Data Science and Predictive Analytics
Academic-Industry Partnering Forum

Stefan Steiner
Department Chair
Statistics and Actuarial Science
Friday, April 27, 2018
Partnering Forum Goals

- Stimulate contact and interaction between companies and academic researchers

- What does longer term success look like?
  - Establish industry research collaborations for faculty and graduate students
  - Provide funding for graduate students
  - Address problems facing industry
  - Develop a talent pipeline for companies
Department of Statistics and Actuarial Science

- 50 research professors and 10 lecturers
- 900+ undergrad and close to 200 grad students
- Research Institute/Groups
  - Waterloo Research Institute in Insurance, Securities and Quantitative Finance (WatRISQ)
  - Business and Industrial Statistics Research Group (BISRG)
  - Propel Centre for Population Health Impact (PROPEL)
  - Centre for Computational Mathematics in Industry and Commerce (CCMIC)
  - Survey Research Centre (SRC)
- Statistical Consulting and Collaborative Research Unit
Department Research Areas

- Actuarial risk management
- Applied probability
- Biostatistics
- Business and industrial statistics
- Computational statistics
- Data science
- Econometrics and quantitative finance
- Risk theory
- Statistical modeling and inference
- Survey methods
Forum Agenda and Logistics

- Morning (9-12:30) – DC 1302 (break at 10:45 in fishbowl)
  - Introduction and successful collaboration showcase
  - Company and faculty member profiles (5 minutes each)
- Lunch (12:30-1:30) – M3 Atrium
  - Presentations by OCE, NSERC and Mitacs regarding funding opportunities
- Afternoon (1:30-3:30) – M3 Atrium
  - Open networking (first 30 minutes)
  - Speed networking (starting at 2pm, scheduled 5 x 15 minute time slots)
  - Closing remarks
Future Engagement/Funding Opportunities

- Collaborative Research
  - OCE, NSERC and Mitacs programs (more on this over lunch)
- Engage with the Statistical Consulting and Collaborative Research Unit
- Hire co-op/internship students
  - Undergraduate students in a large number of programs, e.g. Actuarial Science, Biostatistics, Computer Science, Data Science, Statistics,
  - Graduate students and Post-Doctoral Fellows
- Waterloo-ASA DataFest, May 4-6, 2018 (annual event)
- Actuarial science case competitions
Partnering with UW

The Story

Ella Hilal, PhD.
About the Speaker
Director of Data Science and Engineering
Director of Innovation and Data Intelligence
The Story
The Connected Car
Diverse Data Sources

SPECTRUM OF IN-VEHICLE TECHNOLOGY ENABLERS

- Mobile Only
- Bluetooth® Enabled Assist
- 12V Connector
- Beacon
- OBD Bluetooth
- Full OBD with Cellular
- Pro Install
- OEM

Driver-centric

Vehicle-centric
Scale of Data

Speed of Data Arrival

Volume

Variety

Big Data

Velocity

Veracity

Different Data Forms

Different Data Accuracy
Making Sense of

10 Trillion Data Points
Crafting the Lifestyle Narrative

- Driver Behavior
- Habits
- Consumer Preferences
Crafting the Lifestyle Narrative

Driver Behavior
Habits
Consumer Preferences
Crafting the Lifestyle Narrative

Risk Analysis

Driver Behavior

Habits

Consumer Preferences
Ask the Experts
Great Partners
Working Closely
Major Strides in a **Challenging** Problem Space
Major Strides in a Challenging Problem Space

1. Problem Statement
Major Strides in a Challenging Problem Space

1. Problem Statement
2. Data Assets
Major Strides in a Challenging Problem Space

1. Problem Statement
2. Data Assets
3. Solution with Real-world Constraints
Major Strides in a Challenging Problem Space

1. Problem Statement
2. Data Assets
3. Solution with Real-world Constraints
4. Knowledge Transfer
Success is Sweet
The Story Continues
Shopify is the leading cloud-based, multichannel commerce platform.
Shopify is the leading cloud-based, multichannel commerce platform.

