Big Data Applications and Technologies (X-Informatics)

September 22 2013

Geoffrey Fox
gcf@indiana.edu

http://www.infomall.org/X-InformaticsSpring2013/index.html

Associate Dean for Research, School of Informatics and Computing
Indiana University Bloomington

2013

Big Data Ecosystem in One Sentence

Use Clouds running Data Analytics Collaboratively processing Big Data to solve problems in X-Informatics (or e-X)

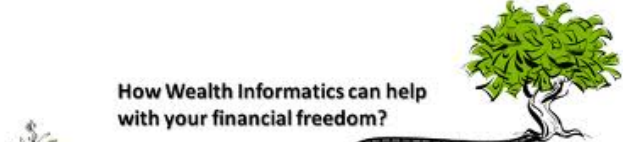
X = Astronomy, Biology, Biomedicine, Business, Chemistry, Climate, Crisis, Earth Science, Energy, Environment, Finance, Health, Intelligence, Lifestyle, Marketing, Medicine, Pathology, Policy, Radar, Security, Sensor, Social, Sustainability, Wealth and Wellness with more fields (physics) defined implicitly

Spans Industry and Science (research)

Education: Data Science see recent New York Times articles
http://datascience101.wordpress.com/2013/04/13/new-york-times-data-science-articles/

X-Informatics Class http://www.infomall.org/X-InformaticsSpring2013/
Big data MOOC http://x-informatics.appspot.com/preview

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.



Xinformatics

NSF 13-093

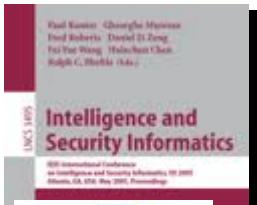
Dear Colleague Letter: OFR-NSF Partnership in Support of Research Collaborations in Finance Informatics

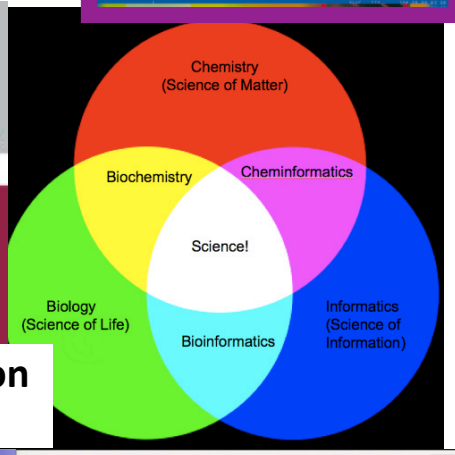
Date: May 14, 2013

Earth Science INFORMATICS

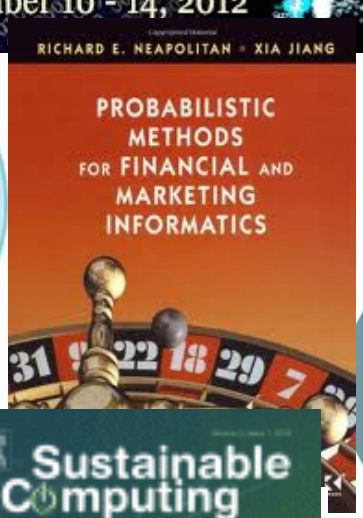
Biomedical Informatics
Computer Applications in Health Care and Biomedicine

