CS60021: Scalable Data Mining

Instructors
Ritesh Sarkhel
Sourangshu Bhattacharya

Course Logistics

Venue

- Classroom: **NR-413**
- Slots:
  - **Wednesday (11:00 – 11:55)**
  - **Thursday (12:00 – 12:55)**
  - **Friday (08:00-08:55)**
- Website: https://dchakravorty2000.github.io/Scalable_Data_Mining_CS60021_2026
- Moodle (for assignment submission): https://moodlecse.iitkgp.ac.in/moodle/

Tentative Schedule (Changeable)

Week	Date-1	Date-2	Date-3	Topics	Instructor
1	22/7	23/7	24/7	Introduction to DM, ML, Stochastic gradient descent.	RS
2	29/7	30/7	31/7	SGD convergence rate, Accelerated SGD	RS
3	5/8	6/8	7/8	SGD variance reduction	RS
4	12/8	13/8	14/8	Distributed Optimization, ADMM	RS
5	19/8	20/8	21/8	Approximate NNS - Locality Sensitive Hashing	RS
6	26/8	27/8	28/8	ANNS - HNSW	RS
7	2/9	3/9	4/9	Map-reduce framework, Hadoop / Spark	SB
8	9/9	10/9	11/9	DL frameworks - Pytorch	SB
	16/9	17/9	18/9	Mid-sem	
	23/9	24/9	25/9	Mid-sem	
9	30/9	1/10	2/10	Subset selection - Coresets, Sub-modular Optimization	SB
10	7/10	8/10	9/10	Subset Selection in AI - Sparse approximation, Bandit Methods	SB
11	14/10	15/10	16/10	Streaming - Sampling, Set Membership, Distinct Count	SB
12	21/10	22/10	23/10	Streaming - Frequency Counting	SB
13	28/10	29/10	30/10	Transformer Models, LoRA Fine-tuning, Flash Attention	SB
14	4/11	5/11	6/11	Long context modelling	SB

Teaching Assistants

- Debamalya Chakravorty
- Srinjoy Das
- Suman Bera

Say hello to your TA's! Reach out to them in case of doubts.

Office hour: Wednesday 1-2PM CSE 312

Course Background

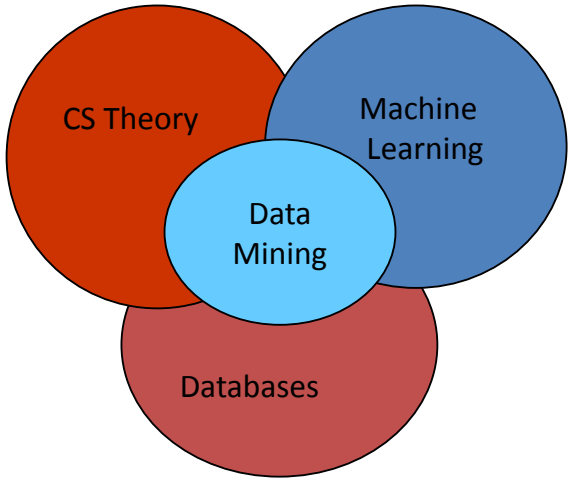
What is Data Mining?

- **Given lots of data discover patterns and models that are:**
  - **Valid:** should hold on new data with some certainty
  - **Non-trivial:** not a common knowledge
  - **Useful:** should be possible to act on the item
  - **Understandable:** humans should be able to interpret the pattern

Data Mining: Cultures

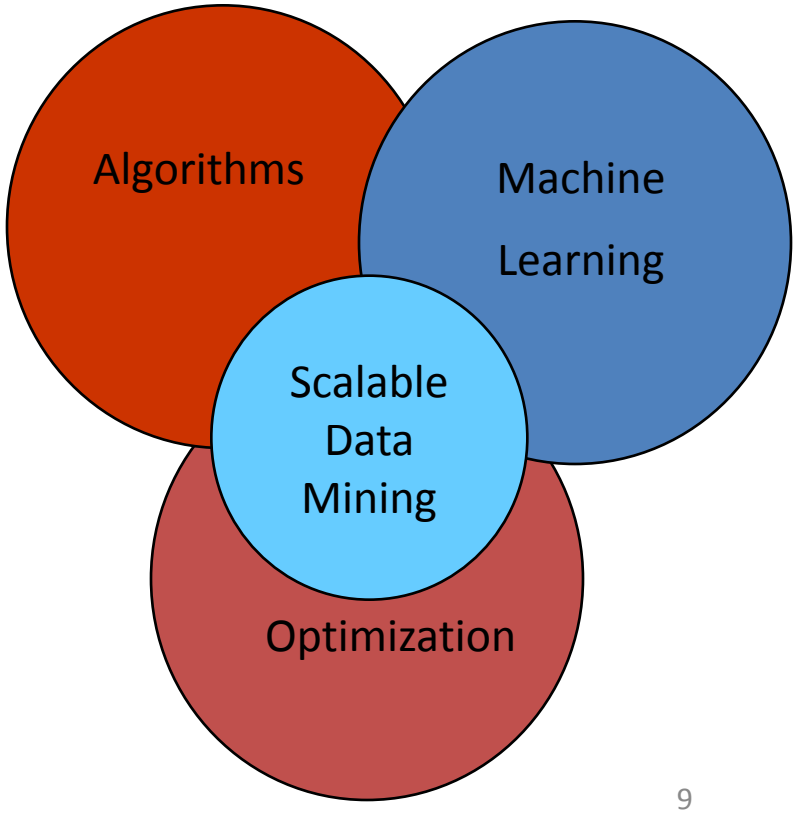
- A lot of the Data Mining Techniques are borrowed from Machine Learning / Deep Learning techniques.

