Final Report for the SAMSI Program on Statistical and Computational Methodology for Massive Datasets (2012 - 2013)

Summary. The Massive Datasets Programs consisted of 10 active research working groups (WGs). The WGs were made up of 5 SAMSI postdocs, 28 SAMSI visitors, and a large number of external and local participants (students, postdocs, faculty, members from government and industry).

The WG activities were supported by 6 workshops (a planning workshop and 5 research workshops). At least one of the WGs will continue operating as part of the 2013-2014 SAMSI Program on Low Dimensional Structure in High Dimensional Systems. A two-day undergraduate workshop formed the education and outreach part of the Massive Datasets Program

Contents

1 Overview

We give a brief summary of the main people involved, workshops, educational activities, postdocs, visitors, and research working groups (WGs) in the Massive Datasets Program.

In the text below, asterisks denote women, or members of under-represented groups.

1.1 Overall Program Leaders

Michael Jordan (EE + CS, UC Berkeley), Karen Kafadar* (Stat, U. Indiana), Michael Mahoney (Math, Stanford), Steve Sain (NCAR), Jiayang Sun* (Stat, Case Western), Alexander Szalay (Physics & Astro, Johns Hopkins)

NAC Liaison: Bin Yu* (Stat, UC Berkeley)
Local Scientific Coordinator: Yufeng Liu (Stat, UNC)
SAMSI Directorate Liaison: Ilse Ipsen* (Math, NCSU)
Web Page: `http://www.samsi.info/MD12`

1.2 Research Workshops

A Planning Workshop was instrumental in developing the Massive Datasets Program, and five research workshops during the program were designed to support the activities of the WGs. The workshops are described in more detail in Section 3.

- Planning Workshop, UC Berkeley, 20 May 20122
  `http://www.samsi.info/MD-planning`

- Opening workshop, Radisson Hotel, 9-12 September 2012
  `http://www.samsi.info/MD-opening-workshop`

- Astrostatistics workshop, SAMSI, 19-21 September 2012
  `http://www.samsi.info/AS12`

- SAMSI-FODAVA Workshop on Interactive Visualization & Analysis of Massive Data, SAMSI 10-12 December 2012
  `http://www.samsi.info/IV12`

- SAMSI-NCAR Workshop on Massive Datasets in Environment & Climate, NCAR, 12-15 February 2013
  `http://www.samsi.info/EC13`

- Transition Workshop, Radisson Hotel, 20-22 May 2013
  `http://www.samsi.info/MD-transition`

1.3 Educational Activities

A two-semester graduate course *Computational and inferential methods for high dimensions and massive datasets* focused on fundamental methodological questions of statistics, mathematics and computer science posed by massive datasets, with applications to astronomy, high energy physics, and the environment. The course coordinator was Naomi Altman* (Stat, Penn State).

An undergraduate workshop on 26-27 October 2012
`http://www.samsi.info/UGF12`

represented the education and outreach portion of the SAMSI Massive Datasets Program, and is described in more detail in Section 2.

1.4 Postdocs

Five postdocs were associated with the Massive Datasets Program. They are listed below, together with the WGs in which they participated and their scientific and SAMSI mentors.

Dorit Hammerling* (Ph.D. Env. Eng, U. Michigan)
WGs: Environment & climate, inference
Scientific mentor: Montserrat Fuentes* (Stat, NCSU)
SAMSI mentor: Richard Smith (STOR, UNC)

David Lawlor (Ph.D. Math, Michigan State)
WGs: Online streaming & sketching, multiscale modeling, graphical models & graphics processors
Scientific mentor: Mauro Magioni (Math, Duke)
SAMSI mentor: Ilse Ipsen* (Math, NCSU)

Garvesh Raskutti (Ph.D. Stat, UC Berkeley)
WGs: Online streaming & sketching, inference
Scientific mentor: Naomi Altman* (Stat, Penn State)
SAMSI mentor: Richard Smith (STOR, UNC)

Yi Grace Wang* (Ph.D. Math, U. Minnesota)
WGs: Imaging, online streaming & sketching
Scientific mentor: Ingrid Daubechies* (Math, Duke)
SAMSI mentor: Ezra Miller (Math, Duke)

Dan Yang* (Ph.D. Stat, UPenn)
WGs: Inference, Imaging
Scientific mentors: Haipeng Shen (STOR, UNC) and Hongtu Zhu (Biostat, UNC)
SAMSI mentor: Richard Smith (STOR, UNC)

1.5 Visitors

The 28 visitors below were at SAMSI for various amounts of time during the Massive Datasets Program.

Naomi Altman* (Stat, Penn State)
Andreas Artemiou (Math, Michigan TU)
Jogesh Babu (Stat, Penn State)
Vinay Bandaru (Stat, Case Western)
Tamás Budavári (Physics & Astro, Johns Hopkins)
Arindam Chatterjee (Stat & Math, ISI, Delhi, India)
Guang Cheng (Stat, Purdue)
Julian Faraway (Stat, U. Bath, UK)
Marco Ferreira* (Stat, U. Missouri)
Karen Kafadar* (Stat, U. Indiana)

Matthias Katzfuß (Math, U. Heidelberg, Germany)
Brandon Kelly (Physics, UC Santa Barbara)
Amy Langville* (Math, College of Charleston)
Tom Loredo (Astro, Cornell)
Qiyi Lu* (Math Sci, SUNY Binghampton)
Ashish Mahabal (Astro, Caltech)
Nuala McCullagh* (Physics & Astro, Johns Hopkins)
Sarah Michalak* (Stat, Los Alamos National Lab)
SaeNa Park* (Stat, Penn State)
Xingye Qiao (Stat, SUNY Binghampton)
Prajval Shastri* (Astrophysics, IIA Bangalore, India)
Jiayang Sun* (Stat, Case Western)
Wei Sun (Stat, Purdue)
Michael Woodroofe (Stat, U. Michigan)
Bowei Xi* (Stat, Purdue)
Ganggang Xu (Math, Texas A&M)
Zhuqing Yu* (Stat, Purdue)
Tao Yu (Stat, U. Singapore)
Lingsong Zhang (Stat, Purdue)

1.6 Working Groups

The ten active working groups (WGs) and their leaders are listed below. The research activities of the WGs are summarized in Section 4.

Inference: Naomi Altman* (Stat, Penn State)
Imaging: Jiayang Sun* (Stat, Case Western), Dani Ushizima* (Lawrence Berkeley Lab)
Environment & climate: Jessica Matthews* (CICS-NC)
High energy physics: Steffen Bass (Physics, Duke), Karen Kafadar* (Stat, U. Indiana)
Online streaming & sketching: Ilse Ipsen* (Math, NCSU), Petros Drineas (CS, RPI), Michael Mahoney (Math, Stanford)
Datamining & clustering: Amy Langville* (Math, College of Charleston), Carl Meyer (Math, NCSU)
Multi-scale modeling: Marco Ferreira* (Stat, U. Missouri)
Stochastic Processes & Astrophysical Inference: Tom Loredo (Astro, Cornell), Robert Wolpert (Stat, Duke)
Discovery & classification in synoptic surveys: Jogesh Babu (Stat, Penn State)
Graphical models & graphics processors: Tamás Budavári (Physics & Astro, Johns Hopkins U.), David Lawlor (SAMSI)

2 Undergraduate Workshop

This forms the education and outreach part of the Massive Datasets Program, and gives the SAMSI postdocs an opportunity to hone their mentoring skills.

Date: 26-27 October 2012, SAMSI
Web page: `http://www.samsi.info/UGF12`

2.1 Talks

Naomi Altman* (Stat, Penn State): *Overview of Statistical and Computational Methodology for Massive Datasets*
Jogesh Babu (Stat, Penn State): *Big Data in Observational Astronomy*
Andreas Artemiou (Stat, Michigan TU): *Reducing the Dimension in Regression Problems*
Bowei Xi* (Stat, Purdue): *Parallel R*
Dorit Hammerling* (SAMSI postdoc), Garvesh Raskutti (SAMSI postdoc), Dan Yang* (SAMSI postdoc): *An Introduction to the R Fields Package: A Practical Tool for Large Spatial Data*
Snehalata Huzurbazar* (SAMSI): *Career Options*
Marco Ferreira* (Stat, U. Missouri): *Looking at the World at Multiple Scales of Resolution*
Nate Burch (SAMSI 2. year postdoc and Math, NCSU): *PCA, SVD, and JPEGs, ASAP!*
David Lawlor (SAMSI postdoc), Yi Grace Wang* (SAMSI postdoc): *Overview of Matlab with Wavelets and Image Processing*

2.2 Panel Discussions

Two panel discussions were an integral part of the Undergraduate Workshop:

The Road to Industry and Academia Panelists: Naomi Altman* (Stat, Penn State), Christopher Gotwalt (Director of JMP Statistical Research and Development, SAS), Richard Smith (SAMSI), Amy Nail* (President of Honestat, LLC)

The Road to Graduate School and Immediately Beyond Panelists: Mary Beth Broadbent* (PhD student, Stat), David Lawlor (SAMSI postdoc), Kevin Penner (PhD student, Math, NCSU), Gina-Maria Pomann* (PhD student, Stat, NCSU)

3 Research Workshops

We give a brief survey of the planning workshop, and the five research workshops that supported the WG activities in the Massive Datasets Program.

3.1 Planning Workshop

The purpose of the Planning Workshop at UC Berkeley on 20 May 2011 was to identify research topics and program leaders for the Massive Datasets Program, and to acquaint the participants with SAMSI.

Date: 20 May 2011, UC Berkeley
Web page: `http://www.samsi.info/MD-planning`

Nine initial research topics were identified, including three application areas (astronomy, high energy physics, and the environment).

3.1.1 Participants

The 25 participants at the Planning Workshop (with 13 coming from the Bay Area/West Coast) represented Mathematics, Statistics, Electrical Engineering & Computer Science, Astronomy, and the National Labs. They are:

Naomi Altman* (Stat, Penn State) Tamás Budavári (Astro & Physics, Johns Hopkins), John Chambers (Stat, Stanford), Dan Crichton (Jet Propulsion Lab), Jim Demmel (EECS, UC Berkeley), Noureddine El Karoui (Stat, UC Berkeley), Alex Gray (EECS, Georgia Tech), Trevor Hastie (Stat, Stanford), Arnie Horta (DoD), Ilse Ipsen* (SAMSI), Bob Jacobsen (Lawrence Berkeley National Lab), Michael Jordan (EECS, UC Berkeley), Karen Kafadar* (Stat, U. Indiana), Yufeng Liu (STOR, UNC), Sarah Michalak* (Los Alamos National Lab), Michael Mahoney (Math, Stanford), Ketan Mane (RENCI), Dan Merl (Lawrence Berkeley National Lab) Steve Sain (NCAR), Richard Smith (SAMSI), Jiayang Sun* (Stat, Case Western), Duncan Temple Lang (Stat, UC Davis) Daniela Ushizima* (Lawrence Berkeley National Lab), John Wu (Lawrence Berkeley National Lab), Bin Yu* (Stat, UC Berkeley)

3.2 Opening Workshop

This is the official start of the year-long SAMSI Research Program on Massive Datasets.

> **Date:** 9-12 September 2012, Radisson Hotel, RTP
> **Web page:** `http://www.samsi.info/MD-opening-workshop`

3.2.1 Organizers

Naomi Altman* (Stat, Penn State), Michael Mahoney (Math, Stanford), Jiayang Sun* (Stat, Case Western), Dani Ushizima* (Lawrence Berkeley Lab)

3.2.2 Tutorials

The Opening Workshop began with an extremely well attended session – over 200 participants – of five tutorial talks on Sunday.

Tamás Budavári (Physics & Astro, Johns Hopkins)
Statistical methods in astronomy
Petros Drineas (CS, Rensselaer Polytechnic Institute)
Mining massive datasets: A (randomized) linear algebraic perspective
Haesun Park* (CSE, Georgia Tech)
Visual analytics for knowledge discovery in high dimensional data
Michael Jordan (EECS & Stat, UC Berkeley)
Resampling methods for massive data
Steve Wright (CS, Wisconsin)
Optimization techniques for statistical analysis of large datasets

Videos of the first four tutorials are available here
`http://www.samsi.info/communications/videos`

3.2.3 Topical Talks

For the remaining days (Monday – Wednesday), each morning and each afternoon were allocated to a particular topic, with a set of three talks followed by a panel discussion. The panel discussions were designed to contribute research ideas for the working groups.

A poster session, preceded by a poster blitz, took place on Monday evening. The Opening Workshop ended with the formation of working groups on Wednesday afternoon.

1. **Inference:**
   Bin Yu* (Stat, UC Berkeley), Xiatong Shen (Stat, U. Minnesota), Brian Caffo (Biostat, Johns Hopkins)
   Panel: Bill Eddy (Stat, CMU), Alex Gray (CSE, Georgia Tech), Karen Kafadar* (Stat., U. Indiana), Bo Li* (Stat, Purdue)

2. **Imaging:**
   Jim Nagy (CS, Emory), Jianqing Fan (Finance & Stat, Princeton), Rollin Thomas (Lawrence Berkeley Lab)
   Panel: Dani Ushizima* (Lawrence Berkeley Lab), Jiayang Sun* (Stat, Case Western), Peihua Qiu (Stat, U. Minnesota), Erkki Somersalo (Math, Case Western)

3. **Environment & Climate:**
   Anna Michalak* (Earth Sciences, Stanford), Dan Crichton (Jet Propulsion Lab), Noel Cressie (Stat, Ohio State and University of Wollongong, Australia)
   Panel: Jessica Matthews* (CICS North Carolina), Amy Braverman* (Jet Propulsion Lab), Steve Sain (NCAR), Richard Smith (Stat, SAMSI and U. North Carolina)

4. **High Energy Physics:**
   Steffen Bass (Physics, Duke University), Kyle Cranmer (Physics, NYU), Luc Demortier (Physics, Rockefeller University)
   Panel: Robert Wolpert (Stat, Duke University), Mandeep Gill (Astrophysics, Stanford), Cosma Shalizi (Stat, CMU), Daniel Whiteson (Physics, UC Irvine)

5. **Streaming, Sketching & Datamining:**
   Michael Mahoney (Math, Stanford), Maryam Fazel* (Electrical Eng., U. Washington), Inderjit Dhillon (CS, UT Austin)
   Panel: Piotr Indyk (CS, MIT), Graham Cormode (AT&T Labs Research), Ashish Goel (Management Science & Eng., Stanford), Michael Mahoney (Math, Stanford)

The session on high energy physics was extremely interesting, with the last two talks discussing the highly complex statistical talks involved in the discovery of the Higgs Boson. But, for some reason we do not understand, the session turned out to be very poorly attended.

SAMSI Blog `http://samsiatrtp.wordpress.com/`
The Blog collects impressions, suggestions and thoughts from the OW attendees.

Andrea Campos: *Impressions of the Massive Datasets Opening Workshop*
Noel Cressie: *Thoughts on Bayesian Statistical Inference for Regional Climate Projections in North America*
Brian Caffo: *Impressions of the Massive Datasets Opening Workshop*
Rollin Thomas: *Astrophysicists are using Palomar Transient Factory Wide-Field Imaging Data*

3.3 Astrostatistics Workshop

The purpose of the SAMSI Astrostatistics Workshop was to bring together astrophysicists and statisticians, so they could brainstorm on advanced topics in statistical inference for modern empirical astrophysics.

The importance of rigorous application of statistical methods to astrophysical data analysis has increased enormously in the last two decades. Large data sets at multiple wavelengths require complex automated processes that involve a diverse range of sophisticated statistical techniques. The current and upcoming astrophysics missions such as the SDSS, Planck, LSST and LIGO will push this trend to even larger data sets and parameter sets.