Merchants can use the software to design, set up and manage their stores.
Shopify is the leading cloud-based, multichannel commerce platform.

Merchants can use the software to design, set up and manage their stores.

The Shopify platform was engineered for reliability and scale.
Shopify is the leading cloud-based, multichannel commerce platform.

Merchants can use the software to design, set up and manage their stores.

The Shopify platform was engineered for reliability and scale.

Shopify currently powers over 500,000 businesses in ~150 countries.
Shopify is the leading cloud-based, multichannel commerce platform.

Merchants can use the software to design, set up and manage their stores.

The Shopify platform was engineered for reliability and scale.

Shopify currently powers over 500,000 businesses in ~150 countries.

Red Bull, LA Lakers, the New York Stock Exchange, GoldieBlox, and many more.
Detection of Check-out Bots
Detection of Flash Sales
Marketing Campaigns

- Potential of Engagement
- Risk of Un-subscription

Photo by: Nicole De Khors
Collaboration is Essential for Advancement & Innovation

@ ella.hilal@Shopify.com
https://ca.linkedin.com/in/allaahilal
https://uwaterloo.ca/scholar/ahilal
/a_hilal
``The right treatment to the right patient at the right time."
Treating the patient, not the diagnosis

Heterogeneity between patients:

0 months 3 months

Alice

Bob

Drug A

Drug A
Treating the patient, not the diagnosis

Heterogeneity between patients:

<table>
<thead>
<tr>
<th></th>
<th>0 months</th>
<th>3 months</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>sad</td>
<td>happy</td>
</tr>
<tr>
<td></td>
<td>Drug A</td>
<td></td>
</tr>
<tr>
<td>Bob</td>
<td>sad</td>
<td>happy</td>
</tr>
<tr>
<td></td>
<td>Drug B</td>
<td></td>
</tr>
</tbody>
</table>
Treating the patient, not the diagnosis

Heterogeneity between patients:

- **Alice**
  - 0 months: Sad face
  - 3 months: Happy face
  - Drug A

- **Bob**
  - 0 months: Sad face
  - 3 months: Happy face
  - Drug B

Heterogeneity within patients:

- **Charles**
  - 0 months: Sad face
  - 3 months: Happy face
  - 6 months: Sad face
  - Drug A
Treating the patient, not the diagnosis

Heterogeneity between patients:

- **Alice**: 0 months - Frowny face, 3 months - Happy face
  - Drug A

- **Bob**: 0 months - Frowny face, 3 months - Happy face
  - Drug B

Heterogeneity within patients:

- **Charles**: 0 months - Frowny face, 3 months - Happy face, 6 months - Frowny face
  - Drug A

- **Charles**: 0 months - Frowny face, 3 months - Happy face, 6 months - Happy face
  - Drug A
  - Drug B
Dynamic treatment regimes (DTRs) ‘formalize’ the process of personalized medicine:

“If patient over 65 prescribe Drug A, otherwise Drug B.”
Dynamic treatment regimes (DTRs) ‘formalize’ the process of personalized medicine:

“If patient over 65 prescribe Drug A, otherwise Drug B.”

Advantages:
- Improved results over standard ‘one size fits all’ approaches.
- Can account for delayed effects of interventions.
- Can learn from observational data.
Personalized Medicine

```
```
```
The right treatment to the right patient at the right time."
```

But: applications not limited to health settings!
Applications possible wherever (sequences of) decisions are made over time to optimize some outcome.
Where next?