- **Data mining overlaps with:**
  - **Databases:** Large-scale data, simple queries
  - **Machine learning:** Small data, Complex models
  - **CS Theory:** (Randomized) Algorithms



This Course

- **This class overlaps with machine learning, statistics, artificial intelligence, databases but more stress on**
  - **Algorithms**
    - **Online / Streaming**
    - **Optimization**
  - **Computing architectures**



Pre-requisites

- Algorithms
- Machine Learning / Data Analytics / Information Retrieval
- Linear Algebra
- Probability, statistics, calculus

Example Applications

Analytics: Word Count Distribution

- Motivation: Compute word-bigram count distribution for wikipedia corpus.
  - 5 million documents
  - 1.9 million unique words, ? bigrams
- Problem: Input, output and intermediate results are large.
  - Algorithm is simple. Implementation.

Analytics: Word Count Distribution

- Motivation: Compute word-bigram count distribution for wikipedia corpus.
  - 5 million documents
  - 1.9 million unique words, ? bigrams
- Problem: Input, output and intermediate results are large.
  - Algorithm is simple. Implementation.
  - Spark: Map-reduce framework.
- Online version: Sketching algorithms for finding frequency of most frequent items:
  - Misra-Gries Sketching
  - Count-min and Count sketch.

Nearest Neighbor Search (LSH)

- Active learning / Subset selection
  - Calculate pairwise similarity between examples
  - Select examples which provide highest improvement in loss function and are most similar to other non-selected examples.

- Compute similarity to all existing examples in dataset and pick the top ones.
  - Fast nearest neighbor search.

Course Syllabus

Tentative Syllabus

- **Optimization and Machine learning algorithms:**
  - **Optimization algorithms**: Stochastic gradient descent, Variance reduction, Momentum algorithms, ADAM.
  - **Algorithms for distributed optimization:** Stochastic gradient descent and related methods.
  - **Indexing:** Data Structures for Indexing and Learned Index Structures

Tentative Syllabus

- **Software paradigms:**
  - **Big Data Processing:** Motivation and Fundamentals. Mapreduce framework. Functional programming and Scala. Programming using map-reduce paradigm. Example programs.
  - **Deep Learning Frameworks (Pytorch):** Motivation, Computation graphs, Tensors, Autograd, Modules, Example programs.

Syllabus

- **Algorithmic techniques:**
  - **Subset Selection:** Formulations, Coresets, Submodular optimization, Orthogonal Matching Pursuit, Online Convex-optimization.
  - **Finding similar items:** Shingles, Minhashing, Locality Sensitive Hashing families.
  - **Stream processing:** Motivation, Sampling, Bloom filtering, Count-distinct using FM sketch, Estimating moments using AMS sketch.

Syllabus

- **LLM fine-tuning:**
  - Transformer architecture, LORA updates., Flash attention
  - Long context modelling.

Reference Textbooks

- Mining of Massive Datasets. 2nd edition. - Jure Leskovec, Anand Rajaraman, Jeff Ullman. Cambridge University Press. http://www.mmds.org/
- Tensorflow for Machine Intelligence: A hands-on introduction to learning algorithms. Sam Abrahams et al. Bleeding edge press.
- Hadoop: The Definitive Guide. Tom White. O'Reilly Press.
- Recent literature.

Evaluation

- Grades:
  - Mid sem, End sem: 60 - 70
  - Class Test + Assignments: 30 - 40
  - Term Project: 10 - 20
- Assignments: 2 - 3
- Both Term Project and assignment will require you to write code.

Attend classes regularly!

Don’t count on material that you can just “cover” before the exam and get a decent grade.