Date: 19-21 September 2012, SAMSI
Web Page: `http://www.samsi.info/AS12`

3.3.1 Organizers

Jogesh Babu (Stat, Penn State), Prajval Shastri* (Astrophysics, IIA Bangalore, India)

3.3.2 Talks

Each morning and afternoon started with a one-hour tutorial, followed by 30-minute topical talks. The second day culminated in a poster session. The talks were a perfect mix of astronomy, statistics, and computer science. The workshop ended with a panel discussion on future directions in astrostatistics.

Jim Berger (Stat, Duke): *Bayesian Statistics Overview*
Fabrizia Guglielmetti* (Max Plank Institute, Germany): *Astronomical Image Analysis Employing Bayesian Mixture Models*
Roberto Trotta (Astrophysics, Imperial College, London, UK): *Bayesian Hierarchical Modeling of Supernovae Type 1a*
Sergey Klimenko (Physics, U. Florida): *Searches for Gravitational Waves with the Ground-Based Interferometer*
Tom Loredo (Astronomy, Cornell): *"Big Data" Issues from a Bayesian Perspective*
Ashish Mahabal (Astronomy, Caltech): *Detecting Faint Intermittent Sources in Large Datasets*
Hakeem Oluseyi* (Physics & Space Sciences, Florida Inst. Tech.): *lessons Learned from Automated Analyses of 36,000,000 Light Curves*
Aneta Siemiginowska* (Astrophysics, Harvard): *Bayesian Methods in High Energy Astrophysics*
Kirk Borne (Astrophysics & CS, George Mason U.): *Learning from Big Data in Astronomy – Overview*
Bhuvnesh Jain (Physics & Astronomy, U. Penn): *The DES and LSST Surveys: Extracting Information from Galaxy Images*
Debbie Bard* (SLAC, Stanford): *Multivariate Techniques in Star and Galaxy Identification*
Risa Wechsler* (Physics, Stanford): *Weighing Galaxies using Inference from Cosmological Simulations*
Dave Donoho (Stat, Stanford): *Sparsity – Overview*

Jean-Luc Starck (Astrophysics, CEA/Saclay, France): *Sparsity and Cosmic Microwave Background Analysis*
Ann Lee* (Stat, CMU): *Improved Estimation for Astrophysical Problems via Nonlinear Transformation and Adaptive Bases*
Brandon Kelly (Physics, UC Santa Barbara): *Analyzing Astronomical Data Sets Using Non-Linear Random Effects Models*
Harrison Prosper* (Physics, Florida State): *Bayes at the Frontier: The Promise and the Challenges*
Eric Ford (Astronomy, U. Florida): *Searching and Characterizing Small Transiting Planets Using NASA's Kepler Mission*
Josh Bloom (Astronomy, UC Berkeley): *Automating Science in the Time-Domain Survey Era: Machine-Learning Challenges*
Joseph Richards (Astronomy, UC Berkeley): *Advances in Discovery and Classification for Time-Domain Astronomy*

One of the highlights was Dave Donoho's standing-room-only talk about sparsity, with a call for reproducible research. An extremely lively discussion about modeling ensued during Jean-Luc Starck's subsequent talk.

SAMSI Blog `http://samsiatrtp.wordpress.com/`

Nuala McCullagh: *Impressions from the Astrostatistics Workshop*
Jogesh Babu: *Astronomy and Big Data*

3.4 SAMSI-FODAVA Workshop on Interactive Visualization & Analysis of Massive Data

The purpose of this workshop was to bring together researchers in Mathematics, Statistics, Computational Science & Engineering, Computer Science, and Visualization, to work on visual analytics for massive-scale data.

Massive-scale data pose challenges because they can originate from different sources, and they can vary in time. They can also be unstructured, high dimensional, noisy, and incomplete. Recent developments in visual analytics demonstrate that adding interactive, visual interfaces to automated data analysis methods can substantially increase our ability to understand the data. Issues to be considered include: Mathematical, statistical, and algorithmic issues in efficient representation and transformation of data. Scalable and dynamic algorithms for real-time interaction. Visual representation in limited screen space.

This 2 1/2 day workshop was organized in conjunction with the joint NSF/DHS Initiative on the *Foundations of Data Analysis and Visual Analytics* (FODAVA). Haesun Park*, the PI and FODAVA lead was the main workshop organizer.

Date: 10-12 December 2012, SAMSI
Web Page: `http://www.samsi.info/IV12`

3.4.1 Organizers.

Haesun Park* (CSE, Georgia Tech), Orly Alter* (Bio. Eng., U. Utah), Kwan-Liu Ma (CS, UC Davis), Mauro Maggioni (Math, Duke), Leland Wilkinson (SYSTAT Software, Inc.)

3.4.2 Talks

The following talks talks were all scheduled for thirty minutes.

Daniel Keim (CS, U. Konstanz, Germany): *The role of visualization and analytics in solving problems based on massive data*
Gilad Lerman (Math, U. Minnesota): *Robust subspace modeling*
Santosh Vempala (CS, Georgia Tech): *On the complexity of statistical algorithms*
Heike Hofman* (Stat, Iowa State): *Interactive graphics for data exploration*
Christopher Healy (CS, NCSU): *The role of perception in visualization and visual analytics*
David Glotz (IBM): *Visual analytics for evidence-based medicine*
Chris Johnson (CS, U. Utah): *Large-scale visual data analytics*
Sabine van Huffel* (Electrical Eng., KU Leuven, Belgium): *Computational signal processing in smart patient monitoring: Algorithms, applications and future challenges*
Orly Alter* (Bio. Eng., U. Utah): *Discovery of mechanisms and prognosis of cancers from matrix and tensor modeling of large-scale molecular biological data*
Sayan Mukherjee (Stat, Duke): *The combinatorial Laplacian and dimension reduction*
Kristen Bennet* (CS/Math, RPI): *TB-Vis: Visualizing TB Patient-Pathogen relationships*
Alex Gray (CSE, Georgia Tech): *VisRR: Visual information retrieval and recommendation system for document discovery*
Fei Sha (CS, USC): *New approaches for nonlinear dimensionality reduction*
Naren Ramakrishnan (CS, Virginia Tech): *New approaches to storytelling from massive contextual datasets*

Alan Qi (CS, Purdue): *Scalable Bayesian learning for matrix and tensors*
Ping Li (Stat, Cornell): *BigData: Probabilistic methods for efficient search and statistical learning in extremely high-dimensional data*

The student poster session at the end of the first day featured about 15 posters, and was preceded by a lively “poster fast forward”, where presenters gave one-minute adverts about their posters.

The second day concluded with a panel discussion chaired by Jimeng Sun (IBM). The panelists were Polo Chau (CSE, Georgia Tech), Daniel Keim (CS, U. Konstanz, Germany), Larry Rosenbaum (NSF), and Leland Wilkinson (SYSTAT Software, Inc.)

3.5 SAMSI-NCAR Workshop on Massive Datasets in Environment & Climate

This 2 1/2 day workshop was organized in conjunction with NCAR, and took place at NCAR. The purpose was to unite researchers from the statistical and climate research communities, so they could address the challenges posed by large-scale data sets from environmental observations and climate models.

Date: 12-15 February 2013, NCAR
Web Page: `http://www.samsi.info/EC13`

3.5.1 Organizers

Jessica Matthews* (CICS-NC), Amy Braverman* (Jet Propulsion Lab), Steve Sain (NCAR), Richard Smith (Stat, UNC)

3.5.2 Talks

The workshop focussed on five major themes: Climate observations, climate models, data assimilation, computing techniques for data producers, and regional climate data sets. All talks were scheduled for 45 minutes.

Amy Braverman* (JPL): *Massive data set analysis for NASA's atmospheric Infrared Sounder*
Elizabeth Mannshardt* (Stat, NCSU): *Analysis of NASA Aquarius sea surface salinity data*
Andrew Finley (Forestry, Michigan State): *Bayesian dynamic modeling for large space-time data sets using Gaussian predictive processes*
Doug Nychka (NCAR): *Mining spatial structure in regional climate*
Claudia Tebaldi* (Climate Central and NCAR) *Multi-model ensemble reduction and expansion: Pattern scaling and resampling*
Emily Kang* (Stat, U. Cincinnati): *Spatial analysis of high-resolution regional climate-change projection in North America*
Christopher Jones (Math, UNC): *Challenges of Lagrangian data assimilation*
Tom Auligne (NCAR): *Data assimilation for numerical weather prediction and climate: Challenges with massive datasets*
Rich Loft (NCAR): *Rethinking cyber-infrastructure for climate data analysis workflows*
Dorit Hammerling* (SAMSI)
Eric Fetzer (JPL): *Merging water vapor, temperature and cloud observations from the A-Train*
Annmarie Eldering* (JPL): *Watching the earth breathe - Measuring global carbon dioxide from space*
Seth McGinnis (NCAR): *Distilling regional climate model data from NARCCAP for use in impacts analysis*
Forrest Hoffman (Oak Ridge National Lab)
Dean Williams* (Lawrence Livermore National Lab): *Climate science responds to Big Data challenges: Accessing and analyzing model output and observations*

The first day concluded with a poster session. Interaction among workshop participants was encouraged with daily structured 1-hour breakout sessions, directed by a leader and recorded by a rapporteur. The results from the discussions of the breakout sessions were reported on the third day, at the end of the workshop.

3.6 Transition Workshop

The Transition Workshop marked the official end of the SAMSI Massive Datasets Program. Its purpose was several-fold: Reunite the participants of all WGs for one last time; report and review the progress of each WG; and plan on how to continue the research activities of the WGs beyond the SAMSI Massive Datasets Program. At the close of registration we had 60 accepted participants.

Date: 20-22 May 2013, Radisson Hotel, RTP
Web Page: `http://www.samsi.info/MD-transition`

3.6.1 Talks

A time slot was allocated for each WG, to present their results. Many WGs made a concerted effort to give early-career participants (graduate students and postdocs are indicated by ^{o}) an opportunity for public exposure.

1. **Environment & climate**
   Chair: Amy Braverman* (JPL)
   Speakers: Dorit Hammerling*^{o} (SAMSI),
   Matthias Katzfuß^o (Math, U. Heidelberg, Germany), Peter Thorne (CICS-NC)

2. **Online streaming & sketching**
   Chair: Ilse Ipsen* (Math, NCSU)
   Speakers: Michael Mahoney (Math, Stanford), David Lawlor^o (SAMSI),
   Garvesh Raskutti^o (SAMSI)

3. **High energy physics**
   Chair: Robert Wolpert (Stat, Duke)
   Steffen A. Bass (Physics, Duke), Peter Marcy^o (Stat, Wyoming),
   Chris Coleman-Smith^o (Physics, Duke), Karen Kafadar* (Stat, U. Indiana)

4. **Inference:**
   Chair: Naomi Altman* (Stat, Penn State)
   Speakers Xia Wang* (Math Sci, University of Cincinnati),
   Tao Yu (Stat, National University of Singapore), Yufeng Liu (STOR, UNC)

5. **Datamining & clustering**
   Chair: Carl Meyer (Math, NCSU)
   Speakers: Shaina Race*^{o} (OR, NCSU), Ralph Abbey^o (Math, NCSU)

6. **Astrostatistics (Stochastic processes & astrophysical inference, Graphical Models & graphics processors)**
   Chair: Jogesh Babu (Stat, Penn State)
   Speakers: Tamás Budavári (Physics & Astro, Johns Hopkins),
   Robert Wolpert (Stat, Duke), Mary Beth Broadbent*^{o} (Stat, Duke)

7. **Discovery & classification in synoptic surveys**
   Chair: Jogesh Babu (Stat, Penn State)
   Speakers: Fabrizia Guglielmetti* (Max Planck Institut, Germany)

8. **Multiscale modeling**
   Chair: Karen Kafadar* (Stat, U. Indiana)
   Speakers: David Dunson (Stat, Duke), Mauro Maggioni (Math, Duke),
   Marco Ferreira* (Stat, U. Missouri)

9. **Imaging**
   Chair: Jiayang Sun* (Stat, Case Western)
   Speakers: Ashish Mahabal (Astro, Caltech), Lingsong Zhang (Stat, Purdue),
   Weihong Guo* (Case Western), Dan Yang*[o] (SAMSI),
   Dani Ushizima* (Lawrence Berkeley National Lab)

A poster reception took place at end of the first day. Furthermore, during each day, ample free time was available for WGs to discuss the continuation of their research beyond the Massive Datasets Program, and also to work on their final reports. In the afternoon of the third day, the SAMSI Director and Associate Director gave information on how to continue the WG activities with the help of SAMSI, and how to stay connected to SAMSI. The Transition Workshop finally ended with a small good bye party.

4 Working Group Reports

Below are short reports submitted by the WGs about their research activities.

4.1 Inference

The goals of the WG are to address statistical inference problems that arise in massive data problems, as well as computation methods to support inference. Many members of the working group are also members of other working groups, establishing collaborative links with most other working groups in the massive data program.

The WG consisted of seven subgroups: Multiple Testing, graphical covariance inference under noise, dimension reduction, inference after dimension reduction, semiparametric modeling, classification and learning, simulation in complex models

WG members: Naomi Altman* (Stat, Penn State), Andreas Artemiou (Stat, Michigan TU), David Banks (Stat, Duke), Guang Cheng (Stat, Purdue), Yufeng Liu (STOR, UNC), Wenbin Lu (Stat, NCSU), Xingye Qiao (Stat, SUNY Binghamton), Ali Shojaie (Biostat, U. Washington), Jiayang Sun* (Stat, Case Western), Xia Wang* (Stat, U. Cincinnati), Bowei Xi* (Stat, Purdue), Kai Zhang (STOR, UNC), Lingsong Zhang (Stat, Purdue), Jian Zou (Stat, Indiana U.-Purdue U. Indianapolis)

Postdocs: Dorit Hammerling* (SAMSI), Alan Lenarcic (Genomics, UNC), Garvesh Raskutti (SAMSI)

Graduate Students: Fernando Bonassi (Stat, Duke), Nicholas Jarrett (Stat, Duke), Wonyul Lee (STOR, UNC), Qiyi Lu* (Math Sci., SUNY Binghamton), Weining Shen (Stat, NCSU), Sunyoung Shin* (STOR, UNC), Wei Sun (Stat, Purdue), Zhuqing Yu* (Stat, Purdue), Xiang Zhang (Stat, NCSU) endenumerate

Publications directly resulting from WG research

1. Wonyul Lee and Yufeng Liu, *Joint Estimation of Multiple Precision Matrices with Common Structures*, Journal of Machine Learning Research, submitted

2. Xiang Zhang and Yichao Wu, *Variable Selection of Support Vector Machine in High Dimensions*, Journal of the Royal Statistical Society, Series B, submitted

3. Xingye Qiao and Lingsong Zhang, *Distance-weighted Support Vector Machine*, Electronic Journal of Statistics, submitted

4. C. Zhang and Yufeng Liu, *Multicategory Large-margin Unified Machines*, Journal of Machine Learning Research, to appear

5. Y. Wu. and Y. Liu, *Adaptively weighted large margin classifiers*, Journal of Computational and Graphical Statistics, to appear

6. Y. Wu and Y. Liu, *Functional robust support vector machines for sparse and irregular longitudinal data*, Journal of Computational and Graphical Statistics, to appear

Other publications related to work done at SAMSI

1. W. Luo and N. S. Altman, *A Characterization of Conjugate Priors in Linear Exponential Families with application to Dimension Reduction*, Statistics and Probability Letters, 83, 650-654, 2013

2. G. K. Smyth and N. S. Altman, *Individual Channel Analysis of Two-Channel Microarrays*, BMC Bioinformatics, 14, 165, 2013, doi:10.1186/1471-2105-14-165

