X-Informatics Course

Geoffrey Fox Instructor, I-400(Undergraduate) I-590(Graduate) Spring 2013

X-Informatics is the application of Information Technology to a broad range of applications denoted generically X. It generalizes "Bioinformatics, Cheminformatics, Health, and Social Informatics" fields in the School of Informatics and Computing. Other well defined X-Informatics fields include Medical, Pathology, Astronomy, Business, Financial, Security, Crisis and Intelligence Informatics. We also can define Life Style or Consumer (the IT underlying modern Web 2.0 and Ubiquitous computing) Radar, and Physics (including techniques used in analyzing data from Large Hadron Collider to discover Higgs Boson) Informatics. X-Informatics is characterized by "Big Data" [1] and the "Fourth Paradigm of Science" [2] where the recent huge increases in data in many areas have changed the approach to discovery (in Science) and decision making (term used in business and consumer applications. The study of X-Informatics is closely related to "Data Science" where it has been recently estimated [3] that by 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as need 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions.

We will define the processing model in X-Informatics with the step wise transformations Observation→Data→Information→Knowledge→Wisdom and build course around this concept with the data deluge driving the increasing use of this pipeline. The course will have a modular structure for each of the different application areas X. There will be a cross cutting discussion of the technology (cyberinfrastructure) needed to implement this pipeline. We will describe general principles including role of provenance (meta-data) to label process; Security; issues around open data and publication; ethics and need to form collaborative teams (i.e. social issues). Technically we will overview role of clouds, grids, parallel programming and workflow, statistics and visualization. Special tools that will be discussed include MapReduce, the statistics language R, histograms, clustering, Support Vector Machines, Dimension Reduction, Hidden Markov Models and Image processing, We distinguish data analysis (refining the expected with excellence) and data mining (looking for the unexpected).

The course could have a practical lab component which will be designed so it can be performed at different levels by students at different maturities e.g. that graduate students will be able to perform more challenging programming. The minimal programming will involve customizing services and forming workflows to produce visualizations. The laboratories will all be set up using virtual machine appliances i.e. pre-configured images that can either run locally or on a cloud environment like OpenStack/Eucalyptus on FutureGrid or Amazon/Azure.

1. Geoffrey Fox, Tony Hey, and Anne Trefethen, *Where does all the data come from?* , Chapter in *Data Intensive Science* Terence Critchlow and Kerstin Kleese Van Dam, Editors. 2011. http://grids.ucs.indiana.edu/ptliupages/publications/Where%20does%20all%20the%20data%20come%20from%20v7.pdf.
2. Jim Gray, Tony Hey, Stewart Tansley, and Kristin Tolle. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. 2010 [accessed 2010 October 21]; Available from: http://research.microsoft.com/en-us/collaboration/fourthparadigm/.
3. James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and A.H. Byers. *Big data: The next frontier for innovation, competition, and productivity*. 2011 [accessed 2012 August 23]; McKinsey Global Institute Available from: http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation.