Data Analytics and its Curricula

Microsoft eScience Workshop

October 9 2012 Chicago

Geoffrey Fox

gcf@indiana.edu

Informatics, Computing and Physics

Indiana University Bloomington

https://portal.futuregrid.org

Data Analytics

- Broad Range of Topics from Policy to new algorithms
- Enables X-Informatics where several X's defined **especially in Life Sciences**
  - Medical, Bio, Chem, Health, Pathology, Astro, Social, Business, Security, Intelligence Informatics defined (more or less)
  - Could invent Life Style (e.g. IT for Facebook), Radar …. Informatics
  - Physics Informatics ought to exist but doesn't
- Plenty of Jobs and broader range of possibilities than computational science but similar issues
  - What type of degree (Certificate, track, "real" degree)
  - What type of program (department, interdisciplinary group supporting education and research program)

Computational Science

- Interdisciplinary field between computer science and applications with primary focus on simulation areas
- Very successful as a research area
  - XSEDE and Exascale systems enable
- Several academic programs but these have been less successful as
  - No consensus as to curricula and jobs (don't appoint faculty in computational science; do appoint to DoE labs)
  - Field relatively small
- Started around 1990
- Note Computational Chemistry is typical part of Computational Science (and chemistry) whereas Cheminformatics is part of Informatics and data science
  - Here Computational Chemistry much larger than Cheminformatics but
  - Typically data side larger than simulations

General Remarks I

- **An immature (exciting) field:** No agreement as to what is data analytics and what tools/computers needed
  - Databases or NOSQL?
  - Shared repositories or bring computing to data
  - What is repository architecture?
- **Sources:** Data from observation or simulation
- **Different terms:** Data analysis, Datamining, Data analytics., machine learning, Information visualization, Data Science
- **Fields:** Computer Science, Informatics, Library and Information Science, Statistics, Application Fields including Business
- **Approaches:** Big data (cell phone interactions) v. Little data (Ethnography, surveys, interviews)
- **Includes:** Security, Provenance, Metadata, Data Management, Curation

General Remarks II

- **Tools:** Regression analysis; biostatistics; neural nets; bayesian nets; support vector machines; classification; clustering; dimension reduction; artificial intelligence; semantic web
- **Some driving forces:** Patient records growing fast (70PB pathology) and Abstract graphs from net leading to community detection
- Some data in **metric spaces**; others very high dimension or **none**
- **Large Hadron Collider** analysis mainly histogramming – all can be done with MapReduce (larger use than MPI)
- **Commercial:** Google, Bing largest data analytics in world
- **Time Series:** Earthquakes, Tweets, Stock Market (**Pattern Informatics**)
- **Image Processing** from climate simulations to NASA to DoD to Radiology (Radar and Pathology Informatics – same library)
- **Financial decision support**; marketing; fraud detection; automatic preference detection (map users to books, films)

School	Program	On-Campus	Online	Degrees
Undergraduate				
George Mason University	Computational and Data Sciences: the combination of applied math, real world CS skills, data acquisition and analysis, and scientific modeling	Yes	No	B.S.
Illinois Institute of Technology	CS Specialization in Data Science CIS specialization in Data Science			B.S.
Oxford University	Data and Systems Analysis	?	Yes	Adv. Diploma
Masters				
Bentley University	Marketing Analytics: knowledge and skills that marketing professionals need for a rapidly evolving, data-focused, global business environment.	Yes	?	M.S.
Carnegie Mellon	MISM Business Intelligence and Data Analytics: an elite set of graduates cross-trained in business process analysis and skilled in predictive modeling, GIS mapping, analytical reporting, segmentation analysis, and data visualization.	Yes		M.S. 9 courses
Carnegie Mellon	Very Large Information Systems: train technologists to (a) develop the layers of technology involved in the next generation of massive IS deployments (b) analyze the data these systems generate			
DePaul University	Predictive Analytics: analyze large datasets and develop modeling solutions for decision making, an understanding of the fundamental principles of marketing and CRM	Yes	?	MS.
Georgia Southern University	Comp Sci with concentration in Data and Know. Systems: covers speech and vision recognition systems, expert systems, data storage systems, and IR systems, such as online search engines	No	Yes	M.S. 30 cr