3. H. Huang, Y. Liu, Y. Du, C. M. Perou, D. N. Hayes, M. J. Todd, and J. S. Marron, *Multiclass distance weighted discrimination*, Journal of Computational and Graphical Statistics, to appear

4. W. Lee, Y. Du, W. Sun, D. N. Hayes, and Y. Liu, *Multiple Response Regression for Gaussian Mixture Models with Known Labels*, Statistical Analysis and Data Mining, 5, 6, 493-508, 2012

5. H. Huang, Y. Liu, and J. S. Marron, *Bi-directional discrimination with application to data visualization*, Biometrika, 99, 4, 851-864, 2012

6. H. Huang, X. Lu, Y. Liu, P. Haaland, and J. S. Marron, *R/DWD: Distance Weighted Discrimination for Classification*, Visualization and Batch Adjustment. Bioinformatics, 28, 8, 1182-1183, 2012

7. W. Lee, and Y. Liu, *Simultaneous Multiple Response Regression and Inverse Covariance Matrix Estimation via Penalized Gaussian Maximum Likelihood*, Journal of Multivariate Analysis, 111, 241-255, 2012

Grant proposals relating to work done at SAMSI

1. Sanjay Rao, William Cleveland, Ramana Kompella, Gene Spafford, Bowei Xi, Baijian Yang, *Endace Traffic Collection Box*, Purdue Research Foundation

2. Bowei Xi and William Cleveland, *Internet Traffic: Modeling Multiplexed Packet Arrivals for Network Engineering, and Connection Level Packet Dynamics*, NSF-DMS

3. Tim Berners Lee, Lalana Kagal, Srini Devadas, Bhavani Thuraisingham, Murat Kantarcioglu, Kevin Hamlen, Daniel Krawczyk, James Bartlett, Robert Morris, Zhiqiang Lin, Sharad Mehrotra, Michael Lee, Jennifer Trueblood, Nalini Venkatasubramanian, Chinya Ravishankar, Weiwei Wang, Vagelis Hristidis, Elisa Bertino, Sonia Fahmy, Dongyan Xu, Bowei Xi, Luo Si, *Cyber Security Collaborative Alliance*, Department of Defense

Conference presentations

1. Topic Contributed Session at Joint Statistical Meetings (JSM), 6 August 2013

   (a) Xia Wang, Ali Shojaie and Jian Zhou, *Bayesian large-scale multiple testing for time series data* (multiple testing)

   (b) Andreas Artemiou and Yufeng Liu, *Adaptively weighted large margin classifiers for Sufficient Dimension Reduction* (dimension reduction)

   (c) Guang Chen and Yufeng Liu, *Two-stage Selection Procedure for Large-margin Classifiers - Wei Sun, Purdue University* (classification and learning)

(d) Xiang Zhang, Lan Wang, Runze Li and Yichao Wu, *Variable Selection for Support Vector Machine on High Dimensions* (classification and learning)

2. Naomi Altman, *Exploratory and Inferential Methods for Massive Data* Joint Statistical Meetings, 8 August 2013 (dimension reduction)

3. Yufeng Liu, *Joint estimation of multiple dependent Gaussian Graphical Models*, Joint Statistical Meetings, 2013 (classification and learning)

Topics investigated by WG

1. *Multiple Testing* (Naomi Altman, Dorit Hammerling, Karen Kafadar, Xingye Qiao, Ali Shojaie, Jiayang Sun, Xia Wang, Lingsong Zhang, Jian Zou

   In massive data problems it is often necessary to perform inference for a large number of parameters via tests or interval estimates. It is essential to adjust the inference to account for the increased probability of observing extreme events when multiple events are observed. The issues of particular interest to the working group included multiple testing adjustments for correlated data, discrete tests and some meta-analytic techniques for combining test results.

2. *Graphical Covariance Inference under Noise* (Alan Lenarcic, Michael Mahoney, Ali Shojaie)

   This is a continuation of a topic originally considered in SAMSI Graphical Models Workshop (2010-2011). We used UNC Killdevil resources to research covariance under noise estimation for 20K node graphs of various hyperbolicity. We demonstrated our algorithm had good performance on tree decomposable graphs, and we also coded a new C++ version of a competing algorithm that, now reimplemented, has competitive performance.

3. *Dimension Reduction* (Naomi Altman, Andreas Artemiou, Dan Yang, Garvesh Raskutti, Qiyi Lu, Bowei Xi, Kai Zhang, Lingsong Zhang)

   When there are a large number of variables or dimensions, statistical visualization is difficult and inference may be inefficient. Sufficient dimension reduction is a set of dimension reduction methods which use information from a response variable to reduce the dimension of a predictor. The group decided to focus at the beginning on functional data and depending on the progress to extend the results to graph based and tree based data. A subgroup worked on generalizations of principal component analysis for non-elliptical distributions.

4. *Inference after Dimension Reduction* (Andreas Artemiou, Guang Cheng, Wonyul Lee, Yufeng Liu, Sunyoung Shin, Bowei Xi, Kai Zhang, Lingsong Zhang)

   Dimension reduction is very important for high dimensional data analysis. However, how to perform valid inference on the final model to correctly account for the impact of dimension reduction remains a challenge. Recently, the work of Kai Zhang and others (Berk et al., to appear in Annals of Statistics) offers a promising approach to perform valid inference for models after arbitrary dimension reduction and variable selection. Despite its validity, the method can be too conservative. The group decided to tackle problems with special structures to obtain more efficient inference procedures. One example is polynomial regression and the other example is the LASSO. The goal is to derive useful inference procedures for these kinds of techniques.

5. *Semiparametric Modeling* (Guang Cheng, Weining Shen, Wenbin Lu, Xingye Qiao, Jiayang Sun, Wei Sun, Zhuqing Yu)

   Parametric modeling provides a convenient inferential setting but is very restrictive. On the other hand, in high dimensional data nonparametric models that impose only loose constraints such as smoothness require very large sample sizes for convergence of estimators and good inferential performance. Semi-parametric methods allow modeling of some parameters parametrically and others nonparametrically, often improving interpretability and performance of the nonparametric portion without sacrificing efficiency of the parametric portion. This topic explored methods for feature screening or selection prior to modeling as well as methods for joint modeling of the parametric and nonparametric parts of the model.

6. *Classification and Learning* (Andreas Artemiou, Guang Cheng, Qiyi Lu, Xingye Qiao, Wei Sun, Lingsong Zhang)

   This subgroup interacts with the semiparametric subgroup and the dimension reduction subgroup. We study the role of statistical learning methods, especially classification methods, in the massive data setting. There are several potential directions of interested, including:

   (a) Classification for very large data sets. The statistical properties for divide-and-conquer approaches.
   (b) Classification for massive data sets with many missing entries. Low rank approximation via nuclear norm. Matrix completion for the purpose of classification
   (c) Real time learning with increasing/evolving dimensions and increasing/streaming data sample sizes.
   (d) Dimension reduction.

7. *Computation for Complex Models* (David Banks, Fernando Bonassi, Nicholas Jarrett)

   This project addressed how to compare different approximations of complex astrostatistical phenomena. The example chosen was the n-body problem, which concerns prediction of the motion of n celestial bodies which have specified initial locations and momenta, and which are influenced only by gravitational forces. The full solution is a system of differential equations—these do not have (in general) a closed form expression, and must be solved numerically. Alternatively, from a statistical perspective, one can use Gaussian Process (GP) models, as fit to data by using Approximate Bayesian Computation (ABC). A third alternative is to employ Agent-Based Models (ABMs), in which each body is an agent that, at any given time steps, calculates its relation to all the other bodies and moves in the appropriate direction at the appropriate speed. With ABMs, the main concern is the size of the time step, which directly controls the accuracy for a given forecast horizon.

Work in progress

1. *Multiple Testing*

   (a) *Bayesian Large-Scale Multiple Testing for Time Series Data* (Ali Shojaie, Xia Wang, Jian Zou)

   In this project, we consider the problem of massive data multiple testing under temporal dependence. The observed data is assumed to be generated from an underlying two-state hidden Markov model. Bayesian methods are applied to develop the testing algorithm by optimizing the false negative rate while controlling the false discovery rate which is comparable to Sun and Cai (2009). Simulation studies show the similarity and the difference between the EM approach used in Sun and Cai (2009) and the Bayesian approach when the alternative has a simple or a mixture distribution. The model is also applied to the Influenza-like illness (ILI) data.

(b) *Testing the Equivalence of Outputs of Climate Models* (Naomi Altman, Dorit Hammerling, Wei Sun)

Multiple climate models are used to evaluate anthropomorphic impacts and for input into other large scale models. To evaluate and calibrate the models, it is necessary to determine locations at which they produce significantly different predictions. The spatial correlation of the predictions and the variance in prediction output both need to be considered. Sun's 2001 tube method for simultaneous confidence intervals is under consideration for this problem.

(c) *Meta-analysis of Multiple Tests* (Naomi Altman, Karen Kafadar)

In bioinformatics studies it is common to test two or more sets of differential outcomes on the same parameters and then present the set of parameters with significant outcomes as a Venn Diagram. Investigators would then like to make statements such as "significantly more parameters are non-zero in one of the studies compared to the others". To make such a statement with statistical support, it is necessary to have a probability model for the outcomes when there is no difference among the studies, and then develop a framework for determining when an outcome has a small probability or posterior odds under the null model. We are developing a frequentist framework for testing the entries in a Venn diagram under conditions that are more general than the unrealistic assumption that the outcomes of the studies are independent, which is often used as the null.

(d) *Two-stage Multiple Tests* (Naomi Altman, Karen Kafadar)

Many studies in bioinformatics, imaging and other disciplines have very high dimensional response and test more than one parameter for each dimension. For example, it is common to use the same experimental designs in these studies that have been used classically for a single response. If there are m variables, and the design has r degrees of freedom for the treatment, then there are rm tests. Typically, for a single response variable, an omnibus test of the r hypotheses, such as an overall F-test in analysis of variance or a chi-squared test for 2-way tables is done first. This is followed up by other tests using the r individual degrees of freedom, such as contrasts or goodness of fit of cells in the table.

In the multiple testing situation, if only π_1 of the overall omnibus tests are significant, then only $m + \pi_1 mr$ tests need to be done, which is a considerable saving in terms of multiplicity if π_1 is small. On the other hand, both false positives and false negatives are introduced at the initial stage of testing. We are building on the work of Jiang and Doerge (2006) which proposes a two-stage testing procedure for high dimensional analysis of variance.

(e) *Multiple Testing for Discrete Data* (Naomi Altman, Isaac Dialsingh)

In high dimensional testing problems π_0, the percentage of null hypotheses that are true is an important parameter. When the test statistics are continuous, the p-values from the truly null hypotheses come from the continuous uniform distribution on $(0, 1)$ and this assumption is key to many estimators of π_0. However, when the test statistics are discrete, the p-values from each truly null hypothesis comes from a discrete distribution with finite support with a positive point mass at $p = 1.0$. Often the null distribution depends on an ancillary statistic such as a table margin that might vary among the test statistics. Accordingly, methods for estimating π_0 developed for continuous test statistics may not perform well when applied to discrete testing problems. This project introduces a number of π_0 estimators that perform well with discrete test statistics and also assesses how well methods developed for continuous tests perform with discrete tests.

2. *Graphical Covariance Inference under Noise* (Alan Lenarcic, Michael Mahoney, Ali Shojaie)

   We continued to work on and present a topic originally considered in SAMSI Graphical Models Workshop (2010-2011). We used UNC Killdevil resources to research covariance under noise estimation 20K node graphs of various hyperbolicity. We demonstrated our algorithm had good performance on tree decomposable graphs, and we also coded a new C++ version of a competing algorithm that, now reimplemented, has competitive performance. Given discussion with other workgroups we are tailoring our research to statistics journals this summer and running additional simulations.

3. *Dimension Reduction*

   (a) *Computing Bregman SVD* (Naomi Altman, Wei Luo, Garvesh Raskutti)

   Collins, Dasgupta and Schapire (2001) suggested using Bregman divergence associated with the likelihood of exponential family distributions as a means of generalizing the singular value decomposition (SVD) to discrete and other non-elliptical data. They suggested a generalized power method to compute the Bregman SVD.

   We have found that this method is slow and often fails to converge, particularly when the eigenvalue gap is small. We have implemented two algorithms both of which appear to improve both speed and convergence. The first algorithm is simply an improved generalized power method and as such has the same theoretical properties as the generalized power method. The second algorithm extracts the singular vectors sequentially. Although this algorithm is much faster and appears to have improved convergence compared to the power method, our theoretical computations to date suggest that it need not converge to the correct answer.

   (b) *Generalizing PCA and SVD* (Naomi Altman, Wei Luo, Garvesh Raskutti)

   Principal components analysis (PCA) and the singular value decomposition (SVD) are both associated with L_2 loss and as such appear to be most appropriate for elliptically distributed data. Placing these methods in the context of maximum likelihood and regression provides a framework for extensions to other distributions and to nonlinear dimension reduction methods. PCA has been shown to provide the maximum likelihood estimator for the factor space in a Normal isotropic factor analysis model (Tipping and Bishop, 1999).

   We extend this to elliptical families. SVD is well-known to provide the latent regressors with smallest squared residual in the L_2 matrix nearness problem. This formulation has been used to suggest methods for robust and sparse PCA. We are working on generalizing these two ideas as a means both of unification of a number of current extensions such as kernel PCA and Bregman SVD and heuristic to suggest other useful extensions of this dimension reduction method.

(c) *Principal Support Vector Machine for Dimension Reduction using Distance-weighted Discrimination* (Andreas Artemiou and Xingye Qiao)

Principal Support Vector Machine (PSVM) is a useful method for sufficient dimension reduction. The basic idea is to slice the response variables into bins and use a modified form of support vector machine to (SVM) find the optimal hyperplanes that separate them. These optimal hyperplanes are then aligned by the principal components of their normal vectors. However, it has been shown that the SVM method displays data piling issue in the high-dimensional, low-sample size setting, which leads to incorrect approximation to the true mean difference directions between bins. We consider replacing the SVM classifier in the PSVM method by the Distance-weighted discrimination (DWD) classifier or its variants to address the data piling issue.

4. *Inference after Dimension Reduction: Post-Selection Inference on Polynomial Regression* (Wonyul Lee, Yufeng Liu and Kai Zhang)

Polynomial regression is an important method to model nonlinear relationships. In common practice, the fitted curve is achieved through a selection process. Due to the dependence on the selection process, inference conditional on the selected model can be invalid. In this project, we propose a new method to perform statistical inference on polynomial regression models, which achieves validity after the dimension reduction process. By taking advantage of the special structure of polynomial regression models, the inference can be much less conservative than some other existing methods. Our numerical studies demonstrate that our new inference method can outperform some existing post-selection inference methods.

5. *Semiparametric Modeling*

(a) *Local Polynomial Estimation of the Semi-nonparametric Models: Joint Asymptotic Studies* (Guang Cheng, Tao Yu)

The local polynomial method has been demonstrated as an efficient tool in the estimation of the partially linear model. The theoretical properties as well as the numerical performance of the local polynomial estimates for both parametric and nonparametric components have been extensively studied in the literature but in a separate manner, e.g., Carroll et al (1997). The research on the *joint* asymptotic behaviors of these estimates is extremely important, but still lacking. We consider two different estimation methods, namely the profile procedure and one-step procedure. We derive the joint limiting distributions for the estimates of the parametric and nonparametric components within the general quasi-likelihood framework.

(b) *Semiparametric regression in real time* (Guang Cheng, Xingye Qiao)

Wand et al. proposed a computational method for semiparametric models to handle data that arrive in separate batches over time. In their implementation, they update the model by Bayesian techniques. Our objective is to relax the strong assumption that the model is fixed for every batch. Rather, we believe it more reasonable that the model itself is dynamic in the sense that the parameter dimension should be also updated when the new batch of data comes.

6. *Classification and Learning*

(a) *Using Adaptively Weighted Large Margin Classifiers for Robust Sufficient Dimension Reduction* (Andreas Artemiou, Yufeng Liu)

Support Vector Machine (SVM) is a popular large-margin classifier. At the same time, sufficient dimension reduction is a powerful idea in dealing with high dimensional data. Recently Li, Artemiou and Li (2011) introduced Principal Support Vector Machine (PSVM), an algorithm which achieves linear and nonlinear dimension reduction under a unified framework by utilizing inverse regression and SVM ideas. Wu and Liu (2012) proposed adaptively weighted large margin classifiers for robust performance of classification in the presence of outliers. In this project we show how the idea of adaptive classifiers can be used in the Sufficient Dimension Reduction framework to improve the performance of PSVM with outliers.

(b) *Large-Margin Classifier Selection via Decision Boundary Stability* (Guang Cheng, Yufeng Liu, Wei Sun)

Large-margin methods have been widely used in classification. However, which classifier should be chosen in practice remains to be an open question. The existing criterion for classifier comparison is merely on competing their Generalization Errors (GE) or excess risks. However, this criterion does not consider various sources of variation in prediction.

We introduce a novel concept of Decision Boundary Instability (DBI) and incorporate it into the classifier selection algorithm in order to obtain an accurate and stable classifier. Specifically, we propose a two-stage classifier selection algorithm. In the first stage, the potentially good classifiers are initially selected as those having relatively small GEs. This is achieved by constructing the confidence interval for the difference of one classifiers GE and the minimal GE. If the confidence interval contains 0, this classifier is potentially good. In the second stage, the optimal classifier is chosen as the most stable one, i.e., the one with smallest DBI, among the potentially good classifiers. The proposed two-stage selection algorithm is shown to be selection consistent in the sense that the selected classifier asymptotically obtains both minimal GE and DBI. Its effectiveness is illustrated in various simulated examples and real data sets.

(c) *Classification Error and Classification Stability in the Weighted Nearest Neighbor Classifier* (Guang Cheng, Xingye Qiao, Wei Sun, Xiang Zhang)

We investigate the relationship between classification error and classification stability in the weighted nearest neighbor classifier. Our goal is to deliver some theoretical insight on the error and stability for classification.

(d) *Variable Selection of Support Vector Machine in High Dimensions* (Xiang Zhang, Yichao Wu)

Support Vector Machine (SVM) is a popular classification tool. However, it selects all variables and can perform poorly in high dimensional space due to noise accumulation. In this research we address the variable selection problem of SVM and prove the oracle property of our procedure. We show that under weak conditions, for a general class of nonconvex penalty, one of the local minimizer of nonconvex penalized SVM is the oracle estimator, that is, we estimate the model as if the true model is known in advance. We also provide a non-asymptotic lower bound for the probability of identifying the oracle from possibly multiple local minima. Furthermore, we give sufficient conditions under which the oracle is found by local linear approximation algorithm with probability tending to one.

(e) *Distance-weighted Support Vector Machine* (Xingye Qiao, Lingsong Zhang)

A novel linear classification method that possesses the merits of both the Support Vector Machine (SVM) and the Distance-weighted Discrimination (DWD) is proposed in this article. The proposed Distance-weighted Support Vector Machine method can be viewed as a hybrid of SVM and DWD that finds the classification direction by minimizing the DWD loss, and determines the intercept term in the SVM manner.

We show that our method inheres the merit of DWD and overcomes the data-piling and overfitting issue of SVM. On the other hand, the new method is not subject to imbalanced data issue which was a main advantage of SVM over DWD. It uses an unusual loss which combines the Hinge loss (of SVM) and the DWD loss through a trick of axillary hyperplane. Several theoretical properties, such as Fisher consistency and asymptotic normality of the DWSVM solution are proved. We use some simulated examples to show that the new method can compete DWD and SVM on both classification performance and interpretability. A real data application further establishes the usefulness of our approach.

7. *Simulation in Complex Models* (David Banks, Fernando Bonassi, Nicholas Jarrett)

   This project addressed the n-body problem, which concerns prediction of the motion of n celestial bodies which have specified initial locations and momenta, and which are influenced only by gravitational forces. For the n-body problem the full solution is a system of differential equations—these do not have (in general) a closed form expression, and must be solved numerically. Alternatively, from a statistical perspective, one can use Gaussian Process (GP) models, as fit to data b y using Approximate Bayesian Computation (ABC).

   A third alternative is to employ Agent-Based Models (ABMs), in which each body is an agent that, at any given time steps, calculates its relation to all the other bodies and moves in the appropriate direction at the appropriate speed. With ABMs, the main concern is the size of the time step, which directly controls the accuracy for a given forecast horizon. We are able to very accurately reproduce the ABM output using the modified GP. Advantages of the GP approximation to the ABM include greater computational speed and the ability to perform multi-step transitions without dealing with intractable systems of differential equations.

   Due to the complexity of the posterior distributions, ABC plays a role in the model calibration; essentially, there is no analytic form for the likelihood function. We also developed an ABC-based method to enhance the ABM simulation. Our analyses found promising and complementary results for both the GP models and the ABMs. Our contribution provides a case study for a classical problem of significant numerical difficulty, and its innovations are to (1) compare different emulation strategies, (2) develop ABC methodology for tuning the simulation, and (3) show how computer-intensive ABM simulation can train a simpler and faster GP model that is still acceptably faithful to reality.

Other on-going work performed at SAMSI

1. *Semi-supervised Sparse Fisher Linear Discrimination* (Qiyi Lu, Xingye Qiao)

   In this project, we consider Sparse Fisher Linear Discriminant (SparseFLD) analysis in the semi-supervised learning setting. The high-dimensional, low-sample size setting has brought some challenges to the classification problem. However, large amounts of partially unlabeled data suggest a promising approach to these issues. Firstly, the global information in the unlabeled data can be properly used to enhance the estimation of the parameters in SparseFLD. Secondly, an alternative objective function is considered which counts for the clustering performance. Lastly, we consider an iterative framework for semi-supervised classification which reinforce the performance of the Semi-Supervised SparseFLD.

2. *Significance Analysis of Multi-class High-Dimensional, Low-Sample Size Data* (Qiyi Lu, Susan Wei, Xingye Qiao)

   The main goal of this project is to study a statistical test which can test equality of means and/or equality of distributions of several classes for a given variable set. In practice, permutation tests are often used for the purpose of testing the class difference, where the null distribution is mimicked by the empirical distribution of the statistic calculated from randomly relabeled data sets.

   However, for high-dimensional data, direct permutation with an arbitrarily-chosen distance measure may not work. This is because when the dimension is greater than the sample size, such a distance measure will be mainly driven by the error aggregated over dimensions, rather than by the true mean differences between classes. To address this issue, a three-step procedure called Direction-Projection-Permutation test (DiProPerm) was proposed and studied for the two class binary setting. We direct our interest to testing the significance difference between multiple classes. The new test is called multi-class DiProPerm (mDiProPerm) test. We compare mDiProPerm with multiple binary DiProPerm tests and other methods, and try to show the conditions under which our tests are valid.

3. *Self-Modeling Generalized Regression* (Naomi Altman, Kalyan Das)

   Self-modeling regression is a class of semi-parametric models for functional data in which the functions are modeled as parametric transformed versions of a baseline nonparametric curve. Altman and Villarreal (2004) worked out the computational details for fitting the model using penalized regression splines for the nonparametric curve along with asymptotic inference for the parametric portion of the model for continuous response. This work extends the model to exponential families by use of generalized linear mixed models and the EM algorithm.

4. *NMR-based Metabolite Profiling of Pancreatic Cancer* (Kwadwo Owusu-Sarfo, Vincent M. Asiago, Siwei Wei, Narasimhamurthy Shanaiah, G. A. Nagana Gowda, Bowei Xi, Elena G. Chiorean, and Daniel Raftery)

   Investigations of serum metabolite profiles in pancreatic cancer (PC) patients were made using 1H nuclear magnetic resonance (NMR) spectroscopy with a focus on the identification of metabolite biomarkers associated with PC pathology. It was found based on univariate and multivariate logistic regression analysis of NMR data of serum from 78 PC patients and 48 healthy controls, that nine metabolites: alanine, histidine, glutamine, valine, citrate, creatinine, formate, glucose and lactate, distinguished PC from controls.

   A cross-validated regression model built using these metabolites from one batch of samples (55 PC; 32 controls) classified the cancer and control groups with an area under the receiver operative characteristic curve (AUROC) of 0.94. Validation of the model using an independent batch of samples (23 PC; 16 controls) provided an AUROC of 0.86. This distinction of pancreatic cancer patients from healthy controls based on serum metabolite profiles demonstrates the potential for metabolite markers to identify patients with pancreatic cancer.

5. *Fast Least Squares Algorithms for Estimating and Monitoring Network Link Losses Based on Active Probing Schemes* (George Michailidis, Vijayan N. Nair, Bowei Xi)

There is considerable interest currently within the network community on tools for estimating and monitoring network utilization and performance. This is partly due to increasingly complex services, such as videoconferencing and IP telephony, that are provided by internet service providers and that require high quality of service guarantees. Estimation and monitoring of QoS parameters are also important for various network management tasks, such as fault and congestion detection, ensuring service level agreement compliance, and dynamic replica management of Web services. The goal of this paper is to develop fast algorithms for estimating and monitoring link-level loss rates in a network with a general topology.

6. *Hypothesis Tests and Classification with Gaussian Mixture Models under Differential Privacy* (Bowei Xi, Murat Kantarcioglu, Xiaosu Tong, Ali Inan)

   A lot of statistical inference over a specific data-set requires obtaining some very basic statistics such as the mean vector and the covariance matrix of the data. This is often a straightforward task. However, when the data-set in question contains sensitive information, special care has to be taken in order to prevent direct access to the data, even for research purposes. Instead of granting direct access, the data users or the researchers are provided with a sanitized view of the database containing private information. We examine the effect of basic statistics generated under differential privacy mechanism for classification and hypothesis tests.

7. *Statistical Analysis and Modeling of Mass Spectrometry Metabolomics Data* (Bowei Xi, Haiwei Gu)

   This is a peer reviewed book chapter, discussing the statistical and machine learning methods used in metabolomics study.

4.2 Imaging

This WG focused on challenges to statistics, mathematics and computer science, arising from imaging science (which is interdisciplinary itself) and related areas. The application and motivation areas of the imaging and related problems include astronomy, geology, neuroscience and medical sciences.

The mission of this WG is to bring together statisticians, mathematicians, computer scientists and domain scientists to

1. Exchange ideas and discuss challenges in modern imaging,
2. Collaborate on imaging research topics, and
3. Learn from each other, in a sustained way.

WG members: Andrea Bianchi* (Physics, Federal University of Ouro Preto, Brazil), Julian Faraway (Stat, U. Bath, UK), Weihong Guo* (Case Western), Hari Khrishnan (Lawrence Berkeley National Lab), Ashish Mahabal (Astro, Caltech), Joerg Meyer (Lawrence Berkeley National Lab), Kamran Paynabar (ISE, Georgia Tech) Jiayang Sun* (Stat, Case Western), Daniela Ushizima* (Lawrence Berkeley National Lab), Xiaofeng Wang (Biostat, Cleveland Clinic Lerner Research Institute), Lingsong Zhang (Stat, Purdue University),

Postdocs: Garvesh Raskutti (SAMSI & STOR, UNC), Yi Grace Wang* (SAMSI & Math, Duke), Dan Yang* (SAMSI & STOR, UNC)

Members active only through Fall semester 2012: Jogesh Babu (Stat, Penn State), James Nagy (Math & CS, Emory), Prajval Shastri* (Astrophysics, IIA Bangalore, India), Erkki Somersalo (Math, Case Western)

Publications somewhat related to work done at SAMSI

1. I.C. Paula Jr., F.N.S. Medeiros, F.N. Bezerra, D.M. Ushizima, *Multiscale Corner Detection in Planar Shape*, Journal of Mathematical Imaging and Vision, March 2013
2. D.M. Ushizima, D. Morozov, G.H. Weber, A.G.C. Bianchi, E.W. Bethel, *Augmented topological descriptors of pore network*, IEEE Trans. Comp Graph. and Vis., 2012
3. L. S. Bruckman, N. R. Wheeler, Junheng Ma, E. Wang, C. K. Wang, I. Chou, Jiayang Sun, and R. H. French, *Statistical and domain analytics applied to PV module lifetime and degradation science*, IEEE Access, June 2013
4. D.M. Ushizima, A.G.C. Bianchi, W. Guo, *Characterization of MRI brain scans associated to Alzheimer's disease through texture analysis*, International Symposium on Bio Imaging (ISBI), April 2013
5. D.M. Ushizima, A.G.C. Bianchi, C. deBianchi, E.W. Bethel, *Material science image analysis using Quant-CT in ImageJ*, ImageJ User and Developer Conference, Luxembourg, October 2012

Conference and Workshop Organization and Presentations
Contributed session on *Challenges and New Developments in Imaging with Large Data Sets*, Joint Stat. Meetings, Montreal, Canada, August 2013
Sponsoring Societies: Section on Statistics in Imaging , Mental Health Statistics Section , Section on Statistical Computing
Organizers: Garvesh Raskutti, Dan Yang, Jiayang Sun
Chair: Timothy Johnson (Biostat, U. Michigan)

1. *Compressive Inference*
(*Weihong Guo*, Garvesh Raskutti, Jiayang Sun, Grace Yi Wang, Dan Yang)

2. *Light Curve Analysis for Classification with Astronomical Data*
(*Ashish Mahabal*, Julian Faraway, Jiayang Sun, Xiaofeng Wang, Yi Wang, Lingsong Zhang)

3. *Forgery Detection in Paintings*
(*Yi Wang*, Ingrid Daubechies, Gungor Polatkan, Sina Jafarpour)

4. *Image Analysis of High-Resolution and High-Throughput Experiments*
(*Daniela Ushizima*, Andrea Bianchi, Hari Krishnan)

5. *Predictive Modeling with High-Dimensional Colorimetric Image Data for Lung Cancer Detection* (*Xiaofeng Wang*, Peter Mazzone)

The most significant transformation, beyond the new collaboration ties and research topics, from this SAMSI experience for some of us is *breaking out of the traditional working box.* Some of us are getting into the habit of working without a limitation of geographical region, in person or virtually, using the modern teleconferencing and social media.

Topics investigated by WG

T1: Compression inference

T2: Multi-frame blind deconvolution and optimization

T3: Segmentations

T4: Study of propagated errors and development of fast de-convolution algorithms, with application to the detection of changes in brightness of stars over time

T5: Statistical methods for image data clustering and classification

T6: Separation of object types, morphological feature detection in astronomical images

T7: Learning false positives from the FDR map

T8: Exploring the integration of images into a regression framework

T9: Statistical methods for detecting lung cancer via a new breathing/imaging device

T10: Computational methods for scaling image representations, processing, and analysis.

These topics evolved into three subgroups with research and developments described below.

1. *Compressive Inference* (Weihong Guo, Garvesh Raskutti, Jiayang Sun, Grace Yi Wang, Dan Yang)

   Advancement of imaging science has allowed us to obtain a large amount of or high-quality images from nanoscales to astronomical scales, which lead to new challenges in processing and analyzing these images. In order to approach the methods capable of reducing data collection (so that the efficiency is gained at the beginning of an investigation) while guaranteeing the resulting quality, this subgroup focused on compressive sensing (CS) algorithms.

