big data

Suggestions:

linear algebra sparse matrices matrix sketching graphs SVD - decompositions optimization, gradient descent

## CSC 134 DBMS

Entity-Relationship (ER) model; relational model; relational database design by ER-to-relational mapping; design of applications using database technology; SQL; schema definition, constraints, and queries; relational algebra; data normalization; access methods such as indexing and hash structures; introduction to transaction processing.

## Data Analytics and Mining.

Theory and implementation of techniques for data analytics and mining with emphasis on big data. Topics include data cleaning, exploratory data analysis, data visualization, feature engineering, classification, clustering, association rule mining, predictive model evaluation, parameter tuning, natural language processing, and selected advanced data mining topics. Design and implementation of systems using contemporary data analysis and mining programming libraries for automatic discovery of patterns and knowledge.

## Timeline

1. Memory Efficient Data Processing (5 weeks)
   1. Streaming statistics and single pass algorithms
   2. Binning and discretization
   3. Pipelines for processing data elements in sequence
   4. Processing large data in chunks
   5. Split-apply-combine strategy for grouped data
2. Writing programs that scale (3 weeks)
   1. Profiling and measurements to diagnose performance problems
   2. Improving performance through parallel programming
   3. Client-server model
   4. Querying remote data sources
   5. Avoiding data movement by running code remotely
3. Handling Complex Data (2 weeks)
   1. Unexpected data and error handling
   2. Text processing and regular expressions
   3. Hierarchical and nested data
4. Applications of big data analysis (5 weeks)
   1. Image processing
   2. Graph representation through sparse matrices
   3. Dimension reduction
   4. Principal components analysis (PCA)
   5. Other data analysis techniques, TBD by class interest

Interactive vs. batch job submissions

Databases: tables, joins, and relational model

Structured Query Language (SQL); declarative versus procedural languages

Nested Data: JSON, XML

Binary Data formats: HDF5, Parquet

Remote API’s: Interacting with remote data sources through a programming language

Parallel Programming (3 weeks)

Functional programming

MapReduce and data parallelism

Multiprocessing in Python