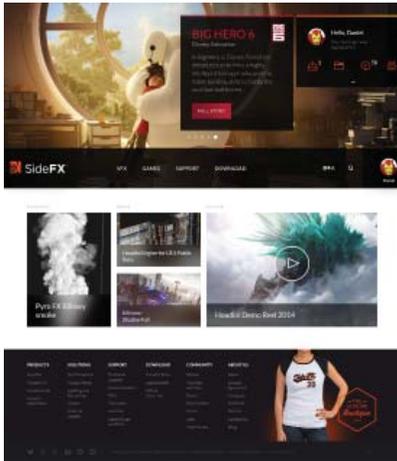
www.sidefx.com is the corporate website for Side Effects Software Inc. and its flagship product Houdini. This project researches the best technologies that will allow for both easy integration into the company's existing technologies and standards, as well as build a strong backend for a new corporate website. Re-architecting Sidefx's website and integrating it with their online store, digital asset store, and software licensing system, requires a solid content management system (CMS). The project explores current technologies used to build such systems in Django, a python web framework. Proof-of-concept methodologies were used to determine which frameworks allowed for a seamless integration and data migration with the current system. The systems were also compared based on how they handle architectural designs found in current research, in particular, the three-tier architectural approach. The goal was to develop an efficient, user-friendly, corporate website, accessible via mobile and desktop devices. In conjunction, a stable, reliable, and scalable CMS was to be designed and implemented. The CMS technology selected for further development from the proof-of-concept systems was Mezzanine. In addition, this project explores novel ways to visualize user discussions and commenting, to build a strong community component to the website. Traditional forums, simple flat forums, knowledge-base forums, etc. were examined based on user feedback. The forums backend had to integrate in with Mezzanine, and be customized to fit Sidefx's brand and new community outlook. If successful, the product of this project's research and development will build a modern and powerful new system that will be flexible, scalable, and easy to maintain.

Project Team
Pirave Eahalaivan, MScAC Student
Luke Moore, Side Effects Software Inc.
Professor **Alan Rosselet**

SKETCHCRAFT – The Jonah Group

SketchCraft is an experimental project that aims to accelerate sketching for ideation in various design contexts, while facilitating visualization in both two and three dimensions. To achieve this goal, SketchCraft looks to create a paradigm for ideating rapidly using a digital stylus that is analogous to the fluid "pencil and paper" sketching experience, removing the tedium and rigidity of typical 3D-modelling software. In developing this new fashion of sketching, a wide variety of stroke types and contexts are to be handled; a set of stroke behaviors are defined to help deal with this miscellany of strokes. These behaviors interpret the incoming strokes and transform them accordingly. For instance, the orthographic behavior takes a stroke that was drawn and projects it perpendicular to the closest stroke it was drawn to. The intention for these behaviors is to design a set of rules that are able to understand what the user is sketching, then intelligently process and alter the input to correspond with the user's vision. Overall, SketchCraft aspires to combine these features with ongoing experimental methods to allow a unique sketching experience to come to life.
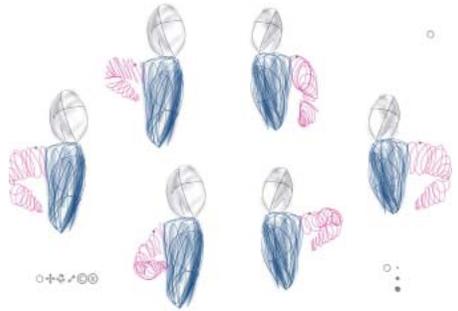
Project Team
Stacey Oue, MScAC Student
Chris Chan, The Jonah Group
Professor **Karan Singh**