   Applications of CS range from astronomy, medical imaging, to sensor networks, especially when dealing with massive data. In medical applications, less data also means less radiation and hence reduces the risk of developing cancer. A challenge to *Compressive Inference* is: how to make statistical inferences about the underlying images from their few incomplete measurements, such as *compressive support detection* and *compressive change detection*.

   Denote f: $\mathcal{X} \to \mathbb{R}$, the intensity function of an image on an image domain $\mathcal{X} \subset \mathbb{R}^d$. Let $\mathbf{f} = (f(x_1), f(x_2), \cdots, f(x_p))$ be the vector formed by the values of f at grids $x_j, j = 1, \ldots, p$. In compressive sensing, the observed data can be modeled from:

   $$\mathbf{y} = A\mathbf{f} + \epsilon$$

   where $A \in \mathbb{R}^{n \times p}$ is a sampling matrix with $n \ll p$, satisfying the so-called *restricted isometry* property, and $\epsilon \sim N(\mathbf{0}, \sigma^2 I_n)$.

   Compressive support detection is the problem of finding the support of a continuous intensity function f in its domain from incomplete discrete measurements $\mathbf{y}$. Thus, statistically, we can 1) test H_0: $f(x) = 0, \forall x \in \mathcal{X}$; and 2) if H_0 is rejected, further derive the support of f: $\{x | f(x) \neq 0\}$ by multiple comparisons.

   Compressive change detection is the problem of detecting changes from one image f_1 to another image f_2 from their incomplete compressive sensing measurements $\mathbf{y_1} = A\mathbf{f_1} + \epsilon_1$ and $\mathbf{y_2} = A\mathbf{f_2} + \epsilon_2$. By taking differences on both sides of the two equations, we get $\mathbf{y_2} - \mathbf{y_1} = A(\mathbf{f_2} - \mathbf{f_1}) + \epsilon_2 - \epsilon_1$. Hence, a test of "$f_1 = f_2$" is equivalent to the test of "$f_1 - f_2 = 0$", by transforming the compressive change detection to the compressive support detection problem.

   This work is in progress and continuing, with promising ground-breaking work done and is expected to have multiple outcomes. Our proposed method involves the following two steps: (a) Estimate $\mathbf{f}$ by a suitable kernel ridge regression; (b) Test H_0 by multiple comparisons using the tube method. This method is by far the first work to make inference on continuous functions from discrete incomplete measurements in imaging. The algorithm for 1D signals has been developed and its comparison with existing methods shows advantages of the proposed method. We will extend it to 2D images, and 3D volumes, and test its performance for real applications.

2. *Image Classification and Clustering* (Julian Faraway, Ashish Mahabal, Jiayang Sun, Xiaofeng Wang, Yi Grace Wang, Lingsong Zhang)

   Clustering and classification are extremely important for classifying transients, separating stars from galaxies, and for identifying tumors and cell types for targeted and personalized medicine. The objective of this subgroup is to classify (in real time) astronomical data (images, light curves, or derived statistics) to different *transient* groups and a big *non-variable* group. This requires developments of new set of measurements (called derived statistics by astronomers), methodologies and algorithms.

This subgroup investigated multiple topics. This is a truly successful collaboration by scientists from statistical and mathematical sciences and domain sciences. New methods are developed, new research directions are formulated and resulting procedures will have applications beyond astronomy. The topics include:

(a) *EDA*
Exploratory data analysis on the *whole* light curves and their derived statistics is an important first step for a sound development of modern classification procedure. In this part, new useful measures are derived from better fitted models to the light curves, and residuals of the fits; they are said to have already been implemented in some astronomy super computers. Multi-scale exploration of the light curves are conducted. Advanced multivariate exploration of the existing (Richard's) and newly derived statistics are performed by projection pursuit to search for useful combinations for efficient classification and dimension reduction. Useful transformation of the data are proposed for better classification.

(b) *Classification based on the derived measures*
In this part, various classification approaches are programmed to analyze the derived measures. Better classification rates have been obtained than either using the previously existing Richard's measures alone. We propose developing an ensemble of classification methods for this type of large data, allowing incremental learning and updates in the future.

(c) *Methodology development*
A novel nonparametric regression method incorporating known variance is proposed. Gaussian Process Regression approach is adapted for model fitting.

(d) *Additional methodology development*
We propose to develop

  i. PF-classification approach for classification of light curve data, and
  ii. Functional centroid classification (FCC) methods for classification of light curve data

For most topics, all members participated in discussion. Sun, Faraway and Zhang worked on EDA; Sun and Faraway were leading the new derived statistics. Sun and X. Wang were leading the known-variance nonparametric regression. Faraway worked on the Gaussian Process Regression approach. Faraway, Zhang, G. Wang and Sun worked on classification. Zhang, Sun and X. Wang will work on the PF-classification and FCC method.

All papers are in the preparation stage. We have code in web pages and in dropbox.com, which will be made available to public after the papers are published.

3. *Image Analysis and Scientific Computing* (Andrea Bianchi, Weihong Guo, Hari Khrishnan, Jiayang Sun, Daniela Ushizima)

The goal of this subgroup was to investigate methods for semi-automatic (with user interaction) and automatic image segmentation methods applied to large image data (by Ushizima), including adapted versions of fast marching, statistical region merging and other machine-learning based algorithms for image partitioning.

Other important topics in this group included protocols for image segmentation benchmarking (by Bianchi) with standard methods (e.g. thresholding), and theoretical foundations in image segmentation, e.g. using compressive inference (by Guo) and mixture distributions (by Sun and Guo).

In this subgroup, substantial progress has been made. We intend to build upon our preliminary results reported in ISBI 2013 conference.

4.3 Environment & Climate

The mission of this WG was to address problems in the analysis of massive data sets used in environmental and climate science. Members of the group narrowed this down to topics of mutual interest described below. Since the group included NASA and NOAA representatives in addition to more traditional university participants, two of the topic areas focussed on statistical problems in areas that are less traditional for statistical methodology research: data analysis in the context of distributed data sources and operational environments, and "crowd-sourced" science analysis. These two areas also emphasized the use of remote sensing data, which are relatively new targets for mainstream statistical research.

The WG sought to identify:

1. Areas where statistical thinking and approaches have not been fully exploited to obtain maximum scientific return from massive datasets, or where potential exists to improve existing methodology, and

2. New statistical research problems motivated by modern data collection, distribution, and analysis technologies such as satellites and distributed data services via the Internet.

Our intention was delineate a research agenda in these areas and begin work on them that would continue beyond the end of this program per se.

WG members: Amy Braverman* (JPL), Jessica Matthews* (NOAA NCDC), Richard Smith (STOR, UNC), Peter Thorne (NOAA NCDC), Yong Wong (Stat, Eastern Kentucky U.)

Postdocs: Dorit Hammerling* (SAMSI), Matthias Katzfuß (University of Heidelberg, Germany), Elizabeth Mannshardt* (Stat, NCSU)

Papers published

1. B.D. Santer, J.F. Painter, C.A. Mears, C.A. and 18 others , *Identifying human influences on atmospheric temperature*, PNAS, 2012 `www.pnas.org/cgi/doi/10.1073/pnas.1210514109`

Grant proposals

1. Pre-proposal in response to DOE's call for *Mathematical and Statistical Methodologies for DOE-Data Centric Science at Scale*, declined

   We proposed to develop and prototype a framework for a new paradigm to analyze massive, distributed scientific data collections that leverages distributed architectures and statistical methods designed for them.

   We intend to pursue an expected proposal opportunity from NASA's Research Opportunities in Space Earth Sciences (ROSES) in early 2014.

Topics investigated by WG

1. *Analysis of distributed spatial data* (Amy Braverman, Dorit Hammerling, Matthias Katzfuß)

   Because bandwidth is increasing at a slower pace than the capacity to generate and store data, it is becoming infeasible in many cases to move large datasets, and statistical analyses often have to be carried out where the data reside. If several datasets stored in separate places are all relevant to a given problem, the challenge is to obtain valid inference based on all data without moving the datasets. This distributed data problem arises in the spatial and environmental sciences, for example, for measurements of total precipitable water made by three different satellite systems.

We have shown that for a very widely used class of spatial low-rank models, which contain a component that can be written as a linear combination of spatial basis functions, computationally feasible spatial inference and prediction for massive distributed data is possible. The required number of computations is linear in the number of data points, while the required bandwidth that does not depend on the data sizes at all. We have recently extended this approach to spatio-temporal problems.

2. *Detection and attribution* (Dorit Hammerling, Matthias Katzfuß, Richard Smith)

   "Detection and attribution" refers to a set of widely used statistical techniques in atmospheric science that are critical to deciding the extent to which climate change is human caused. In a typical analysis, an observational data vector is regressed on a series of "climate signals" representing different forcings of the climate, such as greenhouse gases, other anthropogenic influences such as atmospheric particles, and natural forcings such as solar variation and volcanoes. Control model runs (runs of a climate model in stationary conditions without any forcings) are used to assess the error terms in the regression.

   The problem is complicated because in a typical application, both the X and Y variables of the regression analysis are high-dimensional, so in practice, they are reduced through an empirical orthogonal functions (EOF) expansion. However, the effect of an EOF transformation on the resulting regression introduces further courses of error that have never been fully characterized. An additional complication in some recent detection and attribution studies is that the observations themselves are not treated as fixed, but reconstructed from some observational model that includes unknown parameters. This raises the possibility of varying the unknown parameters in a systematic way to create an ensemble of observational datasets, but how to incorporate that into the traditional detection and attribution algorithms has not been studied.

   The new approach developed within the WG uses a Bayesian hierarchical model formulation to incorporate these sources of uncertainty as well as those in the traditional approach. A theoretical approach has been developed and code written to implement it. The proposed application will apply these methods to a dataset concerned with temperature trends in the lower troposphere, which has been widely studied in recent literature (in particular, Santer *et al.*, 2012). In collaboration with climate scientists Ben Santer and Carl Mears, we have collected a dataset that includes an ensemble of climate model runs as well as an ensemble of observational reconstructions. During the summer of 2013, we plan to complete the analyses and submit a paper for publication to *Journal of Climate* or some similar publication (Hammerling *et al.* 2013).

3. *Climate Mortality* (Tamara Greasby, Kenny Lopiano, Richard Smith)

   There has been extensive discussion of the role of climate change on human health, including the direct influence of temperature increases on mortality. However, research has been limited on projecting future mortality changes, using climate models to project future temperature changes and then temperature-mortality relationships to project mortality.

   In this study, we combine observational temperature and mortality data from a well-known epidemiological dataset, the National Morbidity, Mortality and Air Pollution Study

(NMMAPS), with climate model data from a regional climate model study, the North American Regional Climate Change Assessment Program (NARCCAP). Initial analyses have shown a strong (though nonlinear) relationship between temperature and mortality, and a clear potential for future increases in mortality if temperature rise at a uniform rate. However, projecting future increases using climate models requires some attention to biases between climate model outputs and ground observations.

Preliminary work was presented at the ENAR meeting in Orlando (March 2013), and is currently being written up as a paper (Smith *et al.* 2013). The plan is to complete the analysis and submit a paper by the fall of 2013.

4. *Temperature records* (Elizabeth Mannshardt, Jessica Matthews, Richard Smith, Peter Thorne)

   In June 2012, associated with its monthly state of the climate report, NCDC reported a string of 13 consecutive upper tercile months in the CONUS temperature record. Whilst making clear that this was a naive calculation the straight probability was reported as 1 in 1.3 million. This statistic was subsequently repeated, without the contextual caveat across numerous media outlets causing somewhat of a blog storm and requests from senior NCDC management for AR(1) clarifications.

   Peter Thorne provided these while pointing out that the CONUS record was patently not an AR(1) process. He approached Jessica Matthews for help in determining the true probability. As a result of subsequent discussions there was a teleconference with Richard Smith that instigating a multi-author multi-approach effort to ascertaining realistic estimates of the probability of such a string occurring. At time of writing analysis is ongoing and being prepared for submission to JGR.

5. *Citizen science* (Dorit Hammerling, Carl Schreck)

   Cyclone Center is a unique web interface that invites citizen scientists to help analyze satellite imagery of hurricanes and other tropical cyclones. The project was introduced by Carl Schreck (CICS-NC) as part of the weekly WG meetings. We have started discussions on a possible collaboration to develop an optimal method for evaluating and combining the classifications made by these citizen scientists. This collaboration could result in a publication describing the methodology to obtain on overall classification of the tropical cyclones from individual classifications, as well as a data product of the citizen science classifications.

6. *Modeling multivariate spatial processes* (Yong Wang)

   One key feature of multivariate spatial data is the cross covariance function which, if modeled appropriately, could improve spatial interpolation. Wang and Zhang proposed an approach to constructing the multivariate covariogram which involves a parametric form of marginal covariogram and a semi-parametric cross-covariogram. The cross-covariogram is designed in a special structure that guarantees the validity of the multivariate covariance function. This approach offers the flexibility to accommodate any given marginal covariogram model and much improved predictive performance.

   A recent improvement to this approach takes advantage of the eigen-structure of the marginal covariance matrices to achieve dimension reduction and therefore the ability to handle massive spatial data if it is combined with low-rank models. Wang and Zhang have shown that this approach is compatible with the theoretical framework of many existing multivariate spatial models and results in simpler inferences. Application of this approach to several real data examples has also demonstrated promising predictive performance.

Anticipated future publications directly resulting from WG research

1. Mathias Katzfuß and Dorit Hammerling, *Statistical inference for massive distributed spatial data using low-rank models*
2. Amy Braverman, Dorit Hammerling, Matthias Katzfuß, *Analysis of distributed spatial data*
3. D.M. Hammerling, M. Katzfuß, R.L. Smith, B.D. Santer, C.A. Mears, P. Thorne, P. *A hierarchical statistical model for regression-based climate-change detection and attribution*, to be submitted to Journal of Climate
4. R.L. Smith, T. Greasby, K. Lopiano, *Projecting future temperature-related mortality using regional climate models*
5. P.F. Craigmile, P. Guttorp, R. Lund, E. Mannshardt, J. Matthews, R.L. Smith, P. Thorne, *Assessing the significance of the recent run of warm months in the US national temperature record*, to be submitted to Journal of Geophysical Research
6. Yong Wong, *Modeling multivariate spatial processes*

4.4 High Energy Physics

For decades, high-energy physics has developed a “Standard Model” regarding elementary particles and their interactions. The parameters in this model are estimated and verified by high energy physics (HEP) experiments which produce enormous quantities of data.

For instance, the ATLAS collaboration at the Large Hadron Collider (LHC, CERN) saves 320 MB of data per second after having discarded over 99.99% of the “uninteresting” data. Understanding and analyzing these data require the use of well calibrated models of the underlying physical processes alongside sophisticated data processing, extraction, and manipulation methods. Historically analysis of HEP data has focused on inference for single observables, such as a particle mass or the value of a given constant, over more global analyses. This working group provided a forum for the development of methods for handling, estimating parameters, and approximating models from massive data, to enable new global analyses of many observables over a vast parameter space.

The importance of accurate and precise estimates of these model parameters is reflected in the need for better understanding of the fundamental aspects of particle physics as well as revealing hitherto unknown correlations.

Progress in this direction required the calibration and bias quantification of large-scale HEP models, such as event generators and hydrodynamical simulations, along with the development of novel strategies for observable selection and model to data comparisons which can be carried out in finite computational time.

Topics commensurate with the mission of the WG are:

- Analysis of particle collision data from the LHC.
- Understanding the role of computer models in the current HEP discovery process.
- Potential applications of methods from machine learning, data mining, pattern recognition, and statistical inference on large scale HEP data sets, to extract possibly interesting signals from mountains of noise.
- Quantifying biases in HEP data analysis arising from experimental apparatus (detectors), model approximations from simulated data, and other systematic errors.
- Parameter estimation, observable selection and correlation estimation for very large scale models, including many parameters and many possible observables.
- Aspects of model to data comparisons for very large multi-dimensional data sets: efficient implementations and approximations.
- Methods for comparing models based on very different physics.

