CS60021: Scalable Data Mining

Sourangshu Bhattacharya

COURSE DETAILS

Teachers

- Teacher:
  - Sourangshu Bhattacharya

- Teaching Assistants:
  - Kiran Purohit
  - Anurag P.

Venue

- Classroom: CSE - 107

- Slots:
  - Monday (8:00 - 9:55)
  - Tuesday (12:00 – 12:55)

- Website: TBA

- Moodle (for assignment submission): https://moodlecse.iitkgp.ac.in/moodle/
- Student key: SDBSB2324

Evaluation

- Grades:
  - Tests: 50
  - Term Project / Assignment: 30
  - Class Test: 20
- Number of Assignments: 3
- Both Term Project and Assignment will require you to write code.

COURSE BACKGROUND

What is Data Mining?

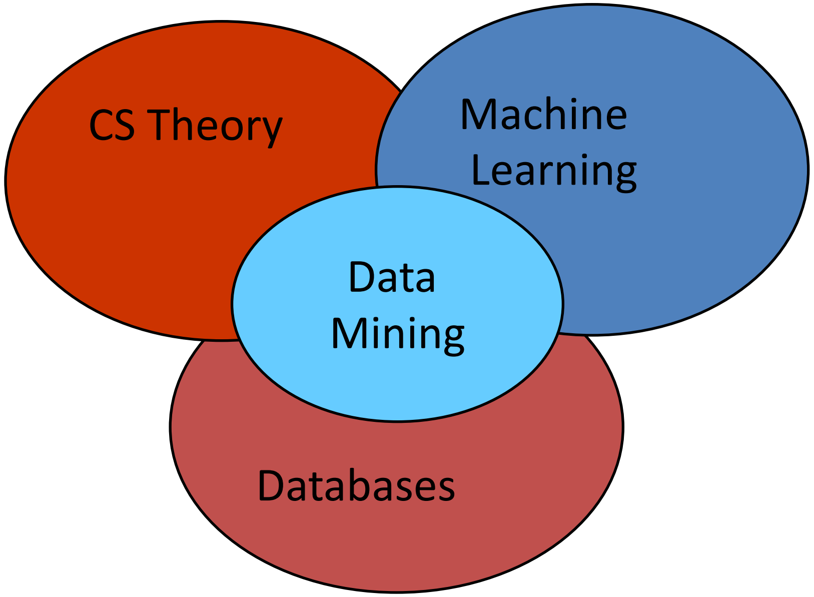
- **Given lots of data**
- **Discover patterns and models that are:**
  - **Valid:** should hold on new data with some certainty
  - **Useful:** should be possible to act on the item
  - **Unexpected:** non-obvious to the system
  - **Understandable:** humans should be able to interpret the pattern

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

Data Mining: Cultures

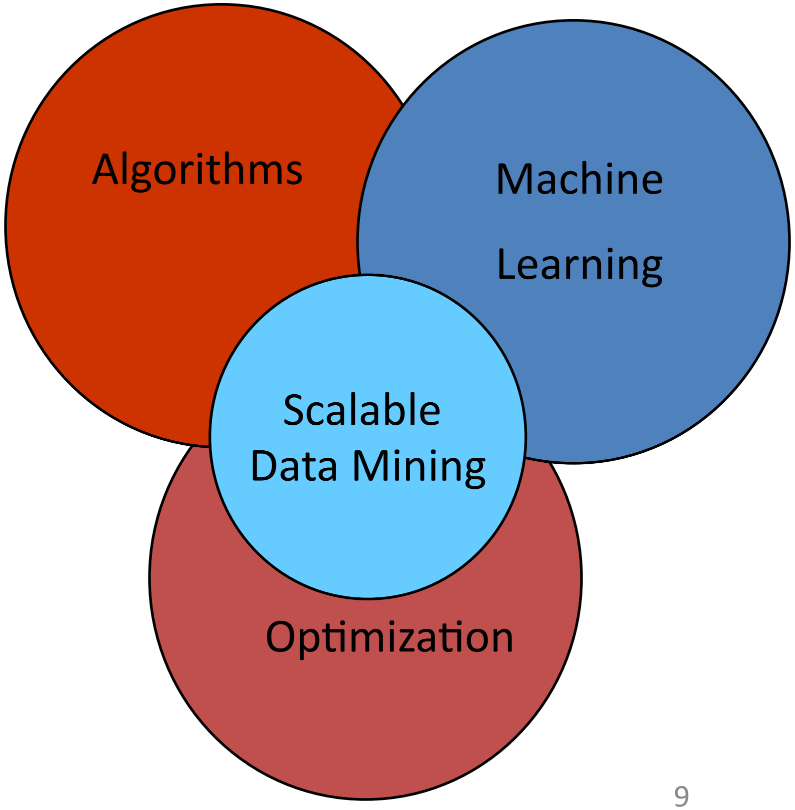
- **Data mining overlaps with:**
  - **Databases:** Large-scale data, simple queries
  - **Machine learning:** Small data, Complex models
  - **CS Theory:** (Randomized) Algorithms

- **In this class, we will explore scalable algorithms and systems for Data Mining.**



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

Word Count Distribution

- 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.
  - You are allowed to use multiple computers.
  - Algorithm is simple.
- Use Map-reduce / Spark

Large Scale Machine Learning

- Train massive deep learning models on massive datasets.
- Dataset too large:
  - Speed up train by speeding up optimization
  - Acceleration techniques.
- Dataset distributed / privacy concerns:
  - Distributed optimization.
  - Federated Learning.
- Model is too complex:
  - Use GPU to train
  - Pytorch.