SYNCING DATA BETWEEN ONLINE AND OFFLINE FRESHBOOKS APPLICATIONS – FreshBooks

Product teams frequently face the problem of syncing data between multiple platforms; web and mobile. Syncing when all devices are online is non-trivial by itself. It becomes more difficult when devices go offline, but continue to produce data. The first challenge is in ensuring that the offline data propagates to all devices in the user's ecosystem. The next challenge deals with managing conflicts while merging offline data with changes from other devices.
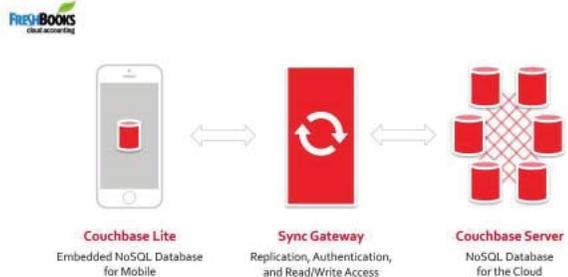
At FreshBooks, our iPhone and iPad users' data gets fully synced every time the app is opened and online. This approach proved to be time-consuming and inefficient. We also restricted offline editing due to the lack of a good sync strategy. The purpose of this project was to create a prototype for a "new sync strategy".

We first determined the desired features of our prototype. First, data had to be synced only when changes occurred. Second, changes had to be propagated as soon as a device switched

state from offline to online. Finally, applying changes and resolving version conflicts needed to be seamless. After researching and evaluating the pros and cons of several options we created a prototype using "CouchBase for Mobile". It satisfied our criteria and turned out to be a popular industry choice as a data sync solution. We also took the opportunity to use the recently released Swift language to create the prototype. The prototype was a big success and revealed the work required for a full-system integration. It also showed us some of the challenges that were inherent to integrating third-party services with FreshBooks' mobile stack.

Project Team
Harsha Balasubramanian, MScAC Student
Ryan Henry, FreshBooks
Professor **Anthony Bonner**

SYNTAX-BASED LINKING OF ASPECT-MENTION PAIRS FOR SENTIMENT ANALYSIS –
Verticalscope Inc.

An online discussion forum is an example of social media that contains vast amounts of unstructured domain-specific information. Being able to extract opinions about different features of the domain-specific products allows better monitoring of product reputation and understanding of the needs of the market. Our generalized task is called aspect-based sentiment analysis (ABSA) (Hu and Liu, 2004), which mines opinions from unstructured text containing specific mentions (a reference to a product or set of products) and their aspects (denote both parts and attributes). To achieve the goal, we split the task into four main subtasks, which are (1) aspect extraction, (2) mention extraction, (3) aspect mention linking, and (4) sentiment analysis. We design and implement a pipeline containing existing methods, such as the named entity extraction, from Stanford's Conditional Random Field (CRF) sequence models (Lafferty, McCallum, and Pereira, 2001) and the sentiment prediction, from Stanford's Recursive Neural Tensor Network (Socher, Richard, 2013). For the aspect-mention linking subtask, we propose a new rule-based linking strategy based on the syntactic structure of the parse tree. We discover that directly matching the aspect with the mention at the sentence level results in more precise sentiment than sentence-based sentiment and document-based sentiment. The system is built on Spark which is scalable to a large amount of data. This aspect-based sentiment analysis system can help Verticalscope Inc. to gain market competitiveness and we hope it will be of interest to other workshop participants.

Project Team
Wenjie Zi, MScAC Student
Afsaneh Fazly and **En-Shiun Annie Lee**, Verticalscope Inc.
Professor **Graeme Hirst**

ABOUT THE PROGRAM

The Master of Science in Applied Computing (MScAC) program is designed to educate the next generation of technical leaders, innovators and entrepreneurs by turning research into practical industry-based applications. Students in this program spend the first eight months studying with some of Canada's leading computer scientists and another eight months in an internship, where they will apply their research to real-world problems.

This brochure highlights the 2014-2015 faculty and industry supervised student projects.

ABOUT THE DEPARTMENT OF COMPUTER SCIENCE

Founded in 1964, the Department of Computer Science at the University of Toronto is home to 1,700 undergraduates, more than 260 graduate students and 63 faculty members. Through its undergrauate, graduate and professional graduate programs, 7,500 active alumni are making further contributions to the field within industry and research.

CONTACT US

Eugene Fiume, Director
MScAC program
elf@cs.toronto.edu
Tel. 416.978.5472

Matt Medland, Associate Director
MScAC program
matt@cs.toronto.edu
Tel. 416.946.5292

www.cs.toronto.edu/mscac

DEPARTMENT OF COMPUTER SCIENCE

University of Toronto
Master of Science in Applied Computing
40 St. George Street
Toronto, ON M5S 2E4
www.cs.toronto.edu