Syllabus for
Deriving Knowledge from Data at Scale (141031/141034)
Data Science (4693 classroom /4694 online)

Seattle classroom (Puget Plaza 401) and online
10/5/15 -12/7/15
Mondays 6:00 PM – 9:00 PM

Instructor: Winson Taam

Email: maatnosniw@gmail.com
Office Hours: 5-6pm W, F

Course Description:

This course is designed to pull together what has been learned so far about the structuring and manipulation of data and core statistical and machine learning techniques and add knowledge on the machinery used to leverage those techniques in real-world scenarios in which data scientists are asked to generate knowledge. Students will be asked to apply the course content to real-life scenarios and think creatively as well as critically through issues. The expectation is that by the end of the class, students will be able to attack data sciences questions impacting their business; establish robust experimental tests of data-driven hypotheses, generate meaningful and reliable findings and communicate them clearly. Information will also be provided during the quarter for students to read further on topics if desired.

Course Learning Objectives:
"At the conclusion of this course, the student will know and/or be able to do…"

- Develop an analytics plan to answer a specific business question, using inferential statistics.
- Understand and can select between different modeling approaches for different problem types – classification, regression, clustering, dimensionality reduction/data compression, topic modeling.
- Implement experimental designs and build models.
- Understand over- and under-fitting, and can make intellectual and mathematical trade-offs using Occam's razor (eg. reformulating maximum likelihood to include BIC or AIC), use of evaluation metrics such ROC curves, confusion matrices to improve model performance.

Course Format:
Each class will include a discussion on the practice of data science, analytic techniques and machine learning to extract insight from data, and hands on modeling. There will be regular reading and data science assignments.

Course Materials:
No text book is required. All reading material will be freely downloadable. The following texts are not required but may be helpful:

Mining of Massive Datasets by Anand Rajaraman and Jeffrey David Ullman, Cambridge University Press 2011 (ISBN-13: 9781107015357); Can be downloaded from: http://infolab.stanford.edu/~ullman/mmds.html

Elements of Statistical Learning: Data mining, Inference, and Prediction by Hastie, Tibshirani and Friedman (2009); it can be download from: http://statweb.stanford.edu/~tibs/ElemStatLearn/

Machine Learning for Hackers by Conway and White (ISBN-13: 978-1-449-30371-6); Can be downloaded from: http://en.tjcities.com/wp-content/uploads/Books/Machine_Learning_for_Hackers.pdf

ROC Curve, Lift Chart and Calibration Plot by Vuk and Curk; Can be downloaded from: (http://mrvar.fdv.uni-lj.si/pub/mz/mz3.1/vuk.pdf)

Technical Requirements:

Access to a personal computer or laptop, suitable to install and run R Studio, a free open source development environment for R available from http://www.rstudio.com/ide/, and either WEKA http://www.cs.waikato.ac.nz/ml/weka/index.html or KNIME http://www.knime.org/, both are open source data processing, analysis, and exploration platforms.

Program Webpage:
Moodle site

Student Assessment: Students must attend 8 of the 10 sessions, satisfactorily complete all weekly assignments, and finish the class capstone project.

Course Topics and Assignments by Date:

	Date	Topic
Week 1		○ Syllabus, expectations, software we will use in the course, housekeeping; ○ Check-in from first two quarters - opportunity to ask lingering questions; ○ Overview of the Data Science process -problem definition to model evaluation and tuning; ○ Motivation & applications of machine learning, definition of machine learning ○ Overview of supervised and unsupervised learning, select practical applications to illustrate value.
Week 2		○ **Algorithms:** Decision Trees and Decision Forests, introduction to ensembling; ○ **Techniques:** Introduction to WEKA, data processing and analysis in WEKA; ○ **Practical Application:** Discuss decision tree data science lab for the week; ○ **Theory:** Handout select paper for reading during the week.
Week 3		○ **Algorithms:** Clustering for exploratory data analysis, K-means & EM. ○ **Techniques:** Data Science hands-on in class, evaluating model performance from ROC Curves, Confusion Matrices, Lift, Precision, Recall. Continue discussion model ensembles; ○ **Practical Application:** Discuss random forest data science lab for the week; ○ **Theory:** Handout select paper for reading during the week.
Week 4		○ **Algorithms:** Time Series Forecasting, ARIMA, Holt-Winter, and ETS. ○ **Techniques:** Motivation & applications of machine learning, definitions of machine learning, overview of supervised and unsupervised learning, select practical applications to illustrate value; ○ **Practical Application:** Discuss forecasting data science lab for the week; ○ **Theory:** Handout select paper for reading during the week.
Week 5		○ **Algorithms:** Linear regression, logistic regression, lasso, ridge regression; ○ **Techniques:** Data wrangling, from analysis, cleaning, dealing with missing values, understanding the data; ○ **Practical Application:** Discuss data cleaning data science lab for the week, along with linear regression data science lab; ○ **Theory:** Handout select paper for reading during the week;
Week 6		○ **Techniques**: Basic machine learning technologies applied to recommendation & personalization e.g. Naive Bayes, Matchbox, market-basket analysis and association rules. ○ **Industry Example**: Building Recommendation Models for Retail ○ **Practical Application**: Discuss optional data science lab for the week **Theory**: Handout select paper for reading during the week
Week 7		○ **Algorithms:** Support vector machines (SVM) and the kernel trick; ○ **Techniques:** Topic modeling as dimensionality reduction, feature selection and creation (lecture 1/2); ○ **Industry Example:** Winning the Netflix Prize; ○ **Practical Application:** Introduce capstone project, mandatory. ○ **Theory:** Handout select paper for reading during the week;
Week 8		○ **Algorithms:** Neural networks, overview of deep learning and deep networks; ○ **Techniques:** Feature selection and creation ○ Discuss a Kaggle contest entry from start to finish; ○ **Industry Example:** Guest lectures from practicing data scientists;
Week 9		○ **Algorithms:** Review and guidelines on rapid experimentation with different models, fine tuning specific models, ○ **Techniques:** Hands-on in class ensemble of models, boosting and bagging;

		○ **Industry Example:** Nonlinear regression in predicting traffic flow**;** ○ **Practical Application:** Discuss progress on capstone project**.**
Week 10		Close the course, reviewing results from capstone project, bring in previous graduates and or data scientists from Microsoft, Amazon, Google, etc. so that they can talk about their experiences.