NEW YORK WORKSHOP ON COMPUTER, EARTH & SPACE SCIENCES 2009

AGENDA & ABSTRACTS

Agenda for New York Workshop on Computer, Earth, and Space Sciences

Friday, February 06, 2009.

Time	Speaker	Title
9:00-9:30	Everyone	Viewpoints Demo, Coffee and Setup
9:30-10:00	Michael Way (NASA/GISS)	Workshop Introduction and Talk on Gaussian Process Regression
10:00-10:20	Liu Jingchen (Columbia - Stat)	Efficient Simulation For Tail Probabilities Of Gaussian Random Fields
10:20-10:40	Michael Blanton (NYU - Physics)	Interpretation of the images of galaxies
10:40-11:00	Mike Bauer (NASA/GISS)	Why Weather Matters to Climate Models
11:00-11:20	Zoltan Haiman (Columbia - Astro)	Can You Spot Dark Energy in a Map With 100 Million Pixels?
11:20-12:10	Kevin Knuth - **Invited Speaker**	Automating the Scientific Method
12:10-12:50	**LUNCH BREAK**	with robot demo by Kevin Knuth
12:50-1:40	Alexander Gray - **Invited Speaker**	How to do Machine Learning on Massive Astronomical Datasets
1:40-2:00	Sunil Bhaskaran (Lehman College)	Integrating Hyperspectral Data and GIS data
2:00-2:20	Jo Bovy (NYU - Physics)	Inferring the velocity distribution of nearby stars from Hipparcos data
2:20-2:40	Arlin Crotts(Columbia - Astro)	Image Subtraction in Astrophysics
2:40-3:00	Ji Meng Loh (Columbia - Stat)	The Marked Point Bootstrap
3:00-3:20	Caleb Scharf (Columbia - Astrolab)	Challenges in understanding the formation of planetary systems
3:20-3:40	Benjamin Herman (CCNY)	Stochastic Monte Carlo methods for non-linear statistical inverse problems
3:40-4:00	David Hogg (NYU - Physics)	Automated calibration, or generating and trusting astronomical meta-data
4:00-4:20	Jianting Zhang (City College - Comp Sci)	Correlations btwn Beta Diversity & Productivity Variations using NASA MODIS

4:20-4:40	Jackie Faherty (AMNH - Astro)	The Brown Dwarf Kinematics Project
4:40-5:00	Rui Castro (Columbia - EE)	Distilled Sensing: Active sensing for sparse recovery
5:00-5:20	David Schminovich (Columbia - Astro)	One trillion time-tagged photons
5:20-5:40	Bodhisattva Sen (Columbia - Stat)	Separating signal from background

9: 30-10:00 - Michael Way – (NASA/GISS)

Title: Using the latest Machine Learning techniques for non-linear regression and newly developed matrix inversion methods to calculate Photometric Redshifts in the Sloan Digital Sky Survey

Abstract: I will quickly introduce galaxy photometric redshifts to this diverse audience. I will then discuss our program of calculating these redshifts using non-linear regression techniques from the Machine Learning Community and newly developed methods for inverting large non-sparse matrices that have made this technique very competitive.

Collaborators:
Ashok Srivastava (NASA/Ames, Intelligent System Division)
Les Foster and students (San Jose State University, Department of Mathematics)
Paul Gazis (NASA/Ames, Kepler Mission)
Jeffrey Scargle (NASA/Ames, Space Sciences Division)

10:00- 10:20 - Liu Jingchen – (Columbia University – Statistics)

Title: Efficient Simulation for Tail Probabilities of Gaussian Random Fields

Abstract: We are interested in computing tail probabilities for the maxima of Gaussian random fields. In this paper, we discuss two special cases: random fields defined over a finite number of distinct point and fields with finite Karhunen-Lo\`{e}ve expansions. For the first case we propose an importance sampling estimator which yields asymptotically zero relative error. Moreover, it yields a procedure for sampling the field conditional on it having an excursion above a high level with a complexity that is uniformly bounded as the level increases. In the second case we propose an estimator which is asymptotically optimal. These results serve as a first step analysis of rare-event simulation for Gaussian random fields.

10:20- 10:40- Michael Blanton – (New York University – Physics)

Title: Interpretation of the images of galaxies

Abstract: Astronomers now have created massive digital data sets of nearby galaxies as well as very distant ones (seen as they were some eight billion years in the past). Determining how the structure of these galaxies has changed over time requires a careful interpretation of their images --- accounting for the variation with distance of the resolution, signal-to-noise ratio, and rest-frame wavelength of the observations. I briefly review the state-of-the-art techniques for doing so, and present a vision of how these approaches can be improved to more precisely understand the growth and evolution of galaxies.

