**Co-op Highlights - Ken**

**Environment Canada (Jan – Sept, 2011):**

1. Modification of new data. Examples include merging, formatting date stamps, partitioning dates, vectorization, and quality control.
2. Created new randomization routine which takes whole/half weeks rather than per hour. Also, includes quality check where a sample is selected from each pre-determined quantile interval of the data set.
3. Predictor selection routine which uses forward and hybrid stepwise selection with BIC criterion to select optimal predictors for linear models.
4. Built multiple linear regression models using optimal predictors for each data set. Also predicted new forecasts using new data brought in.
5. Heat maps to determine sparsity in various columns.
6. Ran neural network and support vector machine predictive models to forecast new predictions. (Models already built)
7. Boosted Regression model presentation. Built models and predicted new forecasts and compared results with in-house methods. Optimization of parameters on number of regression trees, shrinkage, and interaction depth.
8. Thorough interaction with supervisor to implement statistical techniques, and provide statistical advice.
9. Interaction with additional co-op student in atmospheric science to produce verification plots.

**BC Genome Sciences Centre (May – Sept, 2012):**

1. Mutational relation between genes in patients. Fisher’s exact test, phi-coefficient, proportional tests, frequency heatmaps, k-means clustering.
2. Data management and parsing of bio-informatics output of gene mutation data.
3. Group comparison techniques on miRNA/mRNA expression level data on individual or groups of genes. Kolmogorov-Smirnov tests, Wilcoxon tests, t tests. Kernel Density, Box plots, empirical cumulative distribution, and quantile-quantile comparison plots.
4. Survival analysis on various cancer types: Kidney, LAML, LUAD. Kaplan Meier Survival functions and Univariate Cox Proportional Hazards Regression in testing and modeling survival outcomes for different quantiles of expression levels: high/med/low. Followed by multivariate cox, boosted cox, and penalized cox models.
5. Multiple Testing corrections.
6. Work above is from a combination of many post-docs including, Gordon, Jeremy, Linda, Gerben, Fang, Martin, Suzanne of the Karsan lab.

**Environment Canada (Feb – Aug, 2013):**

1. Implemented data imputation methods, in particular the k-nn approach and the interpolation approach to time-series data.
2. Optimized boosted regression tree models for prediction modelling of air quality data. In particular, implemented optimal routines for cross-validation to pick out main parameters of number of trees, tree depth, learning rate, and tree split observation size.
3. Data wrangling to turn data into usable form for analysis and verification. For example, merging, filtering, converting data types, time-stamps, etc.
4. Used R and SQLite for data management. Turned flat files into database format. I used standard relational database type design models. Implemented scripts for creating tables, sql commands to fetch data, setting up drivers, etc.
5. Tested Ridge Regression, Support Vector Machines, box-cox transformation for non-normal data, and principle component analysis for dimensionality reduction.
6. Improved coding style by using plyr, and other R libraries for data formatting.
7. Optimized speed of computation of procedures by switching to parallel computation. In particular, I used foreach and multicore, and plyr enables parallel computing when multicore is set up.
8. Throughout the work terms, I’ve discussed methodologies between my supervisor and I. I provided statistical advice (for example, good statistical practice) whenever possible.