Text as data

- Proliferation of political text
- New Opportunities
- New Challenges

Analysis of event data

- Who did what to whom (where/when)
- High volume w/complex linkages
- Spatio-temporal analysis/forecasting
Data Management in Geoscience Papers
---Xuan Yu Dept. Geological Sciences

- Sea-level Rise Rate
- Topography
- Geology
- Tides
- Soil
- Hydrogeologic field observation
The world is **dynamic in time and space**. So is data!

**Functional data**: data from a sample of **random functions** (i.e., stochastic processes). ∞-dimensional data.


**Functional data analysis (FDA)**: a branch of statistics that analyzes functional data. Representative topics: function estimation, regression, dimension reduction, classification, clustering, network, etc.
Shanshan Ding  
Assistant Professor of Statistics

- **High-dimensional statistical learning for big data** - Dimension reduction, sufficient dimension reduction (SDR), feature screening
  1. Dimension folding PCA and PFC (neuroimaging data)
  2. Tensor sufficient dimension reduction (neuroimaging data)
  3. SDR with simultaneous variable selection in ultra-high dimension (genomic studies)  
     (e.g. biomarker and disease related brain region identification)

- **Parsimonious and efficient statistical modeling and inference** - Envelope models and methods
  1. Envelope matrix-variate regressions (neuroimaging, temporal and spatial data)
  2. Envelope quantile regression (health and behavioral studies)

- **Statistical applications** - biosciences, health and behavioral sciences, environmental studies
  Neuroimaging (EEG, MRI, fMRI), genomic data (RNA-seq, microarray), temporal and spatial, longitudinal data, time series, etc.
Robust Regression Models in Unseen Domains by Wavelet Scale Projection
Research in the Laboratory for Chemometrics

wavelet transform
orthogonal projection
latent variable regression model on each scale

high stability and performance in unseen domains

Research funded by NSF grant 1506853
Examples of topological materials

(a) Quantum Hall effect
(b) Quantum Spin Hall effect
(c) Quantum Anomalous Hall effect

Nikolić group, PRB 95, 201402(R) (2017)
Mul$variate Cogni$ve Metric Trajectory Predic$on in Alzheimer’s Disease
Lev E. Givon

- Huge, growing AD dataset*: 1650 paEents/controls, > 100 biomedical data types, mulE-year longitudinal data.
- Goal: discover interrelaEonships between mulEple biomarker trajectories that shed light on progression of AD.

- CogniEon = learning, memory, language, praxis, orientaEon, ...
- ADNI-trained ConvNet architecture can predict 13 normalized cogniEve metrics up to 3 years into future from minimally preprocessed current sMRI.

Joint work with L.J. Mariano, A.R. Schneider, D. O’Dowd, J.M. Irvine, & Alzheimer’s Disease Neuroimaging IniEaEve*
Charles Boncelet, ECE Dept, boncelet@udel.edu, 831-8008

Conducting research on machine learning applied to

- Information Security, esp. steganography and steganalysis.
- Electric Grid, control and resiliency.
- Signal processing.
- Algorithms for machine learning, e.g., graph based learning and entropy based methods.

Research highlights:

- Many highly cited papers in information security.
- Expert in data compression.
Analyzing Biological Networks Via Machine Learning

Li Liao
Computer & Information Sciences

Biological networks, including Protein-protein interactions (PPI) networks, play critical roles in many biological processes in the cell. Reconstructing and analyzing these networks from the huge amount of data generated from high throughput technologies present tremendous challenges as well as opportunities in both our efforts toward understanding the basic biology and translational research that can impact on human health. Our current research is focused on developing computational methods based on machine learning that can integrate data of different types and overlay multiple layers of mapping onto incomplete network to gain insights and make useful inference and detection of network related properties.