Illinois Institute of Technology	CS specialization in Data Analytics: intended for learning how to discover patterns in large amounts of data in information systems and how to use these to draw conclusions.	Yes	?	Masters 4 courses
Louisiana State University businessanalytics.lsu.edu/	Business Analytics: designed to meet the growing demand for professionals with skills in specialized methods of predictive analytics 36 cr	Yes	No	M.S. 36 cr
Michigan State University	Business Analytics: courses in business strategy, data mining, applied statistics, project management, marketing technologies, communications and ethics	Yes	No	M.S.
North Carolina State University: Institute for Advanced Analytics	Analytics: designed to equip individuals to derive insights from a vast quantity and variety of data	Yes	No	M.S.: 30 cr.
Northwestern University	Predictive Analytics: a comprehensive and applied curriculum exploring data science, IT and business of analytics	Yes	Yes	M.S.
New York University	Business Analytics: unlocks predictive potential of data analysis to improve financial performance, strategic management and operational efficiency	Yes	No	M.S. 1 yr
Stevens Institute of Technology	Business Intel. & Analytics: offers the most advanced curriculum available for leveraging quant methods and evidence-based decision making for optimal business performance	Yes	Yes	M.S.: 36 cr.
University of Cincinnati	Business Analytics: combines operations research and applied stats, using applied math and computer applications, in a business environment	Yes	No	M.S.
University of San Francisco	Analytics: provides students with skills necessary to develop techniques and processes for data-driven decision-making — the key to effective business strategies	Yes	No	M.S.

Certificate				
iSchool @ Syracuse	Data Science: for those with background or experience in science, stats, research, and/or IT interested in interdiscip work managing big data using IT tools	Yes	?	Grad Cert. 5 courses
Rice University	Big Data Summer Institute: organized to address a growing demand for skills that will help individuals and corporations make sense of huge data sets	Yes	No	Cert.
Stanford University	Data Mining and Applications: introduces important new ideas in data mining and machine learning, explains them in a statistical framework, and describes their applications to business, science, and technology	No	Yes	Grad Cert.
University of California San Diego	Data Mining: designed to provide individuals in business and scientific communities with the skills necessary to design, build, verify and test predictive data models	No	Yes	Grad Cert. 6 courses
University of Washington	Data Science: Develop the computer science, mathematics and analytical skills in the context of practical application needed to enter the field of data science	Yes	Yes	Cert.
Ph.D				
George Mason University	Computational Sci and Informatics: role of computation in sci, math, and engineering,	Yes	No	Ph.D.
IU SoIC	Informatics	Yes	No	Ph.D

Informatics at Indiana University

- School of Informatics and Computing
  - Computer Science
  - Informatics
  - Information and Library Science (new DILS was SLIS)
- Undergraduates: Informatics ~3x Computer Science
  - Mean UG Hiring Salaries
  - Informatics $54K; CS $56.25K
  - Masters hiring $70K
  - 125 different employers 2011-2012
- Graduates: CS ~2x Informatics
- DILS Graduate only, MLS main degree

Original Informatics Faculty at IU

- Security largely moving to Computer Science
- Bioinformatics moving to Computer Science
- Cheminformatics
- Health Informatics
- Music Informatics moving to Computer Science
- **Complex Networks and Systems** now largest
- **Human Computer Interaction Design** now largest
- Social Informatics

- Move partly as CS rated; Informatics not
- Illustrates difficulties with degrees/departments with new names

Informatics Job Titles

Account Service Provider
Analyst
Application Consultant
Application Developer
Assoc. IT Business analyst
Associate IT Developer
Associate Software Engineer
Automation Engineer
Business Analyst
Business Intelligence
Business Systems Analyst
Catapult Rotational Program
Computer Consultant
Computer Support Specialist
Consultant
Corporate Development Program Analyst
Data Analytics Consultant
Database and Systems Manager
Delivery Consultant
Designer
Director of Information Systems
Engineer
Information Management Leadership Program
Information Technology Security Consultant
IT Business Process Specialist
IT Early Development Program
Java Programmer
Junior Consultant
Junior Software Engineer
Lead Network Engineer
Logistics Management Specialist
Market Analyst