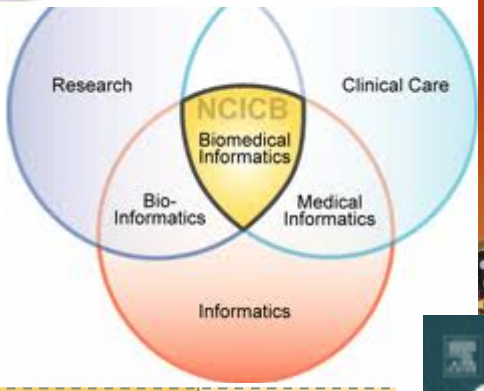
AstroInformatics2012
ber 10 - 14, 2012

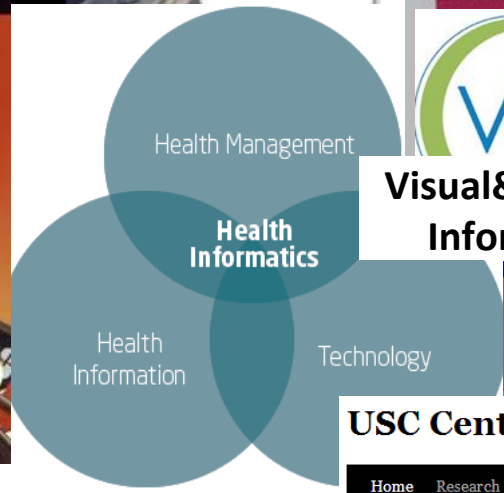




Climate Informatics
network





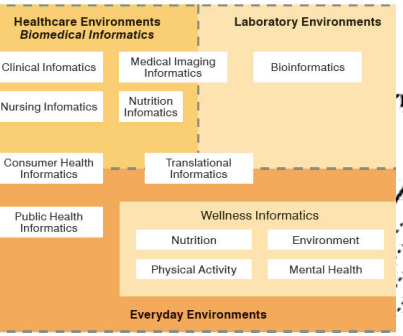


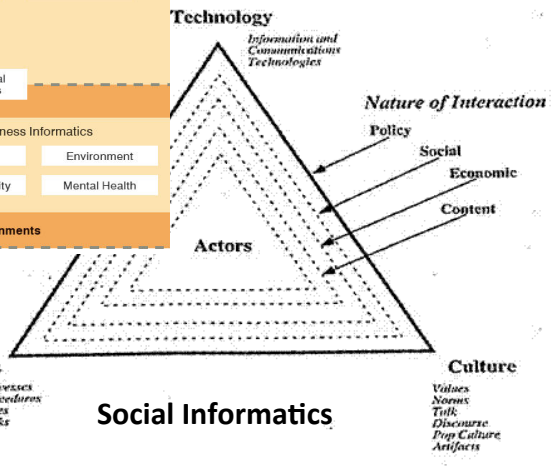
VDI

Visual&Decision Informatics

Opportunities and Challenges in Crisis Informatics

Sustainable Computing
Informatics & Systems



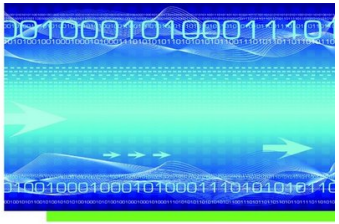


Social Informatics

USC Center For Energy Informatics

Home Research Publications Sma

GEO Informatics
Knowledge for Surveying, Mapping & GIS Professionals



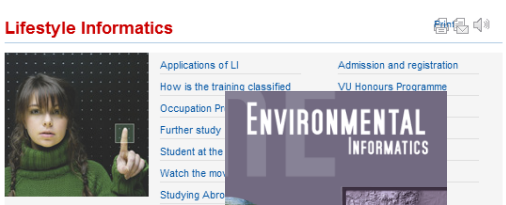
About the Center

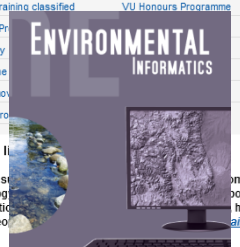
Welcome to the Center For Energy Informatics (CEI) at USC, an Organized Research Unit (ORU) housed in the Viterbi School of Engineering. Energy Informatics is the application of

Noelia Penelope Greer (Ed.)

Business Informatics

Information technology, Management,





The Course

Overview

The Approach

- The Big Data Applications and Technologies (X-Informatics) course is designed as an overview course for a Data Science Curriculum
- The courses covers a mix of **applications** (the X in X-Informatics) and **technologies** (aka data analytics) needed to support the field electronically i.e. to process the application data
- Built around rallying cry: ***Use Clouds running Data Analytics Collaboratively processing Big Data to solve problems in X-Informatics***
- The technologies give an overview of **parallel computing** and **clouds** and discuss some of the **data analytics**. The **X-Informatics** describes the application and its data.
- There are some discussion of **software** and that is arranged in two tracks – **Python** expecting to be run on your local machine (e.g. laptop) or **Java** which can be run on a cloud (FutureGrid (OpenStack), Amazon or Azure) or again on your laptop in less ambitious fashion.

