**Syllabus for CSE 6250**

***Big Data Analytics for Healthcare***

**Course Summary**

Data science plays an important role in many industries. In facing massive amount of heterogeneous data, scalable machine learning and data mining algorithms and systems become extremely important for data scientists. The growth of volume, complexity and speed in data drives the need for scalable data analytic algorithms and systems. In this course, we study such algorithms and systems in the context of healthcare applications.

In healthcare, large amounts of heterogeneous medical data have become available in various healthcare organizations (payers, providers, pharmaceuticals). This data could be an enabling resource for deriving insights for improving care delivery and reducing waste. The enormity and complexity of these datasets present great challenges in analyses and subsequent applications to a practical clinical environment. In this course, we introduce the characteristics of medical data and associated data mining challenges on dealing with such data. We cover various algorithms and systems for big data analytics. We focus on studying those big data techniques in the context of concrete healthcare analytic applications such as predictive modeling, computational phenotyping and patient similarity. We also study big data analytic technology:

* scalable machine learning algorithms such as online learning and fast similarity search;
* big data analytic system such as Hadoop family (Hive, Pig, HBase), Spark and Graph DB

**Learning Outcomes**

By the end of this course, students should be able to:

1. Formulate data analytic problems in health applications, such as clinical predictive modeling, phenotyping.
2. Develop scalable machine learning algorithms that can be used to analyze complex health related data.
3. Gain hands-on experiences with big data tools such as Hadoop and Spark for healthcare applications.
4. Collect, pre-process and analyze health related datasets such as electronic health records and medical claims.

**Prerequisites and Requirements**

Students must choose one of the following courses as a pre-requisite to this course:

CS 7641 Machine Learning or CSE 6740 Computational Data Analysis or CS 7646 Machine Learning for Trading or CSE 6242 Data and Visual Analytics

To be successful in the course, students should have acquired:

* Basic machine learning and data mining concepts such as classification and clustering;
* Proficient programming and system skills in python, java and scala;
* Proficient knowledge and experience in dealing with data (recommended skills include SQL, NoSQL such as MongoDB).

**Assignments and Exams**

There will be 4 homework assignments and one course project.

The 4 homework sets will account for a total of 40%.

A course project, involving health analytics with big data tools, will account for the final 45% of the grade.

The rest 15% will go towards participation and peer feedback.

**Academic Integrity**

Academic dishonesty will not be tolerated. This includes cheating, lying about course matters, plagiarism, or helping others commit a violation of the Honor Code. Plagiarism includes reproducing the words of others without both the use of quotation marks and citation. Students are reminded of the obligations and expectations associated with the Georgia Tech Academic Honor Code and Student Code of Conduct, available online at [www.honor.gatech.edu](http://www.honor.gatech.edu)

**Learning Accommodations**

If needed, we will make classroom accommodations for students with documented disabilities. These accommodations must be arranged in advance and in accordance with the Office of Disability Services ([www.disabilityservices.gatech.edu](http://www.disabilityservices.gatech.edu)).

**Course materials**

The principal resource for the course are the slides used in lecture. We also create [lab materials online](http://www.sunlab.org/teaching/cse8803/lab/) which can be accessed directly. We also provide papers for reading, which be provided as PDFs at the time of lecture.

**Syllabus**

A brief outline of units of the most recently offered section of the class is given below, grouped into 11 parts.

**1. Introduction**

* What is big data analytics?
* Overview of this course
* Examples of health analytic applications

**2. Predictive Modeling**

* Pipeline of predictive modeling
* Predictive modeling evaluation metrics

**3. Classification methods**

* Scalable classification methods
* Stochastic gradient descent
* Ensemble methods

**4. Clustering**

* Scalable clustering methods
* DBScan
* Mini-batch k-means

**5. MapReduce**

* MapReduce programming model
* Hadoop distributed file system (HDFS)

Analytic algorithms using MapReduce

**6. Phenotyping**

* Phenotyping applications
* Phenotyping process
* Phenotyping algorithms

**7. Dimensionality reduction**

* Singular value decomposition
* CUR decomposition
* Tensor factorization

**8. Patient Similarity**

* Locally supervised distance metric
* Graph based similarity metrics

**9. Medical ontology**

* Medical coding schemes for diagnosis, medical, procedure and lab results
* Systematized Nomenclature of Medicine (SNOMED)
* Unified Medical Language System (UMLS)

**10. Graph Analysis**

* Graph algorithms
* Page Rank
* Spectral clustering

**11. Spark**

* Spark programming model
* Resilient distributed dataset (RDD)
* Spark operations (transformation, actions)