“Big Data” at Georgia Tech
GTMI Advisory Board Meeting
April 16, 2014

Richard Fujimoto
Institute for Data and High Performance Computing
Georgia Institute of Technology
Big Data: Transforming Data into Value

Data Producers
- Social networks
- E-Commerce
- GIS
- Surveillance
- Healthcare
- Big science
- Machines/sensors
- Power grid
...

Data Systems
- Databases
- Cloud computing
- Middleware
- Security
- Operating systems
- Communications
- Hardware

Analysts (domain experts)
- Energy
- Health IT
- Internet
- Materials
- Manufacturing
- Security
- Sustainability
...

Data Transformers
- Data Analytics
- Modeling & Simulation
- Optimization
- High perf. computing

applications ← technologies

• Signal Processing
• Visualization
• Control
• Foundations
Sample Research & Education Programs

• FLAMEL: From Learning, Analytics, and Materials to Entrepreneurship and Leadership
• Data and Visual Analytics, High Performance Computing
• Systems Informatics and Control
• Education Programs
FLAMEL (From Learning, Analytics, and Materials to Entrepreneurship and Leadership) offers 25 two-year PhD student traineeships to develop advances in computing to accelerate the design of new materials; begins Fall 2014

GOAL: Accelerate the design of new materials using computation and advanced data analytics

PERSONNEL
Richard Fujimoto (CSE), Terry Blum (Business), Surya Kalidindi (MSE), Wendy Newstetter (BME), Hongyuan Zha (CSE)
~20 other faculty advisors from Math, Computing, MSE, ME, ChBE, Chemistry, Physics, ISyE

Experimental microstructure datasets obtained from modern materials characterization equipment.
Left: Beta-stabilized polycrystalline titanium being studied to develop new materials for lightweight high temperature structural applications (e.g., reduce fuel consumption and lower CO$_2$ emissions).
Right: 2 mm x 2 mm x 2 mm volume with a spatial resolution of 10 nm from a micro-porous layer in a fuel cell diffusion component.

FUNDING: NSF (IGERT Program)
FLAMEL Program Objectives

From Learning, Analytics, and Materials to Entrepreneurship and Leadership

• Students must develop foundational cross-disciplinary knowledge in materials informatics
• Students must develop the ability to formulate relevant problems in material design or manufacturing and work in interdisciplinary teams to develop innovative solutions
• Students must develop the ability to transform research innovations into commercial products and services in a competitive global marketplace
• The program must attract and prepare a diverse set of students to develop careers in materials informatics, and enable them to succeed in launching their careers

Nicholas Flamel (15th century)
FLAMEL Curriculum

• Four tracks (2 courses each)
  – Mathematics and Computation
  – Materials and Manufacturing
  – Entrepreneurship
  – Integration and Synthesis (2 new courses)

• Interdisciplinary Education
  – Problem-based learning
  – Interdisciplinary teams projects
  – GT entrepreneurship programs (e.g., Flashpoint, TI:GER)

• Recurring Themes
  – Communications
  – Cyberinfrastructure (MATIN)
Program at a Glance

Year 1

Curriculum (courses)
- Technical courses (Comp/Math/Mtrl/Mfg)
- Entrepreneurship
- Synth/Integration

Communication
- Workshop
- Comm. Course
- Chalk Talks/Semin.

Experiences
- Industry Internship
- Tech Innov (TI:GER/I-Core)
- Startup (Flashpoint)

Year 2
Engagement with Industry

• Advisory board
• Industry internships (summer)
• Lecture series
  – External visitors, including international
• Lunch series
  – Informal chalk-and-talk lunches
• Annual workshop
OBJECTIVE
Create the mathematical and computational science foundations required to represent and transform all types of digital data in ways to enable efficient and effective visualization and analytic reasoning.

DESCRIPTION
Modern data sets present many challenges in scale, heterogeneity, noisy/missing/time-varying data, etc. In its roles as the lead institution in the NSF/DHS FODAVA program, Georgia Tech is attacking fundamental issues in the visual analytics discipline.