10:40 – 11:00 - Michael Bauer – (NASA/GISS)

Title: Why Weather Matters to Climate Models

Abstract: Among the joys of climate research one rarely lists the flurry of questions from friends, family and others following unusual weather events. These questions range from gleeful challenges like "Where's global warming now?" to breathless worries of "How bad will it get?" Maybe you don't get these. I do. The point is climatologists study climate not weather and as the old saw goes "weather is not climate." Of course this is only half-true as climatologists do study weather, only statistically, and weather is indeed a main ingredient of climate, and yes, even those pesky unusual weather events contribute to it. It shouldn't be surprising then to learn that climate models simulate weather. What may be surprising, though, is that the veracity of this simulated weather is rarely assessed in a direct way. Instead, traditional methods of model validation rely on long-term averages. Which is consistent with the "weather is not climate" sentiment? We will present an alternative approach to model validation, one that makes use of our knowledge of weather processes, such as their patterns of occurrence, structure and behavior to test climate models in a new and informative way. Our aim is to broaden traditional methods of climate model validation not replace them, which is to say that we preserve our climatologist's eye by taking a statistical view of weather rather than the case-by-case perspective of the meteorologist. To do this we have a method for identifying, following and delimiting a target weather phenomenon (in this case mid-latitude cyclones). We will show how this tool can be used to identify specific model deficiencies as well as open up new research possibilities. Examples and pretty pictures will be shown.

11:00 – 11:20 - Zoltan Haiman – (Columbia University – Astronomy)

Title: Can You Spot Dark Energy in a Map With 100 Million Pixels?

Abstract: The apparent photographic shape of a distant galaxy is typically slightly stretched by so--called weak gravitational lensing: the slight bending of light due to the gravitational field of the inhomogeneously distributed foreground mass. Several large astronomical surveys have either been proposed or are underway to measure this effect statistically, over a large fraction of the sky. The properties of the lensing distortion depend on the nature of the mass-energy distribution in the universe. Indeed, weak lensing is currently believed to be one of the most promising approaches to solving the mystery of dark energy, the dominant contribution to the present-day mass-energy. I will describe recent analytical methods and numerical simulations, which attempt to quantify how well we can constrain the properties of dark enery with large maps of the lensing distortion. I will also argue that our existing forecasts have not fully clarified the information content of such maps, and will attempt to provoke computer scientists (and others) to come up with a better statistical approach.

11:20- 12:10 - Kevin Knuth – (University at Albany – Physics)

Title: Automating the Scientific Method

Abstract: In the last decade we have seen computer science, statistics, and the earth, space, life and social sciences come together with a new synergy based on the common goal of data analysis. These multi-disciplinary interactions have become necessary as we pursue both high quality data analysis as well as analysis of extremely large data sets. However, the ultimate goal is more fundamental than mere data analysis. We aim to automate the scientific method itself.

The scientific method relies on the cyclic application of three activities: hypothesis generation, inquiry (experimental design) and inference (data analysis). The majority of our efforts at this point have been focused on the process of automating inference. However, little attention has been paid to automating the processes of inquiry and hypothesis generation.

The most scientifically-useful approach to data analysis is model-based. I will briefly review the methodology behind automating model-based inference with a focus on Bayesian probability theory. I will then introduce a new related methodology called the inquiry calculus, which enables the automation of model-based inquiry. Automated hypothesis (model) generation will be left for another day, as it is the least advanced of the three technologies. I will demonstrate the application of automated inference and inquiry with a robotic scientist that performs its own experiments and analyzes its own data.

12:50- 1:40 - Alexander Gray – (Georgia Institute of Technology)

Title: How to do Machine Learning on Massive Astronomical Datasets

Abstract: I'll describe algorithms and data structures for allowing the most powerful machine learning methods, which often scale quadratically or even cubically with the number of data points, to be performed many orders of magnitude faster than naive implementations. Such techniques can make previously impossible statistical analyses tractable on the scale of entire sky surveys. I will discuss scalable algorithms we have developed for n-point correlations, friends-of-friends, nearest-neighbors, kernel density estimation, nonparametric Bayes classification, principal component analysis, local linear regression, isometric non-negative matrix factorization, hidden Markov models, k-means, support vector machine-like classifiers, Gaussian process regression, and Gaussian graphical model inference, among others. In addition to techniques inspired by computational geometry, fast multipole methods, and Monte Carlo integration, we employ a distributed framework which can be thought of as a higher-order version of Google's MapReduce. Our algorithms have enabled several first-of-a-kind large-scale cosmological analyses.