Specifically, we showcase several such methods that address the following: a) assessment of network evolution; b) inference of de novo edges; and c) detection of disease related nodes.
Astronomy: New Era of Petascale Data Science

John Gizis

Co-Chair, LSST Stars, Milky Way & Local Volume Science Collaboration

• Major new NASA space and NSF ground surveys
• LSST: 10 year survey of the sky, 15 Tb/night, 37 billion stars and galaxies
• Open Data Policy puts premium on data science collaborations and computational resources.
Molecular simulation of membranes with the Anton2 special purpose machine

Our interests:
• Modeling membranes w/ SoA HPC resources
• Fast, scalable algorithms for hydrodynamics
• Petascale HPC for drug binding kinetics

Anton is 100x faster than commodity machines for “all-atom” classical MD

Cholesterol interacting with a Parkinson’s target

Lyman research group
Physics and Astro, Chem and Biochem
This work is directed at data driven instructional environment design, implementation, and formative assessment for effective student centered learning - evidencing the journey to course learning goals (critical thinking, problem solving, and collaboration skills).

Some results:

1. Successfully guiding instruction using item analysis of, in the synchronous case, student clicker responses, and in asynchronous case, quiz/exam responses. Measures for analysis include choice frequency (distractors), difficulty index, and discrimination index.

2. Creation of about 15 hands on guided inquiry activities for SCEN 101 Physical Science; and new course SCEN 115 Origami Science.


Data collected from multiple sources including (1) rubrics in course LMS, (2) homework site, (3) deliverables from individual and group student work, (3) student clickers, and (4) from semester long multi-part projects.

Design principles inspired by Goldberg et. al. (AJP 78, 1265 (2010)), and we focus on the necessity of tools and the need for others/peers for learning. Tools were created, and others (high and low tech.) adapted. Figure on the left shows group topology. There are eight PBL groups (max. 6 members). Each PBL group is split into two experiment based learning (EBL) groups. Inter- and intra-group conversations enrich the learning experience.

Work in progress: Analytics driven constructions for automated individual and group feedback and email prompts - based on a cost function and training set built from past metrics of student performance.
Jeffrey Buler, Ph.D.
Department of Entomology and Wildlife Ecology

- Mapping migratory bird distributions with NEXRAD
- Need to automate real-time processing

NEXRAD – Weather Surveillance Radar

- Macroecology
- Conservation Biology
- Sustainability
CEMA: Center for Environmental Monitoring and Analysis

- Develop, operate, and maintain real-time environmental monitoring resources for Delaware
- Create and maintain value-added environmental data applications for all sectors of the Delaware economy
- Provides environmental data expertise, particularly in weather and climate, for Delaware.

DEOS Network: 70 real-time platforms

Weather

Hydrology

Wave Buoy

Satellite

Data Stakeholders Applications

- Emergency Management
- Transportation
- Natural Resource Managers
- Agriculture
- Public Health
- Researchers
- Consultants

Real-time Snow Monitoring Network

Delaware Water Quality Data Portal

Delaware Irrigation Management System

Coastal Flood Monitoring System

Lima Bean Downy Mildew Risk Tool
Graph Blue Noise and Graph Signal Processing

Gonzalo R. Arce
Institute of Financial Services Analytics, University of Delaware

Motivation

• Network Topology Inference
• Sampling
• Compress signals in irregular domains

Sampling

\[
G \quad \text{Represent } \mathbf{x} \in \mathbb{R}^N \text{ with the samples } \\
\{x(v_1), x(v_2), x(v_3), x(v_4)\} \quad \text{From } \\
\{x(v_1), x(v_2), x(v_3), x(v_4)\} \quad \text{get } \\
\{x(v_i), \forall v_i \in V(G) \text{ (Recovery)}
\]

Network Topology Inference

• Spectral analysis
• Filtering
• Predict evolution of a network process

Graph Signal Processing

Interpolate a brain signal from local observations
Compress a signal in an irregular domain
Locate the source of a tumor

Smooth an observed network profile
Predict the evolution of a network process
Infer the topology where the signals reside

Collaborators: Alejandro Parada-Mayorga (UD), D. Lau (University of Kentucky), S. Segarra (MIT).
Geometric Networks and Graph Limits

Mahya Ghandehari, Department of Mathematical Sciences

In geometric networks nodes embed in $X$

$$u \sim v \text{ iff } d(\pi(u), \pi(v)) \leq 1.$$ 

Ex. Social, biological and neural networks, . . .