The Course Structure

- The current course is set up as a MOOC (massive open online course) with over 25 units (which can expect to change in number and feature choice in units taken) where a unit is from 30-90 minutes of instruction.
- There is an additional motivation unit to explain why topic of course is important
- Each unit is broken up into lessons which are from 5-15 minutes in length
- The homework and mentoring (answering questions, grading) are separate from MOOC lessons
- Some units involve software as indicated later
- This software is offered in tracks – currently for Python and Java and for clients and Clouds
- The mechanics of software are described in SideMOOC's that tell you how to install and set up needed environments.
- Lists of useful resources (largely web sites) are given

The Topics of Big Data & X-Informatics

- Introduction (This Presentation)
- Motivation: Big Data and the Cloud; Centerpieces of the Future Economy
- 3 Units: Introduction: What is Big Data, Data Analytics and X-Informatics
- SideMOOC: Python for Big Data and X-Informatics: NumPy, SciPy, MatPlotlib
- 4 Units: X= LHC Analysis and Discovery of Higgs particle
  - Technology: Explore Events; histograms and models; basic statistics
- SideMOOC: Using Plotviz Software for Displaying Point Distributions in 3D
- 3 Units: X= e-Commerce and Lifestyle
- Technology: Recommender Systems - K-Nearest Neighbors
- Technology: Clustering and heuristic methods
- Parallel Computing Overview and familiar examples
- 3 Units: Cloud Computing Technology for X-Informatics
- 2 Units: X = Web Search and Text Mining and their technologies
- Technology for X-Informatics: Kmeans
- Technology for X-Informatics: MapReduce
- Technology for X-Informatics: Kmeans and MapReduce Parallelism
- Technology for X-Informatics: PageRank
- X = Health
- X = Sensors
- X = Radar for Remote Sensing

Red = Software

The Course

Details of Each Unit

0: Motivation

- Introduction
- Data Deluge
- Jobs
- Industry Trends
- Computing Model: Industry adopted clouds which are attractive for data analytics
- Research Model: 4th Paradigm; From Theory to Data driven science?
- Data Science Process
- Physics-Informatics Looking for Higgs Particle with Large Hadron Collider LHC
- Recommender Systems
- Web Search and Information Retrieval
- Cloud Applications in Research
- Parallel Computing and MapReduce
- Data Science Education: Opportunities at Universities
- Conclusions

3 Units: X-Informatics Introduction: What is Big Data, Data Analytics and X-Informatics? Parts I-III

- What is X-Informatics and its rallying cry
- Jobs
- Data Deluge -- General Structure
- Data Science -- Process
- Data Deluge -- Internet
- Data Deluge – Business
- Data Deluge Science & Research
- Data Deluge Implications for Scientific Method
- Data Deluge Long Tail of Science
- Data Deluge Internet of Things
- Clouds
- Features of Data Deluge
- Processing Big Data
- Data Science Process
- Data Analytics

Python for Big Data and X-Informatics: NumPy, SciPy, MatPlotlib

- Introduction
- Canopy
- Numpy 1
- Numpy 2
- Numpy 3
- Matplotlib 1
- Matplotlib 2
- Scipy 1
- Scipy 2

4 Units: X-Informatics Physics Use Case, Discovery of Higgs Particle; Counting Events and Basic Statistics Parts I-IV

- Looking for Higgs Particle and Counting Errors Introduction
- Physics-Informatics Looking for Higgs Particle Experiments
- Physics-Informatics Accelerator Picture Gallery
- Event Counting
- Examples of Event Counting With Python examples of Signal plus Background
- Fundamental Idea: Random Variables
- Physics and Random Variables
- Statistics of Events with Normal Distributions
- Random Numbers: Generators and Seeds
- Random Numbers: Binomial Distribution
- Accept-Reject
- Monte Carlo Method
- Poisson Distribution
- Central Limit Theorem