Informatics Job Titles

- Marketing Representative
- Mobile Developer
- Network Engineer
- Programmer
- Project Manager
- Quality Assurance Analyst
- Research Programmer
- Security and Privacy Consultant
- Social Media Mgr & Community Mgmt
- Software Analyst
- Software Consultant
- Software Developer
- Software Development Engineer
- Software Development Engineer in Test (SDET)
- Software Engineer
- Support Analyst
- Support Engineer
- System Administrator
- System integration Analyst
- Systems Architect
- Systems Engineer
- Systems/Data Analyst
- Tech Analyst
- Tech Consultant
- Tech Leadership Dev Program
- UI Designer
- User Interface Software Engineer
- UX Designer
- UX Researcher
- Velocity Software Engineer
- Velocity Systems Consultant
- Web Designer
- Web Developer

Undergraduate Cognates

- Biology
- Business
- Chemistry
- Cognitive Science
- Communication and Culture
- Computer Science
- Economics
- Fine Arts (2 options)
- Geography
- Human-Centered Computing
- Information Technology
- Journalism
- Linguistics
- Mathematics
- Medical Sciences
- Music
- Philosophy of Mind and Cognition
- Pre-health Professions
- Psychology
- Public and Environmental Affairs (5 options)
- Public Health
- Security
- Telecommunications (3 options)

Data Science at Indiana University

- Currently Masters in CS, Informatics, HCI, Bioinformatics, Security Informatics and will add Information and Library Science (ILS)
- Propose to add a Masters in Data Science (30 cr.) with courses covering CS, Informatics, ILS
  - Data Lifecycle (~ILS)
  - Data Analysis (~CS)
  - Data Management (~CS and ILS)
  - Applications (X Informatics) (~Informatics)
- Also minor/certificates
- Number of courses in each category being debated
  - Existing programs would like their courses required
  - i.e. political and technical issues in decisions

Massive Open Online Courses (MOOC)

- MOOC's are very "hot" these days with Udacity and Coursera as start-ups
- Over 100,000 participants but concept valid at smaller sizes
- Relevant to Data Science as this is a new field with few courses at most universities
- Technology to make MOOC's
  - Drupal mooc (unclear it's real)
  - Google Open Source Course Builder is lightweight LMS (learning management system) released September 12 rescuing us from Sakai
- At least one model is collection of short prerecorded segments (talking head over PowerPoint)

