Development of a Computational and Data-Enabled Science and Engineering Ph.D. Program

Paul T. Bauman, Varun Chandola, Abani Patra
Matthew Jones

University at Buffalo, State University of New York

EduHPC 2014 Workshop
November 16, 2014
What is CDSE?
Need for Data Science Training

- 140,000–190,000 additional trained personnel for “deep analytical talent positions”, and 1.5 million more “data-savvy managers” needed to take full advantage of big data in the United States.

- National shortage for such talent is at least 60% – McKinsey Global Institute

- “Universities can hardly turn out data scientists fast enough.” New York Times
Need for Data Science Training

- Many new discoveries in science, new constructs in engineering and efficiencies/opportunities in business are made possible by CDSE methodologies e.g.
  - LHC generates 2-6 GB data/second — mining of which led to the Higgs discovery!
  - LSST is hailed as a breakthrough in data enabled science.

- Demand – students in multiple departments (MAE, CBE, CSEE, CSE, Math, Physics, Chemistry, ...) are in ad hoc programs of study aligned with this theme with gaps, poor sequencing and preparation
A Little History!

- Analog with Computational Science program development
- Many scientists doing computing, but little training in numerical mathematics, algorithms, software engineering
- Major advancements came not just from hardware, but from better algorithms

Taken from David Keyes, 
“Computational and applied mathematics in scientific discovery”
Interdisciplinary Approach

- We expect similar trends in the emergence of “Big-Data”
- Currently, trends essentially follow Moore’s Law
- Opportunities for major advancements emerging from interdisciplinary efforts
- We want history to repeat itself here!
- Gains will come from scientists who understand the problems on both sides

Taken from Nambiar and Poess, “Transaction Performance vs. Moore’s Law: A Trend Analysis”, TPCTC 2010
UB CDSE Ph.D. Program

- Students will get heavy doses of graduate applied mathematics/statistics, computer science, and HPC
- Build on a strong foundation in their core science/engineering
- Will have “end-to-end” expertise in innovating, developing, and delivering computational and data science products and tools
Other Data Science Programs

- NYU: Many MS programs, Ph.D. CS specializations
- Columbia: MS Data Science
- Stanford: MS CME, Data Science track
UB CDSE Ph.D. Program

CDSE DISSERTATION

DOMAIN SCIENCE

Applied Mathematics and Numerical Methods
High Performance and Data Intensive Computing
Data Science

MS 30 Hours
42 hrs=(30 from areas above; minimum 9 in each)
FIRST SEMESTER
Select advisor and research committee - committee should have a maximum of 2 members from same dept; core members must be CDSE faculty

SECOND SEMESTER
Decide on a research topic and submit short research plan to committee

THIRD SEMESTER
Pass oral examination on background information - committee provides written list of topics

FIRST FIVE SEMESTERS
finish 30 credit hours of core courses with minimum GPA of 3.2 from 3 areas - Data Science, Applied Mathematics/Numerical Methods, High Performance Computing (minimum of 9 credit hours from each area); See attached list (Sec 3.b). Other courses may be substituted to satisfy this requirement with the consent of committee.

Exam may be retaken once within 12 months

FIFTH SEMESTER
Submit and present refined research dissertation

Revise based on committee feedback

Successfully defend dissertation according to graduate school requirements
Core Courses

Data Science

- CSE 574: Introduction to Machine Learning
- EE 634: Principles of Information Theory and Coding
- IE 512: Decision Analysis
- MAE 674: Optimal Estimation Methods
- STA 502: Statistical Inference
- STA 503: Regression Analysis
- STA 521: Introduction to Theoretical Statistics I
- STA 536: Statistical Design and Analysis of Experiments
- STA 567: Bayesian Statistics

Applied and Numerical Mathematics

- IE 572: Linear Programming
- IE 575: Stochastic Methods
- MAE 702: Applied Functional Analysis
- MTH 539: Methods of Applied Mathematics I
- MTH 540: Methods of Applied Mathematics II
- MTH 543: Fundamentals of Applied Mathematics
- MTH 631: Analysis I
- MTH 649: Partial Differential Equations
### High Performance and Data Intensive Computing

- CSE 562: Database Systems
- CSE 587: Data Intensive Computing
- CSE 603: Parallel and Distributed Processing
- CSE 999: Large-Scale Distributed Data Systems
- IE 551: Simulation and Stochastic Models
- MAE 609: High Performance Computing I
- MAE 610: High Performance Computing II
- STA 545: Data Mining I
- STA 546: Data Mining II
Core Faculty

**Co-Directors**
Venu Govindaraju (CSE), Abani Patra (MAE)

**Lead Faculty**
Gino Biondini (Math), Vipin Chaudhary (CSE), Marianthi Markatou (Biostats), Murali Ramanathan (Pharm), Christian Tiu (Mgt Sci)

**New CDSE Faculty**
Paul Bauman (MAE), Varun Chandola (CSE), Michele Dupuis (CBE)
Affiliated Faculty

School of Engineering and Applied Sciences
A. Aref (CSEE), C.W. Chen (CSE), G. Dargush (MAE), S. Das (MAE), P. Desjardin (MAE), M. Demirbas (CSE), J. Errington (CBE), E. Furlani (CBE), T. Furlani (CCR), K. Hoffmann (Phys & BME), D. Kofke (CBE), T. Kosar (CSE), K. Lewis (CSE), D. Pados (EE), C. Qiao (CSE), D. Salac (MAE), G. Scutari (EE), ...

College of Arts and Sciences
L. Biang (GEO), M. Bursik (GLY), J.-H. Jung (Math), O. Khan (SAP), J. Ringland (Math), S. Metcalf (GEO), C. Renschler (GEO), B. Spencer (Math), G. Valentine (GLY), S. Rapoccio (Phy)

School of Management
S. Das Smith (MSS), R. Ramesh (MSS), H. R. Rao (MSS), A. Zhang (CSE)
Center for Computational Research

- Leading Academic Supercomputing Center
  - More than 14 years experience delivering HPC to campus
- 100 Tflops Total Aggregate Compute Capacity
  - 8000 cores; 600 TB of high performance storage
- Major Upgrades Planned for 2014
  - $1.2M ESD CFA Award
  - $100M Genome Initiative with NY Genome Center in NYC
- Utilization in 2013
  - 454 users (147 faculty), 2.1M jobs
- NYS HPC2 Consortium
  - RPI, UB, Stony Brook, Brookhaven, NYSERNet
  - Bringing HPC to NYS Industry and Academia
- NSF TAS Award
  - 5-year $7.5M XSEDE-based Award
  - XDMoD
http://cdse.buffalo.edu