Data

New problems

Real-world challenges

michael.wallace@uwaterloo.ca  mpwallace.github.io
Wayne Oldford

- Statistical reasoning
- exploratory data analysis
- data visualization
- development of interactive computational environments that support these activities
Survey Methodology

There is designed data collection and organic data collection (Groves, 2011)

Survey methodology research is about:

• the principles of designed data collection
• the combination of designed data and organic data
• in analysis, accounting for the complexity of the design (e.g. targeted sampling; network sampling; longitudinality)
Expertise at UW

• Involvement with large longitudinal surveys:
  – International Tobacco Control Project
  – Canadian Longitudinal Study on Aging

• Analysis of survey data with large numbers of variables; predictive model selection (Wu, Boudreau)

• Machine learning from text data (Schonlau)

• Network sampling (Thompson)
Survey Research Centre

• A full-service survey research organization providing survey design, data collection and top-line analysis since 1999

• Emphasis on collecting high-quality data for scientific and decision-maker use

• Web, telephone, mail and mixed-mode surveys

• Data held on secure servers at the University of Waterloo
STEFAN STEINER
RESEARCH PROFILE
ASSESSMENT OF STREAMING DATA

• Decision support with process monitoring and comparison
  • Monitoring manufacturing processes for upsets
  • Monitoring customer satisfaction measures
  • Comparing medical labs, hospitals, or individual surgeons over time with risk adjustment

• Example application: analysis of automotive telematics data
  • Building accident risk models to identify risky behaviour profiles
  • Developing driver behaviour profiles
  • Providing real-time feedback on driving behaviour
PROCESS IMPROVEMENT AND VARIATION REDUCTION

• Measurement system assessment
  • Develop improved plans incorporating baseline, gold standard assessment, partial verification, specially selected parts, etc.
  • Variety of characteristic types: continuous, binary, diagnostic tests, count, functional, etc.
  • Comparison and calibration of measurement systems – probability of agreement

• Quality/process improvement systems

• Experimental design
Welcome!

I am an Associate Professor of Actuarial Science in the Department of Statistics and Actuarial Science at the University of Waterloo.

I am interested in not only academic research in those areas including probability, statistics, actuarial science, finance and optimization, but also industry consulting projects and commercial adventure in relevant industries.

Office:
Room 3136
M3-200 University Avenue West,
Waterloo, ON, N2L 3G1, Canada.

Tel:
519-888-4567 ext. 31132

Email:
chengguo.weng@uwaterloo.ca

Website:
www.chengguoweng.com

Social links:

- Broad research interests
- Enthusiasm in industrial partnership
- Diversity of Education Experiences

Chengguo.Weng.com
Optimal Decision with Uncertainty

- Optimal reinsurance
- Vast portfolio selection
- Risk prioritization
- Pricing and hedging of insurance/finance products

- Monte Carlo simulation
- Data-driven
- Partial information
- Statistical learning

Predictive Analytics

- Insurance premium rating
- Customers behavior characterization
- Prediction of economic factors

- Personalized prediction algorithms
- Enable price discrimination
- Enable incorporation of large information
- Describe me your **Situations**

- Bring me your **Questions**

- Show me your **Data**
Shoja Chenouri
schenouri@uwaterloo.ca
Department of Statistics and Actuarial Science

Acknowledgment: My research has been funded by
Neural spiking:
- Action potentials (spikes) are nerve impulses
- Spike trains
- Temporal point process. Inference on firing rates

Network data:
- Epidemic networks (directed graphs)
- Predicting links based on covariates on the nodes
- Temporal dynamics of networks

AIS Data:
- Modelling trajectories of vessels
- Identifying dark targets (anomalous vessels)
- Doppelgangers
- Spatiotemporal processes, functional data

Additive manufacturing or 3D printing:
- Design of experiments
- Response surface methodology & optimization
- Deformation and compensation
- Process monitoring
Freight train accidents carrying HazMat

- Modelling probability of a car initiating derailment
- Number of cars derailed
- Data-driven marshalling yard
- Markov chains, GLM and classification algorithms.

Environmental contaminants

- How the implemented policies by Canada and US for acid rain worked? Multivariate Change point detection.
- Below detection data for site characterization and remediation.
- Analysis of left censored data and regression.

Dimensionality reduction

- Represent the high-dimensional data in a low-dimensional form without losing “important information”
- Widely applied to many types of data such as images, videos, texts.
- Many datasets are high-dimensional representations of data from low-dimensional curved manifolds.