WG members: Steffen Bass (Physics, Duke), Karen Kafadar* (Stat, U. Indiana), Luc Demortier (Physics, Rockefeller U.), Robert Wolpert (Stat, Duke), Snehalata Huzurbazar* (Stat, U. Wyoming)

Graduate Students: Chris Coleman-Smith (Physics, Duke), Peter Marcy (Stat, U. Wyoming)

Follow-up activities: The working group will continue to meet on a monthly basis to work on these topics/projects to bring them to a successful conclusion.

Conference presentations

1. Steffen A. Bass, *A Decade of Quark-Gluon-Plasma Physics: what have we learned and what does the future hold?*, Invited lecture at the Kobayashi-Maskawa Institute for the Origin of Particles and the Universe (KMI), Nagoya, Japan, 24 Sept 2012

2. Steffen A. Bass, *What do we know about the viscosity of QCD matter?*, Lecture at the SERC 2013 School at the Variable Energy Cyclotron Centre, Kolkata, India, 1 January 2013

3. Karen Kafadar, *Statistical Methods for Analyzing Massive Data in Genomics and High-Energy Physics*, University of Virginia, Charlottesville VA, 22 March 2013

4. Karen Kafadar, *Statistical and Visualization Methods for Analyzing Massive Data Sets*, Lecture at Utah State University, Logan, UT, 4 April 2013

Topics investigated by WG

1. *Computer experiments when input factors might be correlated* (Karen Kafadar, Luc Demortier)

   A typical problem facing high-energy physicists is the quantification of biases and systematic errors that arise from the experimental hardware (detector). For example, in measuring the mass of a particular particle, say Y, physicists recognize that the measurement will be affected by several factors, such as the energy scale (A), a higher-order correction coming from the theory of quantum chromodynamics (QCD) (B), and levels of background noise (C).

   Typically, physicists will proceed by considering one central value (say, a_0, b_0, c_0) and 3 values for each factor, one at a time, and simulate Y at each of the 3 conditions, holding the other 2 factors constant at their central levels. From the 9 distributions, they then calculate an overall uncertainty in Y as a root-mean-square of the standard deviations from the simulations when varying the 3 factors separately.

   If the input factors A, B, C are independent, this process is a reasonably sensible approach to quantifying the uncertainties in Y as a function of the uncertainties in these factors. If A, B, C are **not** independent (for example, levels of QCD radiation could influence the background noise levels), then the method may return an RMS (root-mean-squared error) that under- or over-estimates the actual uncertainty in Y as a result of uncertainties in A, B, C.

   To handle the problem of potentially dependent input factors in a computer experiment, one could envision three approaches:

   1. Approach: Estimate (from other experiments) the joint distribution of (A, B, C), say $f_{A,B,C}$, and simulate $3n$ (here, 9) distributions of Y using 9 values of (A, B, C) generated from $f_{A,B,C}$ (this approach assumes great confidence in the estimate of the joint distribution $f_{A,B,C}$).

   2. Approach: Find a "realistic region" where (A, B, C) are likely to occur and select 9 values at random from this region to use as inputs for simulating distributions of Y.

   3. Approach: Use Latin Hypercube Sampling (LHS):

   (a) Divide range of A ($a_0 \pm \sigma_a$ or $a_0 \pm 2\sigma_a$) into n intervals of roughly equal density

(b) Choose one point in each of the n intervals: $a_1, ..., a_n$

(c) Repeat for B: $b_1, ..., b_n$

(d) Repeat for C: $c_1, ..., c_n$

(e) Run simulator at n settings (a_i, b_i, c_i), $i = 1, ..., n$

The question discussed concerned the relative advantages of these approaches and whether derivative information about $f(A, B, C)$ would improve the estimates of the uncertainties in Y due to these input factors.

2. *"v_n analysis": Fitting parameters from a parametric (GRW) distribution* (Peter Marcy, Chris Coleman-Smith)

   The scientific objective of this topic was to extract estimates of the shear viscosity of the quark-gluon plasma (QGP) produced in collisions of heavy nuclei via a comparison of computational models to real experimental data.

   Relativistic collisions of heavy ions are usually not directly head-on, but happen at a finite impact parameter, resulting in a spatially anisotropic overlap area for the created high energy-density matter, the quark-gluon-plasma (QGP). As the QGP expands like a near-ideal fluid, this initial spatial asymmetry translates into a momentum-space anisotropic elliptic flow. The azimuthal angular emission of produced particles emerging from the collision zone can be estimated via a Fourier expansion whose coefficients have meaningful physical interpretations. The second coefficient v_2 is proportional to the degree of elliptic flow, and therefore very sensitive to the initial geometry of the collision. The higher order coefficients are sensitive to fluctuations in the initial nuclear density and are somewhat insensitive to the initial geometry.

   Due to fluctuations in the configurations of the colliding nuclei, these flow coefficients fluctuate event-by-event. The distribution of these Fourier coefficients v_n, measured by experiments such as those at the Large Hadron Collider, are highly sensitive to the parameters of the computational models used to describe the data, in particular the specific shear viscosity of the QGP. This is an intrinsic quantity of QCD matter and of great interest to the field.

   In this project we are developing the methodology for a rigorous comparison of a state-of-the-art computer model called VISHNU, based on a hybrid 2+1 relativistic viscous fluid dynamical plus microscopic Boltzmann transport model, to event-by-event data on v_2, v_3, v_4 Fourier coefficients measured by the ATLAS experiment at the LHC. The goal of the analysis is the determination of high probability ranges for the VISHNU parameters, in particular the specific shear viscosity of the QGP. The analysis will most likely require Gaussian Process interpolation due to the high computational cost associated with the model. Furthermore a proper understanding of the v_n data (provided by the experiment in the form of various histograms) is vital for making legitimate, relevant comparisons.

3. *Fitting pdfs using binned vs continuous data:* The v_n analysis necessitates analyzing different data types. The experimental data comes in the form of histograms, while the simulation runs of the computer model are estimated distributions comprised of many data points. We found a common parametric density function with which to compare the two data sets. In the experimental histogram (binned data) case, the density must be fit by a minimum distance method. In the other (continuous data) case, we can directly use maximum likelihood methods.

4. *Emulators using derivative data:*

   The Gaussian process (GP) framework on which computer model emulators rely can incorporate the use of derivative information. While currently such derivative data are not easily obtained, this aspect of the framework may be needed in future work.

   In computer model calibration it is known that conditioning on prior partial derivatives can help to mitigate the effects of the discrepancy function being non-identifiable. That is, the discrepancy function describing the difference between experimental and simulated data can become more well-defined by conditioning on certain partial derivatives being zero in certain parts of the space.

5. *Understanding the form of the "data"*

   As mentioned before, we sought an appropriate parametric family of pdfs for fitting both experimental and simulated data; we found a suitably flexible class which we call the Generalized Reverse Weibull (GRW). We will continue to compare the difference in parameter estimates for the two cases as more data become available.

   One future goal is to understand further the "unfolding" process which has been applied to the experimental data by the physicists before it came to us. The working group is investigating the incorporation of this particular unfolding mechanism used by the physicists, which is either "bin-by-bin correction" or "Bayesian unfolding".

Work in progress: The numbering refers to the above topics being investigated by the WG

1. Computer experiments for assessing uncertainties due to systematic errors (Luc Demortier, Karen Kafadar, Peter Marcy)

2. and 5. v_n analysis (Steffen A. Bass, Chris Coleman-Smith, Snehalata Huzurbazar, Karen Kafadar, Peter Marcy)

We anticipate that the results of the analysis will be published either in *Physical Review C* or *Journal of Physics G.*

3. Comparing statistical estimates from both binned and continuous data (Snehalata Huzurbazar, Karen Kafadar, Peter Marcy)

4. Gaussian process emulators (Chris Coleman-Smith, Peter Marcy)

Review paper: Statistical issues in high-energy physics (All)

Anticipated future publications and software directly resulting from WG research
The numbering refers to the above topics being investigated by the WG

1. and 4. Computer experiments for simulations in high-energy physics

2. Fitting distributions of v_n Fourier coefficients

3. Rigorous model-data comparisons in high-energy nuclear physics for the extraction of quark-gluon-plasma properties

Statistical issues arising in problems of high-energy physics

R code for *Fit parameters of GRW by optimizing (i.e., minimizing Hellinger dist) between fitted pdf and smoothed histogram (PSO: particle swarm optimization)* `http://sites.google.com/site/peterwmarcy/research`

4.5 Online Streaming and Sketching

The mission of this WG was the development and analysis of fast randomized algorithms for computing leverage scores, and their application. Specific topics include

- Approximating leverage scores for L2 and other regression problems: Online, streaming, incremental streaming algorithms
- Numerical analysis: Sensitivity of leverage scores, numerical stability of algorithms
- Applications in astronomy: Characterization of streaming and time dependent aspects of low rank approximations, incremental computation of leverage scores
- Applications in feature selection: How to distinguish among almost identical columns with high leverage scores (RRQR factorization, clustering), derivation of formal bounds

WG members: Naomi Altman* (Stat, Penn State), Tamás Budavári (Physics & Astro, Johns Hopkins), Petros Drineas (CS, RPI), Ilse Ipsen* (Math, NCSU), Michael Mahoney (Math, Stanford) Xingye Qiao (Stat, SUNY Binghampton)

Postdocs: David Lawlor (SAMSI & Duke, Math), Garvesh Raskutti (SAMSI & STOR, UNC), Ching-Wa Yip* (Physics & Astro, Johns Hopkins)

Graduate students: John Holodnak (Math, NCSU), Kevin Penner (Math, NCSU), Thomas Wentworth (Math, NCSU)

Occasional members: Andreas Artemiou (Stat, Michigan TU), Alan Lenarcic (Genetics, UNC), Mauro Maggioni (Math, Duke), Ali Shojaie (Biostat, U. Washington), Yi Grace Wang* (SAMSI & Math, Duke), Lingsong Zhang (Stat, Purdue)

Follow-up activities: The WG will continue its operation, under the same name, throughout the 2013/14 SAMSI Program on Low-Dimensional Structure in High-Dimensional Systems.

Conference presentations

1. Ilse Ipsen, *Accuracy and stability issues for randomized algorithms*, SIAM Conference on Computational Science and Engineering, Boston, 26 February 2013
2. John Holodnak, *Accuracy of a Randomized Algorithm for Computing Leverage Scores*, NCSU Graduate Student Research Symposium, Raleigh, NC, 19 March 2013
3. Ilse Ipsen, *Sensitivity of leverage scores and coherence for randomized matrix algorithms*, Workshop on Advances in Matrix Functions and Matrix Equations, University of Manchester, UK, 10-12 April 2013
4. Ilse Ipsen, *Sensitivity of leverage scores and coherence for randomized matrix algorithms*, Meeting of the International Linear Algebra Society (ILAS), Providence, RI, 4 June 2013
5. John Holodnak, *Accuracy of a randomized algorithm for computing leverage scores*, SIAM Annual Meeting, San Diego, CA, 9 July 2013
6. David Lawlor, *Robust PCA for Massive Data*, SIAM Annual Meeting, San Diego, CA, 9 July 2013

7. Garvesh Raskutti, *Subsampling, regularization and leverage scores*, SIAM Annual Meeting, San Diego, CA, 9 July 2013

8. Thomas Wentworth, *Sensitivity of leverage score estimation*, SIAM Annual Meeting, San Diego, CA, 9 July 2013

Topics investigated by WG

1. *Robust PCA for massive data* (David Lawlor, Tamás Budavári, Ching-Wa Yip, Mauro Maggioni)

   We are investigating a method for robust PCA based on randomized outlier removal. In this framework the probability of a data point's exclusion is proportional to its leverage score, a concept which has recently led to fast randomized algorithms for matrix computations. We have implemented a preliminary version of the algorithm and am testing its performance on different classes of synthetic data. Future research plans include evaluating the performance of the method on astronomical spectra provided by Tamás Budavári; replacing the exact computation of leverage scores by a fast approximation; and obtaining rigorous proofs of the method's accuracy and speed for certain classes of statistical models.

2. *Regression in high dimensions via geometric multiresolution analysis* (David Lawlor, Mauro Maggioni, Tamás Budavári, Ching-Wa Yip)

   We are investigating a multiscale geometric framework for regression on massive and high-dimensional data sets. We have implemented a preliminary version of this algorithm and tested it extensively on astronomical spectra, in which the goal is to predict the redshift of a galaxy. We have also tested the algorithm on synthetic data and compared its performance against other prediction methods, in which our approach appears quite favorable. Future work includes fine-tuning the local regression estimates, perhaps using partial or total least squares; on-line implementation of the algorithms using cover trees; and obtaining rigorous proofs of the method's accuracy and speed for certain classes of functions.

3. *Leverage score approximation* (John Holodnak, Ilse Ipsen)

   We investigated a randomized algorithm by Drineas et al. for approximating the leverage scores of a matrix. The algorithm uses a random projection to reduce the dimension of the matrix. We expanded on the bounds in the original paper by identifying a deterministic condition that must be satisfied by the random projection in order for the algorithm to obtain relative accuracy. We also analyze the deterministic condition for the special case where the random projection is chosen to be the Subsampled Randomized Hadamard Transform. In addition, we computed the deterministic condition using specific data sets to show that the algorithm often performs much better than the worst case bounds indicate.

4. *Randomized subset selection* (John Holodnak, Ilse Ipsen, Kevin Penner)

   We analyzed a subset selection algorithm that chooses columns, in a deterministic fashion, corresponding to the highest leverage scores of a matrix. This algorithm is a variant of an algorithm by Boutsidis et al. The original algorithm samples a subset of columns from the matrix and then uses a deterministic subset selection algorithm by Gu and Eisenstat to return a specified number of columns. We present an example that shows that choosing the columns corresponding to the highest leverage scores rather than using random sampling can make worst case bounds arbitrarily bad. However, numerical experiments run on a range of test matrices indicate that this sampling strategy is often more effective in practice than sampling according to a probability distribution based on leverage scores.

5. *Leverage score sensitivity* (Ilse Ipsen, Thomas Wentworth)

   For real matrices with orthonormal columns, we derived perturbation bounds for the sensitivity of leverage scores and coherence in terms of the angles between the column spaces of the exact and perturbed matrices. The bounds imply that the perturbed leverage scores are close to the exact ones if the principal angles between the column spaces are small. For general matrices of full column rank, we derived perturbation bounds for leverage scores and coherence in terms of the norm of the perturbation and the condition number, with respect to inversion, of the exact matrix. The bounds imply that the leverage scores and coherence are insensitive if the exact matrix is well conditioned with respect to inversion.

6. *Algorithmic and statistical perspectives for leverage-score sampling for large-scale least-squares* (Garvesh Raskutti, Michael Mahoney)

   Leverage-score sampling has recently been proven to be a useful method for improving the speed of least-squares solvers. Past work has focussed on run-time and worst-case relative error guarantees from an algorithmic perspective. As yet, there has been little work on the performance of leverage-score sampling in the context of a statistical model. In our current work, we put leverage-score sampling into a statistical framework and assess the mean-squared error performance for different leverage-score based estimators.