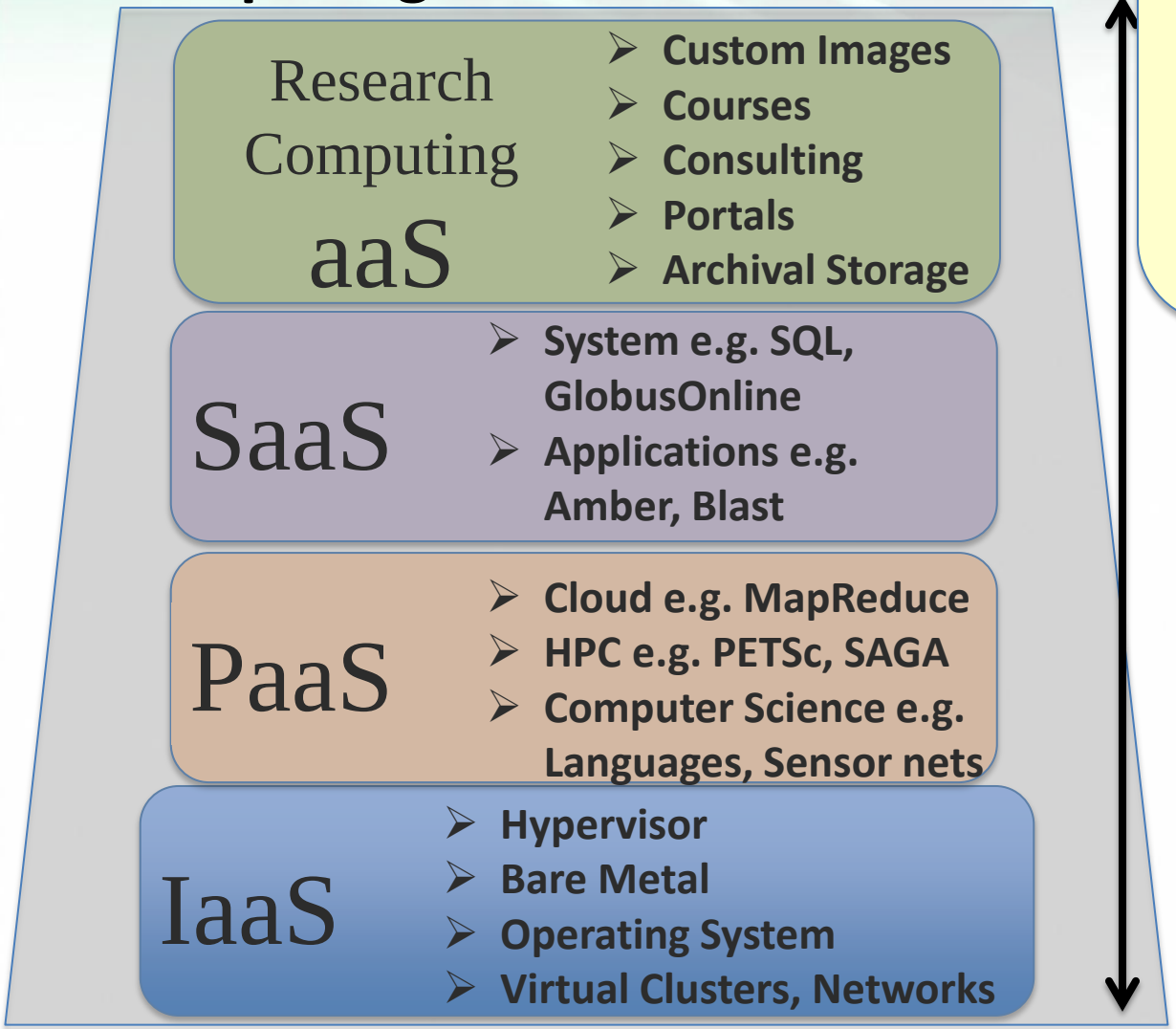
I400 X-Informatics (MOOC)

- General overview of “use of IT” (data analysis) in “all fields” starting with data deluge and pipeline
- Observation→Data→Information→Knowledge→Wisdom
- Go through many applications from life/medical science to “finding Higgs” and business informatics
- Describe cyberinfrastructure needed with visualization, security, provenance, portals, services and workflow
- Lab sessions built on virtualized infrastructure (appliances)
- Describe and illustrate key algorithms histograms, clustering, Support Vector Machines, Dimension Reduction, Hidden Markov Models and Image processing

Data Analytics Futures?

- **PETSc** and **ScaLAPACK** and similar libraries very important in supporting parallel simulations
- Need equivalent **Data Analytics libraries**
- Include **datamining** (Clustering, SVM, HMM, Bayesian Nets …), **image processing**, **information retrieval** including **hidden factor** analysis (LDA), **global inference**, **dimension reduction**
  - Many libraries/toolkits (R, Matlab) and web sites (BLAST) but typically not aimed at scalable high performance algorithms
- Should support **clouds and HPC; MPI** and **MapReduce**
  - Iterative MapReduce an interesting runtime; Hadoop has many limitations
- Need a **coordinated Academic Business Government Collaboration to build robust algorithms that scale well**
  - Crosses Science, Business Network Science, Social Science
- Propose to build community to define & implement **SPIDAL** or **Scalable Parallel Interoperable Data Analytics Library**

FutureGrid offers Computing Testbed as a Service



FutureGrid Uses Testbed-aaS Tools

- Provisioning
- Image Management
- IaaS Interoperability
- IaaS tools
- Expt management
- Dynamic Network
- Devops

FutureGrid Usages

- **Computer Science**
- **Applications** and understanding **Science Clouds**
- **Technology Evaluation** including XSEDE testing
- **Education and Training**