**STUDENT SPEAKERS:**

**Chunyi Wang,** Ph.D. Candidate, Department of Statistics
*Approximating the Likelihood for the Hyper-parameters in Gaussian Process Regression*

In the MCMC with temporary mapping and caching scheme, we need a distribution that is similar to the distribution we would to sample from while easier to compute. For the typical application of Gaussian Process Regression problems, there are quite a few approximation methods in the literature, but most of them are proposed for prediction and inference purposes. I'd like to give a brief review of some of these methods, with emphasis on their applicability of the MCMC with mapping-caching scheme.

**Ilya Sutskever,** Ph.D. Candidate, Department of Computer Science
*Generating Text with Recurrent Neural Networks*

Recurrent Neural Networks (RNNs) are very powerful sequence models that do not enjoy widespread use because it is extremely difficult to train them properly. Fortunately, recent advances in Hessian-free optimization have been able to overcome the difficulties associated with training RNNs, making it possible to apply them successfully to challenging sequence problems. In this paper we demonstrate the power of RNNs trained with the new Hessian-Free optimizer (HF) by applying them to character-level language modeling tasks. The standard RNN architecture, while effective, is not ideally suited for such tasks, so we introduce a new RNN variant that uses multiplicative (or "gated") connections which allow the current input character to determine the transition matrix from one hidden state vector to the next. After training the multiplicative RNN with the HF optimizer for five days on 8 high-end Graphics Processing Units, we were able to surpass the performance of the best previous single method for character level language modeling – a hierarchical non-parametric sequence model. To our knowledge this represents the largest recurrent neural network application to date.
(Joint work with James Martens and Geoffrey Hinton).

**Alexander Shestopaloff,** Ph.D. Candidate, Department of Statistics
*Markov Chain Monte Carlo with Constellations of Points*

Constellation MCMC methods are an example of ensemble MCMC methods introduced by Neal (2010), and are based on the notion that, in certain models, once we have computed the likelihood at a single point, we can get the likelihood at certain other points at little additional computational cost. I will discuss some aspects of constellation MCMC methods, and demonstrate their application to Bayesian classification and regression.

**Billy Chang,** Ph.D. Candidate, Department of Public Health Sciences
*Regularization for Nonlinear Dimension Reduction by Subspace Constraint*

Sparked by the introduction of Isomap and Locally-Linear-Embedding in year 2000, nonlinear approaches to dimension reduction have received unprecedented attention during the past decade. Although the flexibility of such methods has provided scientists powerful ways for feature extraction and visualization, their applications are focused mainly on large-sample and low-noise settings. In small-sample, high-noise settings, model regularization is necessary to avoid over-fitting. Yet, over-fitting issues for nonlinear dimension reduction have not been widely explored, even for earlier methods such as kernel PCA and multi-dimensional scaling.

Regularization for nonlinear dimension reduction is a non-trivial task; while an overly-complex model will over-fit, an overly-simple model cannot detect highly nonlinear signals. To overcome this problem, I propose performing nonlinear dimension reduction within a lower-dimensional subspace. As such, one can increase the model complexity for nonlinear pattern search, while over-fitting is avoided as the model is not allowed to traverse through all possible dimensions. The crux of the problem lies in finding the subspace containing the nonlinear signal, and I will discuss a Kernel PCA approach for the subspace search, and a principal curve approach for nonlinear basis construction.

## POSTER PRESENTATIONS:

**Anita Xia**
*Adaptive population divergence and frost tolerance in Ipomoea hederacea*

This project examines the relationship between seed, leaf, and flowering traits of Ipomoea hederacea. Seeds of 20 distinct populations of Ipomoea hederacea were collected from the United States and planted in a greenhouse. Eight physical traits were measured as they grew. There is evidence that the anther stigma distance and the percent of anthers above the stigma are correlated. There is strong evidence that the first leaf growth rate and all the flowering traits vary by population.

**Yidan Zhang & Jing Shi**
*Study Shows No Variation in the Growth Rates of Elliptio Complanata in a Large Canadian Shield Lake*

This project considers whether the growth rates of juvenile freshwater mussels have a variation across the 6 different sites, or are affected by substrate characteristics. By applying the log model to achieve the growth rates, there is only weak evidence of variation in the mussels' growth rates from different sites. There is also some weak evidence showing that the mussels' growth rates are affected by the OM factor, but no evidence showing that the growth rate varies from different Phi values.