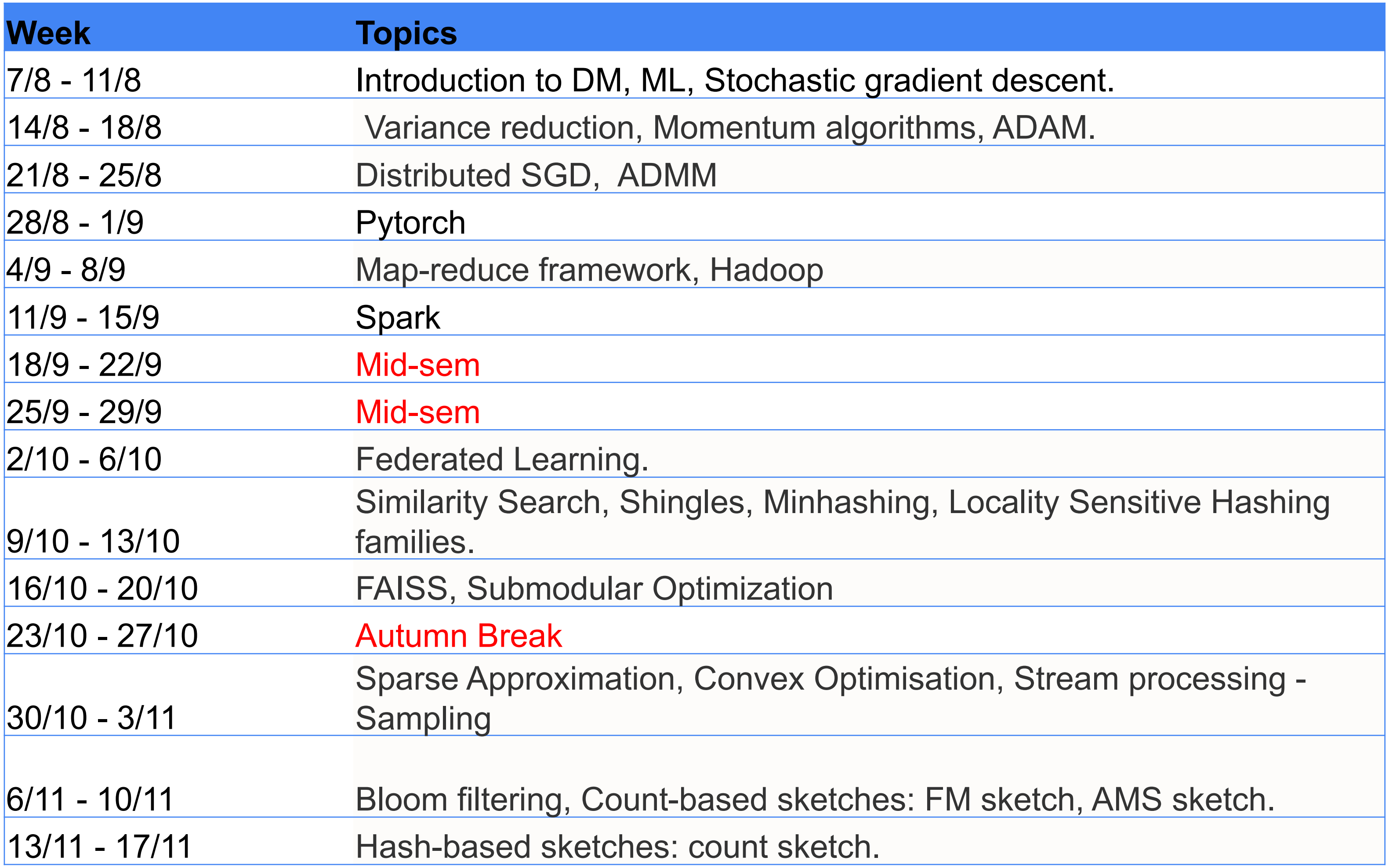
Algorithmic Techniques

- Distinct items in a stream:
  - Count number of distinct IP addresses passing through a server.
  - Streaming model.
  - Problem: 128^4 IP addresses
  - Approximate sketching: FM sketch, count-min sketch.
- Fast nearest neighbor search.
  - Compute similarity to all existing examples in dataset and pick the top ones.
  - Locality sensitive hashing.
  - FAISS

Subset Selection

- Data subset selection:
  - Select a subset of data which is most informative
  - Measure of “informativeness”
  - Diversity ?
  - Fast algorithms:
    - Submodular
    - Sparse approximation
    - Convex Optimization
- Applications:
  - Filter-selection in neural networks
  - Selecting frames to skip in streaming videos.

Tentative Syllabus



Week	Topics
7/8 - 11/8	Introduction to DM, ML, Stochastic gradient descent.
14/8 - 18/8	Variance reduction, Momentum algorithms, ADAM.
21/8 - 25/8	Distributed SGD, ADMM
28/8 - 1/9	Pytorch
4/9 - 8/9	Map-reduce framework, Hadoop
11/9 - 15/9	Spark
18/9 - 22/9	Mid-sem
25/9 - 29/9	Mid-sem
2/10 - 6/10	Federated Learning.
9/10 - 13/10	Similarity Search, Shingles, Minhashing, Locality Sensitive Hashing families.
16/10 - 20/10	FAISS, Submodular Optimization
23/10 - 27/10	Autumn Break
30/10 - 3/11	Sparse Approximation, Convex Optimisation, Stream processing - Sampling
6/11 - 10/11	Bloom filtering, Count-based sketches: FM sketch, AMS sketch.
13/11 - 17/11	Hash-based sketches: count sketch.

THANKS !