SCALABLE DATA SCIENCE

PROF. ANIRBAN DASGUPTA
Department of Computer Science and Engineering IIT Kharagpur

PROF. SOURANGSHU BHATTACHARYA
Department of Computer Science and Engineering IIT Gadhinagar

INTENDED AUDIENCE : Computer Science & Engineering

PRE-REQUISITES : : Algorithms, Machine Learning

INDUSTRIES APPLICABLE TO : Google, Microsoft, Facebook, Amazon, Flipkart, LinkedIn etc.

COURSE OUTLINE :

Consider the following example problems: One is interested in computing summary statistics (word count distributions) for a set of words which occur in the same document in entire Wikipedia collection (5 million documents). Naive techniques, will run out of main memory on most computers. One needs to train an SVM classifier for text categorization, with unigram features (typically ~10 million) for hundreds of classes. One would run out of main memory, if they store uncompressed model parameters in main memory. One is interested in learning either a supervised model or find unsupervised patterns, but the data is distributed over multiple machines. Communication being the bottleneck, naïve methods to adapt existing algorithms to such a distributed setting might perform extremely poorly. In all the above situations, a simple data mining / machine learning task has been made more complicated due to large scale of input data, output results or both. In this course, we discuss algorithmic techniques as well as software paradigms which allow one to develop scalable algorithms and systems for the common data science tasks.

ABOUT INSTRUCTOR :

Prof. Anirban Dasgupta is currently an Associate Professor of Computer Science & Engineering at IIT Gandhinagar. Prior to this, he was a Senior Scientist at Yahoo! Labs Sunnyvale. Prof. Anirban works on algorithmic problems for massive data sets, large scale machine learning, analysis of large social networks and randomized algorithms in general. He did his undergraduate studies at IIT Kharagpur and doctoral studies at Cornell University.

Prof. Sourangshu Bhattacharya is an Assistant Professor in the Department of Computer Science and Engineering, IIT Kharagpur. He was a Scientist at Yahoo! Labs from 2008 to 2013, where he was working on prediction of Click-through rates, Ad-targeting to customers, etc on the Rightmedia display ads exchange. He was a visiting scholar at the Helsinki University of Technology from January - May 2008.

COURSE PLAN :

Week 01 : Background: Introduction | Probability: Concentration inequalities | Linear algebra: PCA, SVD | Optimization: Basics, Convex, GD | Machine Learning: Supervised, generalization, feature learning, clustering.

Week 02 : Memory-efficient data structures: Hash functions, universal / perfect hash families | Bloom filters| Sketches for distinct count | Misra-Gries sketch | Statistical Mechanics an overview.

Week 03 : Memory-efficient data structures (contd.): Count Sketch, Count-Min Sketch | Approximate near neighbors search: Introduction, kd-trees etc | LSH families, MinHash for Jaccard, SimHash for L2.

Week 04 : Approximate near neighbors search: Extensions e.g. multi-probe, b-bit hashing, Data dependent variants | Randomized Numerical Linear Algebra Random projection.

Week 05 : Randomized Numerical Linear Algebra CUR Decomposition | Sparse RP, Subspace RP, Kitchen Sink.

Week 06 : Map-reduce and related paradigms Map reduce - Programming examples - (page rank, k-eans, matrix multiplication) | Big data: computation goes to data. + Hadoop ecosystem.

Week 07 : Map-reduce and related paradigms (Contd.) Scala + Spark (1 hr) Distributed Machine Learning and Optimization: Introduction | SGD + Proof.

Week 08 : Distributed Machine Learning and Optimization: ADMM + applications | Clustering | Conclusion.