1:40 – 2:00 - Sunil Bhaskaran – (Lehman College)

Title: Integrating Hyperspectral Data and GIS data for modeling vulnerability from natural disasters: A Case Study from Australia

Abstract: Natural disasters such as Hail Storms can devastate a region within few hours and can cost billions of dollars in damage. Rapid decisions have to be made with limited resources. A geospatial model can be very useful during emergencies for providing a better understanding of infrastructure and demographic characteristics in disaster impacted regions. The project aims to combine airborne hyperspectral data analysis and spatial analysis of surface data for developing a geospatial disaster assessment model in Sydney, Australia. Methodology included spectral analysis of roof materials and integration with census data from the Australian Bureau of Statistics (ABS). Results included a roof distribution and vulnerability map that may be used by emergency services for determining resource allocation, visualization and what if scenario simulations. This integrated database product, which merges high quality spectral information and cartographic GIS data, has vast potential to assist emergency organizations, city planners and decision makers in formulating plans and strategies for resource management.

2:00 – 2:20 - Jo Bovy – (New York University – Physics)

Title: Inferring the velocity distribution of nearby stars from Hipparcos data

Abstract: We present a three-dimensional reconstruction of the velocity distribution of nearby stars ($\lesssim 100$ pc) using a maximum likelihood density estimation technique applied to the two-dimensional tangential velocities of the stars. The underlying distribution is modeled as a mixture of Gaussian components. The algorithm reconstructs the error-deconvolved distribution function, even when the individual stars have unique error and missing-data properties. We apply this technique to the tangential velocity measurements from a kinematically unbiased sample of 11,865 main sequence stars observed by the Hipparcos satellite. We explore various methods for validating the complexity of the resulting velocity distribution function, including criteria based on Bayesian model selection and minimum coding inference, as well as how accurately our reconstruction predicts the radial velocities of a sample of stars from the Geneva-Copenhagen survey. Thus, we can quantify the information content of the radial velocity measurements, which is interesting in the light of the upcoming Gaia mission. We find that the mean amount of new information gained from a radial velocity measurement of a single star is significant, which strongly argues for a complementary radial velocity survey to Gaia. We also confirm the existence of `moving groups" in the velocity distribution of the disk of the Galaxy, quantifying their statistical significance for the first time. We find that the color-magnitude diagrams of most of the moving groups are inconsistent with being trails of evaporating, young clusters, which favors their interpretation as being due to dynamical resonances or non-axisymmetry and time-dependence of the disk potential.

2:20-2:40 - Arlin Crotts – (Columbia University – Astronomy)

Title: Image Subtraction in Astrophysics: Applications Over 14 Orders of Magnitude in Distance

Abstract: We have been conducting imaging surveys for transient sources and writing image subtraction software pipelines to isolate these sources among terabytes of images. We will describe our techniques and results.

2:40 – 3:00 - Ji Meng Loh – (Columbia University – Statistics)

Title: The Marked Point Bootstrap

Abstract: The bootstrap is a popular procedure for obtaining standard errors of estimates. The original bootstrap was meant to be applied to independent data, but various extensions to the bootstrap have been developed for use with dependent data such as time series and spatial data. Examples are subsampling and the block bootstrap. I will introduce the marked point bootstrap as yet another method for bootstrapping spatial data. It has a relatively simple formulation, can be much faster than block bootstrap, and has some desirable properties.

3:00-3:20 - Caleb Scharf – (Columbia University – Astrophysics Lab)

Title: Challenges in understanding the formation of planetary systems

Abstract: Although over 300 planets are now cataloged around other stars, a complete model for planet formation is currently lacking. A number of characteristics of exoplanets - including orbital ellipticities, and orbital periods - indicate that strong dynamical (gravitational) interaction may have taken place amongst objects in a majority of known systems. I will outline a computational approach to investigating this possibility that involves the high precision simulation of hundreds to thousands of planetary systems over tens of millions of years in order to derive statistical expectations that can be tested against observation. Challenges include the need to sample the distribution "tails" of the planetary population – in particular the planets on very large orbits that are just beginning to be successfully imaged.

3:20 – 3:40 - Benjamin Herman – (City College of New York)

Title: Stochastic Monte Carlo methods for non-linear statistical inverse problems

Abstract: Non-linear inverse problems using sparse measurement sets suffer from difficulties of non Gaussian uncertainty distributions and multiple maximum a posteriori solutions. As an example, lidar instrumentation can estimate aerosol extinction and backscatter coefficients at a limited number of wavelengths, typically lower than the number of degrees of freedom of aerosol models. Accurate uncertainty assessment is crucial to situations such as this one. Uncertainty is represented in the form of a posterior probability density function (PDF). This can include a prior PDF which can be incorporated as a more robust alternative to parameter constraints. The posterior PDF can theoretically be used to assess uncertainty of specific aerosol properties in the form of a derived marginal PDF; however it is not computationally practical with PDFs having more than a few dimensions. In the approach that I present for performing uncertainty assessments the Metropolis-Hastings Markov chain Monte Carlo algorithm is used to generate samples of aerosol model parameters congruent with their posterior PDF. The method is effectively a blending of Metropolis-Hastings, genetic algorithm, and Gauss-Newton inverse method, and can deal with the difficulties put forth.