... Question: How to retrieve the geometric reality of a network?
  e.g. metric space, geometric placement?

... Method: Graph limit theory.

... Results: Good “measure” of geometricity.
  Computable, robust to noise, continuous.

... Significance: Identify nearly geometric networks and uniformity of their processes.
100 Earths project
Goal is to discover 100 Earthlike planets
Humans need to look at every spectrum
Solution: undergrad data analysis lab

Always remember the **First Rule of Data: Look at the data!**
Multiscale Complex Fluids Modeling and Simulations
Antony N. Beris, UD Chemical and Biomolecular Engineering

High performance computations and data analysis of Non-Newtonian flows
Direct Numerical Simulations of viscoelastic turbulence using Spectral methods
Karhunen-Loeve, Principal Component Analysis of the results
Analysis, modeling and simulation of viscoelastic porous media flows
Macroscopic modeling of the flows of multiphase systems, like emulsions and suspensions

Modeling and simulation of thixotropic flows
Evaluation of yield stress and time-dependent hysteresis in aggregating suspensions
Modeling of blood rheology and multiscale simulations of blood arterial flow
Multiscale transport modeling using mesoscopic methods
Lian-Ping Wang, UD Mechanical Engineering

Data-intensive scalable computational methods
Boltzmann-equation based mesoscopic methods (lattice-Boltzmann, gas kinetic schemes)
Computationally scalable and physically flexible

Research questions related to the water cycle and water quality
How does air turbulence affect the collision rates of cloud droplets (warm rain initiation)?
What is the fate of contaminants when released to the soil environment?
How to model transport and retention of contaminants?
KEY MESSAGE
Making sense of Big Data is even more important than the data itself
We convert Data to meaning via System-Level Models:
Connectivity maps, Math Models, Artificial Intelligence, other tools

Petrochemical Refinery Data
50,000 data points every millisecond

Human Gut Microbiome
100 trillion cells, 3 million genes

Prasad Dhurjati, Professor of Chemical & Biomolecular Engineering, UD
http://www.che.udel.edu/dhurjati
Research on Process Analytics since 1982, Health Analytics since 1995
**IT HPC — Community Clusters**

### 1st Community Cluster
- AMD processors
- 200 nodes, 5000+ cores
- 40 Gbps InfiniBand network
- 256 TB high-performance storage (Lustre)
- **Currently end-of-life**

### 2nd Community Cluster
- Intel Xeon E5 v2 processors
- 192 nodes, 3300+ cores
- 56 Gbps InfiniBand network
- 256 TB high-performance storage (Lustre)
- nVidia GPU, Intel Phi options
- In year 3 of 4 year lifespan

---

**Timeline:**
- **2012**
- **2013**
- **2014**
- **2015**
- **2016**
- **2017**
- **2018**
- **2019**

**Clusters:**
- **Mills**
- **Farber**
IT HPC — Community Clusters

3rd Community Cluster
- Higher-density construction
- Intel Xeon E5 v4 processors
- 100 Gbps OmniPath network
- nVidia GPU, Intel Phi options
- Maximize reusable infrastructure
  - Rolling upgradeable
- More flexible buy-in options

To be named

Gen 1  Gen 2  Gen 3  Gen 4  Gen 5

IT HPC — Community Clusters

➢ Make your voice heard! Let us know you need:

http://www.udel.edu/003818

http://www.udel.edu/research-computing/contact/

it-hpc-interest@udel.edu

➢ For information on Research Computing at the University:

http://www.udel.edu/research-computing/