Data and Visual Analytics involves the synergistic integration and iteration of analysis methods, computational algorithms, and visualization techniques to extract knowledge and understanding from large, complex interrelated data sets.
Challenges and Computational Solutions for High Dimensional Large-scale Visual Analytics

• Challenges
  – Data challenges
    (4V’s of Big Data: Volume, Velocity, Variety, Veracity)
      • Massive, High-dimensional, Nonlinear
      • Vast majority of data is unstructured
      • Time varying, dynamic
      • Heterogeneous format/sources/reliability
      • Noisy, errors and missing values are inevitable in real data set…
  – Visualization challenges
    • Screen Space and Visual Perception
    • Interaction: Speed necessary for real-time, interactive use

• Visual Analytic Solutions via Computational Advances
  – Dimension reduction methods that preserve key information
  – Informative representation (e.g. Clustering, Computational Zooming in and out, …)
  – Fast/scalable algorithms (e.g. Adaptive algorithms, on-line algorithms, …)
  – Information fusion
  – VA systems integrating foundational methods
FODAVA Research Test-bed for Visual Analytics of High Dimensional Data

Software: Available at http://fodava.gatech.edu/fodava-testbed-software
(and PyDAVA in progress)

- Library of key computational methods for visual analytics of high dimensional data
  - Foundational data analysis methods (dimension reduction, clustering, classification), their visual interactions, and visual comparisons through alignments
- Modular: A base for specialized VA systems (e.g. iVisClassifier, iVisClustering, VisIRR, UTOPIAN, …)
- Application domains: document analysis, bioinformatics, seismic data analysis, healthcare, communications, computer vision, …

- Language used:
  Computational library in Matlab
  GUI in JAVA
- System support:
  Windows 32/64 bit, Linux 32/64 bit
- PyDAVA: in Python, in progress in collaboration with GTRI
All involve analyzing massive streaming complex networks:

- **Health care** → disease spread, detection and prevention of epidemics/pandemics (e.g. SARS, Avian flu, H1N1 “swine” flu)
- **Massive social networks** → understanding communities, intentions, population dynamics, pandemic spread, transportation and evacuation
- **Intelligence** → business analytics, anomaly detection, security, knowledge discovery from massive data sets
- **Systems Biology** → understanding complex life systems, drug design, microbial research, unravel the mysteries of the HIV virus; understand life, disease,
- **Electric Power Grid** → communication, transportation, energy, water, food supply
- **Modeling and Simulation** → Perform full-scale economic-social-political simulations

Sample queries:
- **Allegiance switching**: identify entities that switch communities.
- **Community structure**: identify the genesis and dissipation of communities
- **Phase change**: identify significant change in the network structure

Ex: discovered minimal changes in O(billions)-size complex network that could hide or reveal top influencers in the community

**REQUIRES PREDICTING / INFLUENCE CHANGE IN REAL-TIME AT SCALE**
On-Line Analytics: Self-Adaptive Transportation Systems  
(Hunter, Alexopoulos, Fujimoto, Guensler)

Using Video Streams to Drive Distributed Simulation

Travel Time: On-Line Simulation vs. Field Experiments

Individual Vehicles Simulate Local Area of Interest

State prediction of each simulator must adapt to
- Live sensor data indicating current system state
- Predicted state provided by other simulators

Massive Data from Phases of Product Realizations
Nagi Gabraeel and JianJun Shi

Q1: How to model and represent those data for given objective?
Q2: How to develop unified approach for quality improvements?
Massive Data Generated at Enterprise Level

Q1: How to provide right data to the right place (people), with the right format, for the right objective?

Q2: How to make (close to real-time) adequate decisions with informed risks considering data uncertainties and actuation uncertainties?
Big Data Predictive Analytics

- Predictive Analytics for improving reliability and sustainability of manufacturing and service systems.
Big Data Education Programs

Building a workforce pipeline

Curriculum emphasizing foundational mathematics, computing knowledge and skills, applications

• Undergraduate minors for non-computing students
  – Scientific & Engineering Computing
  – Computational Data Analysis
• Undergraduate “thread” in CS program
  – Thread in modeling and simulation
• Proposed: Masters Program in Analytics (ISyE, CSE, College of Business)
• Interdisciplinary MS and PhD degree programs in Computational Science and Engineering
  – Jointly offered by three colleges: Computing, Sciences, and Engineering
  – MS offered through distance learning