3:40 – 4:00 - David Hogg – (New York University – Physics)

Title: Automated calibration, or generating and trusting astronomical meta-data

Abstract: We have built a reliable and robust system that takes as input an astronomical image, and returns as output the pointing, scale, and orientation of that image (the astrometric calibration or WCS information). The system uses computer vision techniques; it requires no first guess, and works with the information in the image pixels alone. The success rate is better than 99.9 percent for contemporary near-ultraviolet and visual imaging survey data, with no false positives. We can also calibrate (at low accuracy) the date at which the image was taken, the bandpass, and the photometric sensitivity. We are using this system to generate consistent and standards-compliant meta-data for all digital and digitized astronomical imaging, no matter what its archival state, including imaging from plate repositories, individual scientific investigators, and amateurs. This is necessary for worldwide systems like the "virtual observatory" that depend on reliable meta-data but have no explicit trust model. For us it is the first step in a program of making all of the world's heterogeneous astronomical imaging data interoperable.

4:00 – 4:20 - Jianting Zhang – (City College – Computer Science)

Title: Testing the Correlations between Beta Diversity and Productivity Variations using NASA MODIS Global NDVI datasets

Abstract: Despite considerable amount of research on the relationships between species diversity and productivity at different spatial, ecological, and taxonomic scales, the overall trend of the correlation at the global scale still remains sketchy. In this study we use NASA MODIS NDVI as the surrogate of productivity, and the World Wild Fund (WWF) species distribution data to test correlations between beta diversity and productivity variations at different taxonomic ranks at a global scale. Matrix correlation is performed between species composition measured as beta diversities using SÃ¸rensen similarity index and MODIS NDVI/productivity measured as Bhattacharyya distances through Mantel permutation tests. The correlation coefficients and Mantel test significance levels are reported at the global ecoregion, biogeographical realm, and biome levels, respectively.

4:20 – 4:40 - Jackie Faherty – (American Museum of Natural History – Astrophysics)

Title: Testing the Correlations between Beta Diversity and Productivity Variations using NASA MODIS Global NDVI datasets

Abstract: Despite considerable amount of research on the relationships between species diversity and productivity at different spatial, ecological, and taxonomic scales, the overall trend of the correlation at the global scale still remains sketchy. In this study we use NASA MODIS NDVI as the surrogate of productivity, and the World Wild Fund (WWF) species distribution data to test correlations between beta diversity and productivity variations at different taxonomic ranks at a global scale. Matrix correlation is performed between species composition measured as beta diversities using SÃ¸rensen similarity index and MODIS NDVI/productivity measured as Bhattacharyya distances through Mantel permutation tests. The correlation coefficients and Mantel test significance levels are reported at the global ecoregion, biogeographical realm, and biome levels, respectively.

4:40 – 5:00 - Rui Castro – (Columbia University – Electrical Engineering)

Title: Distilled Sensing: Active sensing for sparse recovery

Abstract: A selective sampling methodology called \emph{Distilled Sensing} (DS) is proposed for recovering sparse signals in noise. DS exploits the fact that it is often easier to rule out locations that do not contain signal than it is to detect the locations of non-zero signal components. We formalize this observation and use it to devise a sequential selective sensing strategy that focuses sensing/measurement resources towards the signal subspace. This adaptivity in sensing results in rather surprising gains in sparse signal recovery compared to non-adaptive sensing. We show that exponentially weaker sparse signals can be recovered via DS compared with conventional non-adaptive sensing.

5:00 – 5:20 - David Schminovich – (Columbia University – Astronomy)

Title: One trillion time-tagged photons

Abstract: The Galaxy Evolution Explorer is an ultraviolet space telescope that has been mapping the UV sky for over five years. Unbeknownst to most astronomical users, GALEX images are produced using a raw data set of nearly one trillion time-tagged photon positions. I will describe the computational and algorithmic challenges posed by this vast data set, several solutions, and some unique but still largely unexplored astronomical applications. These methods are becoming increasingly relevant with the advent of low light level CCDs and other highly efficient photon-counting detectors.

5:20 – 5:40 - Bodhisattva Sen – (Columbia University – Statistics)

Title: Separating signal from background:

Abstract: A complicating feature in certain astronomical data sets is that there is contamination by foreground/background objects. We consider a sample of stars from the dwarf spheroidal Sextans (with contamination), where some stars are foreground stars, located along the line of sight. We develop an algorithm for estimating parameters of the distribution sampled with contamination. Our approach is based on the well-known "expectation maximization" (EM) algorithm. Given models for both member and contaminant populations, the EM algorithm iteratively evaluates the membership probability of each discrete data point, then uses those probabilities to update parameter estimates for member and contaminant distributions. We also discuss some non-parametric extensions of this approach.