**Jon Balon**
*Sexual Dimorphism in a Wind-Pollinated Annual Herb: The Role of Allocation Trade-offs and the Costs of Reproduction*

There are numerous factors which have an effect on resource allocation to reproductive, root, and vegetative biomasses and how they influence Sexual Size Dimorphism during the life-cycle of a certain wind-pollinated plant. It was determined that each of the proportions of biomasses can be explained by many factors, such as nitrogen level, light level, gender, and grandmother. I also consider the difference between the sexes in the costs of reproduction. By comparing the costs of reproduction of male and female plants, there is evidence that male plants have higher costs of reproduction compared to female plants.

**Craig Burkett**
*Saving the World's Fish: One Logistic Model at a Time*

This report will show that it is possible to predict whether or not a species of fish is likely to be listed as endangered in the near future, using information currently available on that species. This study looked at several factors including the maximum length, intrinsic vulnerability, and trophic level of a species of fish. There is strong evidence that the maximum length of the fish is the most useful predictor of endangered status.

**Matt Asher & Jisun Kim**
*Groundhog Day for Prediction Model Development*

Traditional data analysis looks to find a single model capable of explaining as much variation as possible without overfitting. In the context of dynamic, highly noisy data, where the best one can do is identify itinerant, uninterperatable patterns, a different approach is called for. For this project we created a model-agnostic framework to evolve solutions to the problem of asset price prediction in complex markets. After training the system on price data from 2002 to 2006, tests of ensemble predictions for the next 5 years generated compound annual returns of 26%.

**Jerry Smith**
*Seeing Orange: Prawns Exploit a Sensory Bias of the Trinidadian Guppy*

This project was completed for STA490. Trinidadian guppies were exposed to model prawns with three different spot colours on their pincers. It is known that these guppies have a sensory bias for the colour orange. One hypothesis was that the guppies would spend more time near the prawns with orange spots. Another hypothesis was that the species of guppies that co-exist with the predatory prawns would spend less time near the prawns than other species of guppies. A survival analysis was carried out using the proportional hazards model. It was concluded that there is strong evidence to support both the hypotheses.

## KEYNOTE SPEAKERS:

**Joanna Mills Flemming**
Assistant Professor, Department of Mathematics and Statistics, Dalhousie University

*Challenges in marine statistical ecology - from sea cucumbers to grey seals*
Statisticians are making meaningful contributions to important projects in ecology and the environmental sciences. Three such projects, each of which is computer intensive with regard to the amount of data required for analysis and/or the methodology itself, are described. The first involves estimation of critical population dynamics in the Hudson Estuary using multivariate state space models to assess temporal dynamics in the annual reproductive rates of marine fish. The second presents the first comprehensive analysis of invertebrate fisheries based primarily on a global catch database and includes new methodologies to quantify temporal and spatial trends in fishery development. Finally, the third project evaluates encounter data on marine species collected from roving acoustic receivers (bioprobes) that are part of the Ocean Tracking Network (OTN), a global project that aims to provide a permanent platform to monitor the movements and interactions of marine species.

**Diane Lambert**
Research Scientist, Google

*Statistics at Google Scale*
From the outside, a search engine looks like a statistical analysis system that marries data and computing. On the inside, Google is even closer to a statistical analysis system because it relies on continuous measurement, estimation, experimentation and learning. This talk will show how Google uses well-established statistical principles, along with huge amounts of data and computing, to improve search and ads for users, advertisers and publishers.

**Robert Gentleman**
Senior Director, Bioinformatics and Computational Biology, Genentech

*Computationally Intensive Biology Problems*
I will describe three different computational problems that are relevant to biostatistics and computational biology:
1) The analysis of flow cytometry data, where a very large number of cells are measured (in the millions) on relatively few measures (typically 10 or fewer).
2) The analysis of genomic data. Here the computational efforts are typically associated with large data volumes, complexity of mapping reads to genomic coordinates etc. But there are also many statistical problems that arise.