7. *Random sampling for classification* (Xingye Qiao, Ali Shojaie, Lingsong Zhang)

   Random sampling techniques are applied for classifying massive and streaming data to achieve approximate results in a scalable and efficient manner

Work in progress

1. *Randomized subset selection* (John Holodnak, Ilse Ipsen)

   We have a draft manuscript based on current results. However, in order to deal with the "counter example" where the deterministic leverage score-based greedy algorithm can fail, we are also investigating a criterion based on determinant maximization. We are also investigating the relation between randomized column subset selection for a matrix A on the one hand, and randomized Monte Carlo matrix multiplication of AA^T on the other hand,. This can have important consequences for singular value bounds for sampled matrices, which are widely used in the analysis of randomized algorithms.

2. *Leverage score sensitivity* (Ilse Ipsen, Thomas Wentworth)

   We are preparing a manuscript based on the current results. Items still to be investigated include sensitivity of leverage scores associated with truncated SVDs (PCA), and extensive numerical experiments to illustrate the tightness of the bounds.

3. *Robust PCA for massive data* (David Lawlor, Tamás Budavári, Ching-Wa Yip, Mauro Maggioni)

   We are preparing a manuscript for submission to the astronomical community, in which we present an empirical evaluation of a number of algorithms for robust linear modeling on both simulated data and astronomical spectra. Preliminary results suggest that different methods perform better or worse depending on the structure of the inliers and outliers, and we expect to make this distinction clear in the near future.

4. *Regression in high dimensions via geometric multiresolution analysis* (David Lawlor, Mauro Maggioni, Tamás Budavári, Ching-Wa Yip)

   We are preparing a manuscript for submission to the astronomical community, in which we present an empirical evaluation of a number of algorithms for high-dimensional regression of galactic redshifts on astronomical spectra. Preliminary results suggest that our method, based on multiscale geometric ideas, performs better in terms of both error and runtime than global principal components regression, k-nearest neighbors regression, and diffusion maps regression.

4.6 Datamining and Clustering

The goal for this group was to explore various algorithms for data mining, ranking, and clustering with emphasis on linear algebraic techniques.

Methods discussed included: Data mining, OD and stochastic ranking & rating systems, recommender systems, dimension reduction & principal component analysis in clustering, reverse Simon–Ando clustering, graph partitioning, K-means, Fiedler partitioning, nonnegative matrix factorizations

WG members: Naomi Altman* (Stat, Penn State), Rada Chirkova* (CS, NCSU), Anjela Govan* (St. Thomas More Academy) Xiangqian Hu (SAS), Amy Langville* (CS, College of Charleston), Xiangxiang Meng (SAS), Carl Meyer (Math, NCSU), Chuck Paulson (Puffinware), Xingye Qiao (Stat, SUNY Binghamton),

Graduate students: Ralph Abbey (Math, NCSU), Hansi Jiang (Math, NCSU), Kevin Penner (Math, NCSU), Shaina Race* (OR, NCSU)

Publications directly resulting from WG research

1. Carl Meyer, Shaina Race, and Kevin Valakuzhy, *Determining the Number of Clusters via Iterative Consensus Clustering* (refereed paper), Proceedings of the 2013 SIAM International Conference on Data Mining (SDM13)

Conference presentations

1. Shaina Race, *Determining the Number of Clusters via Iterative Consensus Clustering*, 2013 SIAM International Conference on Data Mining (SDM13), Austin, Tx, May 2–4, 2013.

Topics investigated: A Scalable Streaming Algorithm for Latent Dirichlet Allocation (LDA), A Generalized Hebbian Algorithm for Incremental Latent Semantic Analysis, Incremental Singular Value Decomposition in Natural Language processing, Principal Points, PCA, SVD and MLE (the connections between PCA and maximum likelihood estimation for the purpose of data analysis), K-means Clustering via Principal Component Analysis, Graph Databases, Recommendation and the Use of Stochastic Clustering, Graph Visualization Algorithms, Use of eigenvalue gaps for clustering, Collaborative filtering, Clustering streaming data, All flavors of Laplacians for graph partitioning, Community Structure in Multislice Networks, Reverse Simon-Ando clustering, Psychometrics and Educational Data Mining

4.7 Multi-Scale Modeling

The goal of this WG is the development of multiscale modeling and methodology for massive and highly structured datasets.

WG members: David Dunson (Stat, Duke), Marco Ferreira* (Stat, U. Missouri), Mauro Maggioni (Math, Duke), Ji Meng Loh (Math, New Jersey IT), Joshua Vogelstein (Stat, Duke)
Postdoc: David Lawlor (SAMSI & Math, Duke)
Graduate students: Francesca Petralia* (Stat, Duke), Yuan Cheng* (Math & Stat, U. Missouri)

Papers published

1. Dmitriy Karpman, Marco A. R. Ferreira, Christopher K. Wikle, *A Point Process Model for Tornado Report Climatology*, Stat, vol. 2, 1-8, 2013

   We propose a point process model with multiplicative risk for the study of tornado reports in the United States. In particular, we implement a rigorous statistical procedure to evaluate whether tornado report counts are significantly related to topographic variability. The model we propose also includes flexible nonparametric components for spatial and seasonality effects. We apply the proposed model and methodology to the analysis of tornado report data from 1953 to 2010 in the United States. Our analysis shows that in addition to the spatial and seasonal effects, the topographic variability is an important component of tornado risk.

2. G. Chen, M. Iwen, S. Chin, and M. Maggioni,

   A fast multiscale framework for data in high dimensions: Measure estimation, anomaly detection, and compressive measurements, In: Visual Communications and Image Processing (VCIP), 2012 IEEE, pages 1-6, 2012.

   We study the geometry and distribution of high-dimensional data sets, and the relationships between the two. Here we model data as independent samples $X_n = \{x_1, ..., x_n\}$ from a probability measure μ in R^D. We will assume that μ may be well-approximated by a measure which has support on a set of dimension $d \times D$; this assumption is justified by many observations, empirical and, in some cases, theoretical. This setting has been considered in much existing work on dimension reduction, where the ambient space is compressed to lower dimension, under a constraint of small distortion of the distances between data points, or in manifold learning, where one seeks parametrizations of the data with a small number of parameters (ideally $O(d)$). Another approach is that of working directly in the high-dimensional space, while using appropriate constructions to exploit the low intrinsic dimension

3. Mark A. Iwen and Mauro Maggioni, *Approximation of points on low-dimensional manifolds via random linear projections*, Inference & Information, 2013. ArXive Preprint, arXiv:1204.3337v1.

   This paper considers the approximate reconstruction of points $x \in R^D$ which are close to a given compact d-dimensional submanifold, M, of R^D using a small number of linear measurements of x. In particular, it is shown that a number of measurements of x which is independent of the extrinsic dimension D suffices for highly accurate reconstruction of a given x with high probability. Furthermore, it is also proven that all vectors, x, which are sufficiently close to M can be reconstructed with uniform approximation guarantees when the number of linear measurements of x depends logarithmically on D. Finally, the proofs of these facts are constructive: A practical algorithm for manifold-based signal recovery is presented in the process of proving the two main results mentioned above.

Papers submitted

1. Fonseca and Ferreira, *Dynamic Multiscale Spatiotemporal Models for Poisson Data*, Journal of the American Statistical Association

   We propose a new class of dynamic multiscale models for Poisson spatiotemporal processes. Specifically, we use a multiscale spatial Poisson factorization to decompose the Poisson process at each time point into spatiotemporal multiscale coefficients. We then connect these spatiotemporal multiscale coefficients through time with a novel Dirichlet evolution. Further, we propose a simulation-based full Bayesian posterior analysis.

   In particular, we develop filtering equations for updating of information forward in time and smoothing equations for integration of information backward in time, and use these equations to develop a forward filter backward sampler for the spatiotemporal multiscale coefficients. Because the multiscale coefficients are conditionally independent a posteriori, our full Bayesian posterior analysis is scalable, computationally efficient, and highly parallelizable. Moreover, the Dirichlet evolution of each spatiotemporal multiscale coefficient is parametrized by a discount factor that encodes the relevance of the temporal evolution of the spatiotemporal multiscale coefficient. Therefore, the analysis of discount factors provides a powerful way to identify regions with distinctive spatiotemporal dynamics.

   Finally, we illustrate the usefulness of our multiscale spatiotemporal Poisson methodology with two applications. The first application examines mortality ratios in the state of Missouri, and the second application considers tornado reports in the American Midwest.

2. Vogelstein, Petralia, and Dunson, *Multiresolution Scalable Bayesian Conditional density estimation*, Neural Information Processing Systems

   Nonparametric estimation of the conditional distribution of a response given high-dimensional features is a challenging problem. In many settings it is important to allow not only the mean but also the variance and shape of the response density to change flexibly with features, which are massive-dimensional with a distribution concentrated near a lower-dimensional subspace or manifold. We propose a multiresolution model based on a novel stick-breaking prior placed on the dictionary weights. The algorithm scales efficiently to massive numbers of features, and can be implemented efficiently with slice sampling. State of the art predictive performance is demonstrated for toy examples and a real data application.

3. Ferreira and Sanyal, *Bayesian optimal sequential design for nonparametric regression via inhomogeneous evolutionary MCMC*, Statistical Methodology

   We develop a novel computational methodology for Bayesian optimal sequential design for nonparametric regression. This computational methodology, that we call inhomogeneous evolutionary Markov chain Monte Carlo, combines ideas of simulated annealing, genetic or evolutionary algorithms, and Markov chain Monte Carlo. Our framework allows optimality criteria with general utility functions and general classes of priors for the underlying regression function. We illustrate the usefulness of our novel methodology with applications to experimental design for nonparametric function estimation using Gaussian process priors and free-knot cubic splines priors.

Grants submitted

1. Marco Ferreira, *Bayesian Multiscale Spatiotemporal Modeling for Massive Datasets*, NSF-DMS, declined.

   Progressively larger spatiotemporal datasets require new and innovative methods for data analysis. In particular, massive datasets pose unique challenges because the statistical methodology has to be fast, scalable, and able to identify important low-dimensional patterns hidden under a deluge of data. To address this tremendous challenge, in this work the investigator develops new multiscale models and methods for the statistical analysis of massive spatiotemporal datasets.

2. Joshua Vogelstein and David Dunson, *An Integrated Bayesian Approach for Revealing Multiscale Biomarkers in Psychiatry*, NIH R01, pending.

   We propose to develop an integrated Bayesian approach for revealing multiscale biomarkers in psychiatry. According to the American Psychiatric Association's Consensus Report from July 2012, "there are currently no brain imaging biomarkers that are currently clinically useful for any diagnostic category in psychiatry". We conjecture that this lacking is at least partially due to the statistical challenges faced by large multimodal images. We therefore adopt a multiscale approach, developing biophysically inspired models for:

   (a) Multivariate temporal dynamics of the functional MRI data,
   (b) Brain connectivity of the diffusion MRI data,
   (c) Three-dimensional shape models for the structural MRI data,
   (d) An integrated model fusion of the various modalities.

3. Joshua Vogelstein, David Dunson, Michael Milham, and Randal Burns, *Learning Data-Driven Mental Health Diagnostic Taxonomy via Multimodal Fusion*, NIH R01, pending.

   According to the most recent World Health Organization's Global Burden of Disease reports, psychiatric conditions are responsible for almost 11% of human disease burden worldwide. Substantial technological advances across several treatment modalities (for example, pharmacological) have resulted in meaningful improvements in clinical outcomes for several psychiatric conditions. However, the full impact of these advances has not yet been fully realized, as evidenced by continuing high rates of treatment failures across the spectrum of psychiatric disorders.

   Treatment outcome is inextricably linked to diagnostic status. New treatments are assessed by identifying appropriate groups of patients, and comparing the relative efficacy of the treatment to placebo in this population. While human-guided groupings using patterns of symptoms has yielded tremendous progress, this approach is limited in two ways. First, it does utilize of modern computational capabilities for structure learning in large heterogeneous data. Second, it ignores the wealth of recently available neurobiological data.

   We propose to address this challenge by implementing two conceptual advances that we have applied successfully in other disciplines. First, we will develop and apply a statistical inference engine to learn structure from large heterogeneous, etiologically relevant, data. Second, we will feed these algorithms with three complementary and disparate modalities of each individual: (i) brain imaging, (ii) deep phenotyping, and (iii) expert diagnoses.

4. David Dunson and Barbara Engelhardt, *Bayesian epigenomics*, NIH R01, submitted.

   We propose to develop scalable Bayesian epigenomic methods to improve understanding of how epigenetic variability interacts with DNA sequence to modify expression and predict complex diseases. Our emphasis is on developing fundamentally new approaches, which do not rely naively on methods developed for gene expression data, but are tailored to adapt to the complexities of DNA methylation arrays. Improved understanding of how the DNA methylation profile changes with disease and environmental processes in relation to DNA sequence is critical in designing better preventive and treatment therapies to reduce risk and improve response to disease.

   The high-dimensionality and complex non-Gaussian methylation profile leads to unique data analytic challenges we will address with novel Bayesian nonparametric methods. In particular, we will focus on:

   (a) Learning the joint distribution of epigenetic modifications across the genome in relation to disease factors;
   (b) Screening for locations of epigenetic change between diseased and non-diseased individuals;
   (c) Identifying locations of epigenetic change as disease progresses within an individual;
   (d) Learning networks between epigenetic and genetic sequences predictive of disease.

   All of these directions will lead to innovative general purpose statistical and computational methods for big data, while being particularly targeted to challenges that arise in assessing epigenetic effects. We will assess the performance of our methods through a multi-tiered approach including

   (a) Theoretical assessments appropriate to the big data paradigm;
   (b) Simulation studies designed to mimic real multimodal epigenomic and genomic data as closely as possible;
   (c) n-fold cross validation applied to multiple data sets.

   Although we focus on Bayesian hierarchical modeling, we are strongly motivated to obtain approaches that have frequentist optimality properties in testing, estimating predictive distributions for disease phenotypes and in learning low-dimensional structure while appropriately penalizing for computational time. Hence, we avoid traditional approaches to Bayesian computation, and instead rely on fast approximations that can be routinely implemented. All of our code will be made publicly available to facilitate reproducibility, applications and extensions.

5. Mauro Maggioni, *Data Sketches for Dynamic Data Sets in High Dimensional Spaces*, NSF BIG DATA program.

   In the past decade the analysis of massive, high-dimensional data sets has become a necessity for a large number of researchers, confronted with massive amounts of data from genomics, finance, economics, imaging, spectroscopy, documents and web pages, computer vision, robotics, molecular biology, etc.

   The mathematical foundations of this type of analysis is still underdeveloped, in part because of the great heterogeneity of data sets and applications from which they emerge, that may lead one to suspect that finding common structures and models may be difficult if not impossible, against the spirit of mathematics that seeks general explanatory principles as much as possible. Mounting evidence and observations across several disciplines and dramatically different data sets in fact suggest the existence of complex low-dimensional geometric structures in the data, and that modeling and exploiting such structures for prediction, modeling and information extraction often leads to outperforming results.

The PI proposes work at the frontline of research at the intersection of low- and high-dimensional geometry, probability, harmonic analysis, estimation, machine learning and data structures. The first observation is that "big" in "big data" typically refers to some rather naive measure of size, such as the size a data matrix X which represents n data points in R^D. However, in general this has nothing to do with the inherent "complexity" of the data, or its "information content", or "compressibility". While the terms in quotes are ill-defined, and at any rate most often application- dependent, we will consider particular problems and applications where they may rigorously defined, in ways discussed below.

In any case, we posit that in many instances the "complexity" of the data - in fact, "complexity up to precision ρ" - can be much lower than the worst possible allowed by the ambient dimension D, and this phenomenon should be exploited as much as possible in order to reduce sampling requirements in estimation problems, running cost of algorithms, space dedicated to data storage.

6. Mauro Maggioni and I. Daubechies, *Structured Dictionary Models and Learning for High Resolution Images*, NSF DMS.

   We will construct new computationally efficient algorithms to analyze large collections of images (or, more generally, data sets), based on novel constructions of representations of the data, fast transforms mapping data to sparse representations, compressed and compressive probabilistic models, and new metrics for comparing data sets and portions thereof, taking into account known invariances of the data.

