**Indian Institute of Technology, Kanpur**

**Proposal for a New Course**

1. Course No: CS4xx

2. Course Title: Algorithms for Big Data

3. Per Week Lectures: **03** (L), Tutorial: **0** (T), Laboratory: **0** (P), Additional Hours[0-2]: **0** (A), Credits (3*L+2*T+P+A): **09** Duration of Course: **Full Semester**

4. Proposing Department/IDP : **Computer Science and Engineering**

   Other Departments/IDPs which may be interested in the proposed course:

   Other faculty members interested in teaching the proposed course:

5. Proposing Instructor(s): **Sumit Ganguly**

6. Course Description:

   A) Objectives:

Today applications such as machine learning and numerical linear algebra work with massive data. In this course we will cover algorithmic techniques, models, and lower bounds for handling such data. A common theme is the use of randomized methods, such as sketching and sampling, to provide dimensionality reduction. In the context of optimization problems, this leads to faster algorithms, and we will see examples of this in the form of least squares regression and low rank approximation of matrices and tensors, as well as robust variants of these problems. In the context of distributed algorithms, dimensionality reduction leads to communication-efficient protocols, while in the context of data stream algorithms, it leads to memory-efficient algorithms. We will study some of the above problems in such models, such as low rank approximation, but also consider a variety of classical streaming problems such as counting distinct elements, finding frequent items, and estimating norms. Finally we will study lower bound methods in these models showing that many of the algorithms we covered are optimal or near-optimal. Such methods are often based on communication complexity and information-theoretic arguments.

   B) Contents (preferably in the form of 5 to 10 broad titles):

| S. No | Broad Title | Topics | No. of Lectures |
|---|---|---|---|
| 1 | Sketching and Dimensionality Reduction | Estimating $F_0$ −count of distinct elements in a data stream, estimating $l_2$ norm, Johnson-Lindenstrauss' (J-L) Lemma with multiple proofs, Fast J-L (Subsampled Randomized Hadamard Matrix) and Sparse J-L (CountSketch) constructions, $l_p$ norm estimation for $p \in (0,2)$, heavy-hitters via CountSketch and a deterministic algorithm. | 10 |

| | | | |
|---|---|---|---|
| 2 | Lower Bounds | Lower bounds for $F_0$, Johnson-Lindenstrauss' Lower bound, Information theory, Information Theory, Distances Between Distributions, Indexing, streaming lower bounds for norms. | 6 |
| 3 | Approximate Linear Algebra via Dimensionality Reduction | Least Squares Regression, Subspace Embeddings, epsilon-Net Argument, Matrix Chernoff Bound, Affine Embeddings, Approximate Matrix Product, Low Rank Approximation, Sketching-based Preconditioning, Leverage Score Sampling, Distributed Low Rank Approximation, weighted low rank approximation, M-Estimators | 16-18 |
| 4 | Compressive Sensing | Compressive sensing, RIP, L1 minimization, Sparse recovery using sparse matrices, RIP1 | 4 |
| 5* | Sparse Fourier | Sparse Fourier Transform | 3 |
| 6* | Computational Geometry and graphs | Streaming Algorithms for Geometric problems, streaming algorithms for certain graph problems. | 4 |
| **Total** | | | **43-45** |

*Lectures are assumed to be 80 mins each. Topics 5 and 6 may be covered depending on time. Certain topics in 3 are also dependent on available time.

C) Pre-requisites, if any (examples: a- PSO201A,or b- PSO201A or equivalent):

Solid understanding of Probability (CS203M, MSO201 or equivalent) and Linear Algebra (MTH201A or equivalent).

D) Short summary for including in the Courses of Study Booklet.

This is a mathematically rigorous course on developing a variety of algorithms used for processing massive data, based on random sampling and sketching, dimensionality reduction, compressed sensing and sparse Fourier transform. Topics: Estimating $F_0$ and $l_0$-number of distinct items in a data stream, estimating $l_2$and the Johnson-Lindenstrauss' Lemma, estimating $l_p$ using p-stable distributions for $p \in (0,2)$, estimating $l_p$ for $p > 2$ and closely related problems. Johnson-Lindenstrauss' Lemma: Fast J-L, Sparse J-L, Applications to Numerical linear algebra: linear regression and subspace embeddings, affine embeddings, approximate matrix product, low rank approximation, leverage score sampling. Compressive Sensing RIP, $l_1$minimization, sparse recovery using sparse matrices, sparse Fourier transform, streaming algorithms for geometric problems and graph problems. Lower bounds for $F_0$, Johnson-Lindenstrauss' Lower bound, Information theory, Indexing, lower bounds for estimating norms in the streaming model.

7. Recommended books:

   Reference book: Sketching as a Tool for Numerical Linear Algebra by David Woodruff,

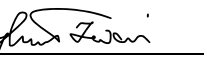
   arXiv:1411.4357.

   Other useful resources: Materials from the following related courses overlap significantly and may be useful in various parts of the course:

   1. Algorithms for Big Data 2025 David Woodruff (CMU)
   2. Algorithms for Big Data 2015 Jelani Nelson and Piotr Indyk (Harvard and MIT)
   3. Algorithmic Techniques for Massive Data: Alexandr Andoni (Columbia).
   4. Algorithms for Big Data: Chandra Chekuri (UIUC).
   5. Algorithms for Big Data: Grigory Yaroslavtsev (UPenn).
   6. Algorithms for Modern Data Models: Ashish Goel (Stanford).
   7. Data Streams Algorithms: Andrew McGregor (UMass Amherst).
   8. Data Stream Algorithms: Amit Chakrabarti (Dartmouth).
   9. Dealing with Massive Data: Sergei Vassilvitskii (Columbia).
   10. Randomized Algorithms for Matrices and Data: Michael Mahoney (UC Berkeley).
   11. Sublinear algorithms: Piotr Indyk, Ronitt Rubinfeld (MIT).
   12. Sublinear and streaming algorithms: Paul Beame (University of Washington).
   13. The Modern Algorithmic Toolbox: Tim Roughgarden, Gregory Valiant (Stanford).
   14. Sublinear Algorithms for Big Data: Qin Zhang (University of Indiana Bloomington)

8. Any other remarks: 1. Versions of this course were taught in the semesters 2019 Sem II, 2021 Sem 1 and II as CS698C. 2. Evaluation will consist of assignments, a mid-semester and end-semester exam, and a paper reading project/experimental project with class presentation and viva.

Dated: 27-02-2026 Proposer: Sumit Ganguly

Dated: 12-03-2026 DUGC/DPGC Convener: 

**The course is approved / not approved**

**Chairman, SUGC/SPGC**

**Dated:**