Using Plotviz Software for Displaying Point Distributions in 3D

- Motivation and Introduction to use
- Example of Use I: Cube and Structured Dataset
- Example of Use II: Proteomics and Synchronized Rotation
- Example of Use III: More Features and larger Proteomics Sample
- Example of Use IV: Tools and Examples
- Example of Use V: Final Examples

3 Units: X-Informatics Case Study: e-Commerce and Life Style Informatics: Recommender Systems Parts I-III

- Recommender Systems as an Optimization Problem
- Recommender Systems Introduction
- Kaggle Competitions
- Examples of Recommender Systems
- Netflix on Recommender Systems
- Consumer Data Science
- Recap of Recommender Systems
- Examples of Recommender Systems
- Vector Space Formulation of Recommender Systems
- User-based nearest-neighbor collaborative filtering
- Item-based Collaborative Filtering
- Technologies: K-Nearest Neighbors and High Dimensional Spaces

2 Units: X-Informatics Technologies : K-Nearest Neighbor Algorithms and Clustering

- Python K-Nearest Neighbor Algorithms
- Visualization
- Testing K-Nearest Neighbor Algorithms
- Kmeans Clustering
- Clustering of Recommender System Example
- Clustering of Recommender Example into more than 3 Clusters
- Local Optima in Clustering
- Clustering in General
- Heuristics

Parallel Computing: Overview of Basic Principles with familiar Examples

- Decomposition
- Parallel Processing in Society
- Parallel Processing for Hadrian's Wall

3 Units: X-Informatics Cloud Technology Parts I-III

- Cyberinfrastructure for e-moreorlessanything or moreorlessanything-Informatics
- What is Cloud Computing: Introduction
- What is Cloud Computing: Several Other Views
- Gartner's Emerging Technology Landscape
- Gartner's Emerging Technology Landscape for 2012
- Gartner's General Technology Landscape: Other Observations
- Simple Examples of use of Cloud Computing
- What is Cloud Computing in more detail
- Introduction to Cloud Architecture: NaaS IaaS PaaS SaaS
- Platform as a Service
- Data in the Cloud
- Cloud (Data Center)Architectures
- Cloud Industry Players
- Cloud Applications
- Security
- Comments on Fault Tolerance and Synchronicity Constraints
- Big Data Processing from an application perspective

2 Units: X-Informatics Web Search and Text Mining Parts I and II

- Data Mining Survey
- Web and Document/Text Search: The Problem
- Web Search Solution in General starting with history
- Information Retrieval (Web Search) Technology
- Boolean Query
- Fuzzy Index
- Vector Space Model
- Probabilistic Models
- Frequency v. Bayes
- Data Analytics for Web Search
- Document Preparation
- Inverted Index
- Index Construction
- Query Structure and Processing
- Link Structure Analysis including PageRank
- Summary Issues
- Crawling the Web
- Web Advertising and Search
- Clustering and Topic Models

X-Informatics Technologies

- **Technology for X-Informatics: Kmeans (Python Track)**
- Kmeans in Python
- Analysis of 4 Artificial Clusters
- **Technology for X-Informatics: MapReduce**
- MapReduce: Introduction
- MapReduce: Advanced Topics
- **Technology for X-Informatics: Kmeans and MapReduce Parallelism (Python Track)**
- MapReduce Kmeans in Python
- **Technology for X-Informatics: PageRank (Python Track)**
- PageRank in Python: Calculate PageRank from Web Linkage Matrix
- Calculate PageRank of a real page

X-Informatics: Health Informatics

- Big Data and Health
- McKinsey Report on the big-data revolution in US health care
- Microsoft Report on Big Data in Health
- EU Report on Redesigning health in Europe for 2020
- Clouds and Health
- Genomics, Proteomics and Information Visualization

X-Informatics: Sensors

- Internet of Things
- Sensor Clouds
- Earth/Environment/Polar Science data gathered by Sensors
- Ubiquitous/Smart cities
- U-Korea (U = Ubiquitous)
- Smart Grid

X-Informatics: Radar Informatics (with application to glaciology)

- Motivation
- Background – Remote Sensing
- Background – Global Climate Change
- Ice Sheet Science
- Radar Overview
- Radar Basics
- What Are We Doing?
- How Are We Doing It?