   Algorithms of this type are needed in a wide range of applications, in e.g. statistical signal processing, computer vision, and machine learning. We focus here on applications to data bases of images, and we illustrate our project on special collections of images (high resolution digitizations of art paintings) arising from our collaboration with art historians: collections of art paintings. As museums turn increasingly to digital (rather than photographical) means of documenting their collections, analyzing, organizing, searching, visualizing and extracting information from these large data sets may become possible at a scale previously undreamt of.

   To achieve this, we first have to meet serious challenges. Mathematically, these include the design of new fast transforms specialized for certain image classes (conceivably tailored to each collection), the definition of new metrics between images, and their use for the organization (to make them searchable) of large collections. Computationally, the challenges are no less daunting: high resolution scans and new cameras produce digital images each of which can have 108 ? 109 pixels, and the target collections can include as many as 104 such images.

7. Mauro Maggioni, *Information, Approximation, and Fast Algorithms for Data in High Dimensions*, AFOSR.

   In the past decade the analysis of massive, high-dimensional data sets has become a necessity for a large number of researchers. Data sets are often modeled by large matrices X which represent represent n data points in R^D. Mounting evidence and observations across several disciplines and dramatically different data sets in fact suggest the existence of complex low-dimensional geometric structures in the data, and that modeling and exploiting such structures for prediction, modeling and information extraction often leads to outperforming results. This key observation is at the foundation of many successful ideas, and related algorithms, in machine learning and statistics and signal processing. We are interesting in better understanding and exploiting the inherent ”complexity”, ”information content”, ”compressibility”, and ”low-dimensionality” of the data.

The PI proposes work at the frontline of research at the intersection of low- and high-dimensional geometry, probability, harmonic analysis, estimation, machine learning and data structures. At a high level, the objectives of the proposed work are:

(a) Novel multi scale geometric estimators for probability distributions in high-dimensions, that are supported near low-dimensional sets.

(b) Fast randomized algorithms producing the estimators above, that will be implemented and tested on large data sets.

(c) New metrics between probability distributions will be introduced, that may be thought of as multi scale versions of Wasserstein distances.

(d) Online algorithms for time-varying data sets that estimate the change, at multiple scales and ”locations” in the data sets, for time series of data sets.

(e) Integration of all of the above with a novel framework inspired by compressive sampling.

Work in progress

1. Cheng, Ferreira, and Vogelstein, *Multiscale spatiotemporal modeling for fMRI data*

   In this project we develop multiscale spatiotemporal modeling for resting state fMRI data.

2. David Lawlor, Mauro Maggioni, Tamas Budavari, and Ching-Wa Yip, *Regression in High Dimensions via Geometric Multiresolution Analysis*

   We are investigating a multiscale geometric framework for regression on massive and high dimensional data sets. We have implemented a preliminary version of this algorithm and tested it extensively on astronomical spectra, in which the goal is to predict the redshift of a galaxy. We have also tested the algorithm on synthetic data and compared its performance against other prediction methods, in which our approach appears quite favorable. Future work includes fine-tuning the local regression estimates, perhaps using partial or total least squares; on-line implementation of the algorithms using cover trees; and obtaining rigorous proofs of the method’s accuracy and speed for certain classes of functions. We are preparing a manuscript for submission to the astronomical community, in which we present an empirical evaluation of a number of algorithms for high dimensional regression of galactic redshifts on astronomical spectra. Preliminary results suggest that our method, based on multiscale geometric ideas, performs better in terms of both error and runtime than global principal components regression, k-nearest neighbors regression, and diffusion maps regression.

3. David Lawlor and Mauro Maggioni, *Fast Multiscale regression in high dimensions, with applications to astronomical data analysis*

   We exploit multi-resolution data structures in high dimensions to perform regression with fast greedy top-bottom algorithms and in adaptive fashion, with complexity dependent uniquely dependent on the intrinsic dimension of the data x and of the regularity of the regression function $\mathbf{E}[y|x]$. We aim at proving finite sample guarantees for our estimators, and to include work on vector- valued regression.

4. Mauro Maggioni, *Geometric Measure Estimation in high dimensions*

   We introduce a novel class of estimator of probability measures in high-dimensions with finite sample guarantees for how close the estimated measure is in Wasserstein distance to the true measure. Fast greedy algorithms are associated to these estimators. We present applications to dynamic data sets and anomaly detection.

5. Maggioni and Monson, *Multiscale Classification in high dimensions*

   We exploit multi-resolution data structures in high dimensions and extend them to perform classification tasks.

6. Gerber and Maggioni, *Fast multiscale computation of Wasserstein distances and optimal transport plans in high dimensions*

   We introduce novel multiscale algorithms for quickly computing the Wasserstein distances between samples of probability distributions in high-dimensions, and discuss applications to comparing high-dimensional data sets.

4.8 Stochastic Processes and Astrophysical Inference

The WG's mission is to pursue research on challenging informatics and statistics problems that arise in the analysis of time-domain astronomical data: high-energy transients, particularly gamma-ray bursts. Its goal is to bring together astronomers with novel and challenging problems in data analysis, some with massive data, and statisticians experienced at constructing and analyzing mathematical and statistical models, to work together in developing new insight into the scientific problems that underlie the data. This collaboration is intended to enable new science that would not be possible to pursue using conventional methods and algorithms. Particular topics include:

- New pulse shapes for Gamma Ray Burst (GRB) constituents
- Development of a new method (Parallel Thinning) for making inference with large datasets modeled with infinitely divisible distributions
- Uncovering the Intrinsic Shapes of GRB Pulses
- Time/Energy Domain GRB Models
- AGN Variability

WG members: Tamás Budavári (Physics & Astro, Johns Hopkins), Eric Feigelson (Astro, Penn State), Jon Hakkila (Physics & Astro, College of Charleston), Brandon Kelley (Physics, UC Santa Barbara), Tom Loredo (Astro, Cornell), Jeffrey Scargle (NASA Ames), Aneta Siemiginowska* (Harvard-Smithsonian Center for Astrophysics), Robert Wolpert (Stat, Duke)

Graduate student: Mary Beth Broadbent* (Stat, Duke)

Publications directly resulting from WG research

1. Jon Hakkila, Thomas J. Loredo, Robert L. Wolpert, Mary Broadbent, Robert D. Preece, *Uncovering the Intrinsic Shapes of GRB Pulses*, submitted

Grant proposals from WG research

1. R. L. Wolpert and T. J. Loredo, *Beyond source detection: Quantifying and propagating uncertainty throughout the photometric discovery chain*, NASA Astrophysical Data Analysis Program, under review
2. J. Hakkila, R. L. Wolpert, T.J. Loredo, *Using SWIFT to calibrate dynamic, spectrum-based GRB redshift estimators*, NASA/Goddard Astrophysics Science Archive Research Center, rejected

Conference presentations

1. Tamás Budavári, *Statistical Methods in Astronomy*, Joint Statistical Meetings, 8 August 2013

Work in progress

1. *New Pulse Shapes* (Broadbent, Wolpert, Loredo, Hakkila)

   Since the publication in APJ (2005) of *Long-Lag, Wide-Pulse Gamma-Ray Bursts* by Norris, Bonnell, Kazanas, Scargle, Hakkila and Giblin, the most common way of modeling GRB light curves has been a linear combination of one or more "*Norris kernel*" curves. This simple four-parameter family of functions exemplifies the "fast rise, exponential decay" or "FRED" shape astronomers have recognized in GRBs since their discovery. WG members are exploring the use of wider classes of curves in the hope of better representing complex GRB shapes with fewer constituent pulses.

2. *Parallel Thinning* (Broadbent, Wolpert)

   Many commonly-used statistical methods break down when applied to problems with copious data. Likelihood functions become so peaked (perhaps at multiple maxima) that simulation-based (*e.g.*, MCMC) methods get "stuck" in parameter space, mixing so slowly that inference becomes impractical. We are exploring a variation on Geyer's "parallel tempering" approach, applicable only for data following infinitely-divisible (ID) distributions. We construct Markov chains whose stationary distributions are proportional to the likelihood functions for "thinned" data-sets constructed by drawing periodic sub-samples from the conditional distribution of the rate-p summand.

3. *Intrinsic Pulse Shapes* (Hakkila, Loredo, Wolpert, Broadbent)

   Gamma-ray burst pulse light curves are characterized by five distinct segments: Precursor Shelf, Rapid Rise, Peak Plateau, Rapid Decay, and Extended Tail. Despite these distinct temporal segments, pulses are almost universally characterized by hard to soft evolution. These segments can give a single pulse the appearance of having three distinct localized peaks, which can lead to ambiguities in pulse-fitting if an incorrect single-pulse model is used. Gamma-ray burst pulses are more similar than they are different, indicating that a single process with few free parameters is responsible for producing pulses spanning a tremendous range of durations, luminosities, and spectral hardnesses. However, we show that a pulse's asymmetry is at least partially the result of instrumental sampling biases. Finally, we discuss how multi-episodic pulse emission places additional observational constraints on theoretical models.

4. *Time/Energy Domain GRB Models* (Broadbent, Wolpert, Loredo, Hakkila)

   The BATSE (Burst and Transient Source Experiment) instrument on NASA's Compton Gamma-Ray Observatory detected and recorded 2704 observations of GRBs, pulsars, and other transient phenomena during its operation from April 1991 to June 2000. BATSE's detectors cannot detect directly the energy level of an incident gamma-ray photon; instead, they record the amount of energy transferred from that gamma ray to a thallium-doped Sodium Iodide crystal scintillator, and re-emitted as photons captured by a photomultiplier tube. Earlier modeling has concentrated on the output from one of the detector's four energy channels; we are now exploring modeling a *incident* photon's energy and arrival time, using the detector's known response functions to support inference about gamma rays at a more basic level than in previous work.

5. *AGN Variability* (Kelley)

Active Galactic Nuclei (AGNs) or Quasars, the supermassive black holes at the center of massive galaxies, are the most intrinsically luminous non-transient objects in the sky, with luminosities that appear variable with time scales on the order of hours to months. Kelley is exploring the X-ray flux of AGNs and also of X-ray binary stars ("microquasars"), a proxy with similar behavior but more complete data. Flux distributions are well-fit by log-normal distributions, with power spectra (Fourier transform of autocorrelation) showing a "bump" due to the quasi-periodic oscillation (QPO). Traditional statistical tools are inadequate for quantifying the aperiodic variability; Kelley proposes a new approach modeling the departures of observed light-curves from their mean with Ornstein-Uhlenbeck processes.

4.9 Discovery and Classification in Synoptic Surveys

The aim of this WG is to pursue research on challenging informatics and statistics problems that arise in the analysis of astronomical data. The focus of the working group is on classes of problems that are cross-cutting, arising in diverse astrophysical applications. It brings broad expertise to the collaboration from other fields that are addressing discovery and analysis challenges with large data sets. The collaboration enables new science that would not be possible to pursue using conventional methods and algorithms.

WG members: Andreas Artemiou (Stat, Michigan TU), Jogesh Babu (Stat, Penn State), Eric Feigelson (Astro, Penn State), Fabrizia Guglielmetti* (MPI for Extraterrestrial Physics, Germany), Woncheol Jong (Stat, Seoul National U., Korea), Brandon Kelly (Physics, UC Santa Barbara), Jeff Kern (National Radio Astronomy Observatory), Alex Kim (Lawrence Berkeley National Lab), Ji MengLoh (Math, New Jersey IT), Tom Loredo (Astro, Cornell), Ashish Mahabal (Astro, Caltech), David Ruppert (OR & IE, Cornell), Prajval Shastri* (Astrophysics, IIA, Bangalore, India), Jiayang Sun* (Stat, Case Western), Robert Wolpert (Stat, Duke)

Graduate Student: SaeNa Park* (Penn State University)

Topics investigated by WG

1. *Characterizing light curves for determining type and distinguishing features* (Jogesh Babu, David Ruppert, Ashish Mahabal, SaeNa Park, Tom Loredo)

   Jogesh Babu, SaeNa Park and Ashish Mahabal focused on the Catalina Realtime Transient Survey (CRTS) light curves data. The aim is to explore the data and develop statistical methods and algorithms to classify and discover features of light curves. David Ruppert focused on Functional data analysis to the light curves. The lessons learned and the methods developed will aid in the future large surveys like LSST, and SKA.

2. *Image analysis for faint source detection (especially transients) or noise characterization, including multiple frequencies/bands, epochs, and subsets of baselines* (Fabrizia Guglielmetti, Ashish Mahabal and Jogesh Babu)

   The focus is on propagation of uncertainties in the measurement process, and faint source detection in multi-wavelength and multi-epoch data.

Work in progress

1. *Characterizing light curves for determining type and distinguishing features*

   Jogesh Babu, Ashish Mahabal and SaeNa Park explored the light curves from CRTS data. They studied the 6 and 8 dimensional clustering plots, identified potential outliers, and made movies of 6 dimensional and 8 dimensional clustering plots. These will aid in developing statistical procedures for automatic classification and identification of the sources based on their light curves. The graphs were made by SaeNa Park from the CRTS data under the direction of Jogesh Babu and Ashish Mahabal.

   Jogesh Babu gave a presentation on Clustering and classification. SaeNa Park worked on light curves data and made several graphs under the direction of Jogesh Babu and Ashish Mahabal. Ashish Mahabal gave 3 presentations on the nature of CRTS data, how transients are found, the variety and rarity of transients, and how current rates compare with future surveys which forms an important basis as to why we need more statistical and computational methods. In addition, Ashish Mahabal provided data on areas around several random pointings in the sky for clustering purposes. The objects are mostly non-variables with a few variable sources. Light curves as well as sets of statistically derived parameters were provided. Images for these areas are also to be provided for applications like co-addition and looking for fainter sources.

The potential number of light curves from CRTS is 500 million which can all be used once the various techniques we have discussed are developed. David Ruppert presented material on functional data analysis including testing hypotheses about the mean, feature extraction, classification, clustering, and principal components analysis on February 26, 2013. He also gave a presentation on nonparametric estimation covering local fitting, regression and penalized splines, confidence bands, density estimation, and SiZer on March 19, 2013. Future work will explore the application of statistical methods for sparsely observed functional data, especially techniques for clustering, classification, and testing, to light curves. The group may also apply SiZer to light curve analysis to detect increasing or decreasing trends in luminosity.

2. *Image analysis for faint source detection (especially transients) or noise characterization, including multiple frequencies/bands, epochs, and subsets of baselines*

   For faint source detection in astronomical images, the following challenges have been discussed:

   - Proper propagation of uncertainties in the measurement process;
   - Transients detection with multi-epoch and multi-wavelength data, leading also to extreme cases as the detection of variable objects in crowded regions;
   - Robust noise estimate, as represented by radio interferometric data.

   All the listed challenges are described by specific multi-frequency data sets. These data sets are often corrupted by the calibration process and are analyzed by a weak source detection procedure. Fabrizia Guglielmetti is planning to use the Background-Source separation algorithm (Guglielmetti F., et al.,2009, MNRAS,396,165) and adapt it for the challenges addressed by the new data sets.

4.10 Graphical Models and Graphics Processors

The purpose of this WG is to write Cuda code for statistical computations in astronomy.

WG members: Tamás Budavári (Physics & Astro, Johns Hopkins), Brandon Kelley (Physics, UC Santa Barbara), David Lawlor (SAMSI & Math, Duke), Tom Loredo (Astro, Cornell)

Publicly available software

1. *Cuda routines for MCMC sampling of hierarchical models on GPUs*
   `https://github.com/bckelly80/CUDAHM`