CS60021: Scalable Data Mining

Sourangshu Bhattacharya

COURSE BACKGROUND

What is Data Mining?

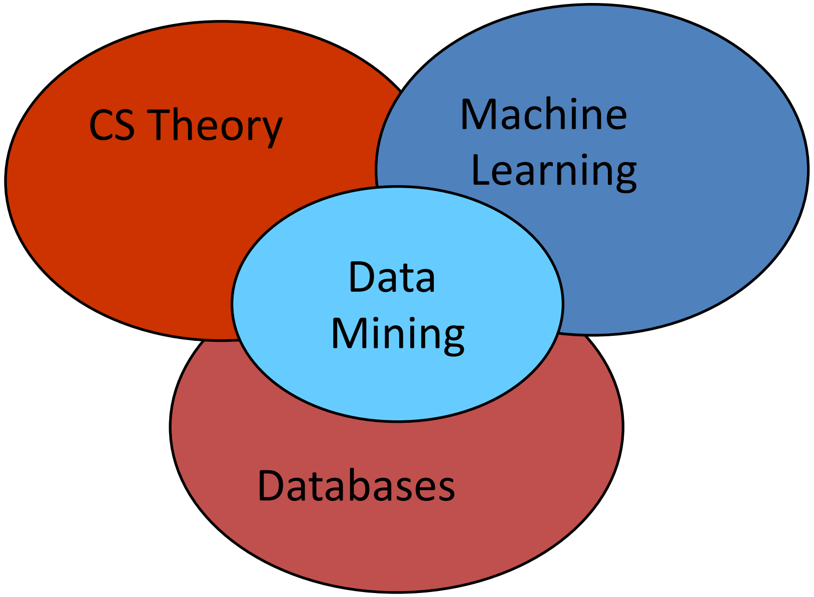
- **Given lots of data**
- **Discover patterns and models that are:**
  - **Valid:** should hold on new data with some certainty
  - **Useful:** should be possible to act on the item
  - **Unexpected:** non-obvious to the system
  - **Understandable:** humans should be able to interpret the pattern

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

Data Mining: Cultures

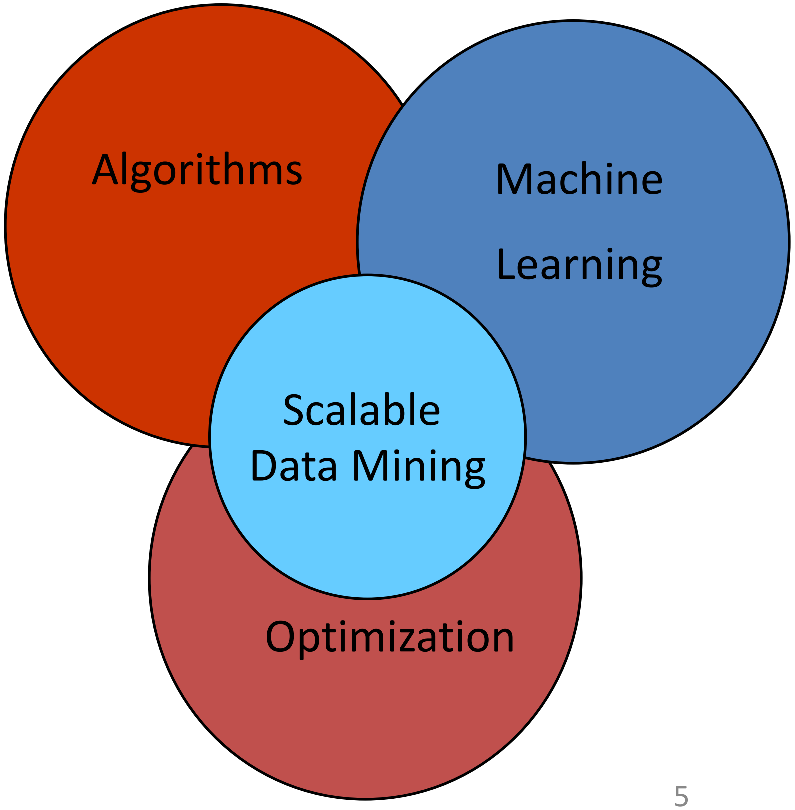
- **Data mining overlaps with:**
  - **Databases:** Large-scale data, simple queries
  - **Machine learning:** Small data, Complex models
  - **CS Theory:** (Randomized) Algorithms

- **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.
- Algorithm is simple.

Large Scale Machine Learning

- Train Massive deep learning models on massive datasets.
- Dataset too large:
  - Speed up train by speeding up optimization
  - Acceleration techniques
  - Distributed optimization.
- Model size too big:
  - Reduce redundant parameters using LSH
  - Change model architecture.

Distinct items

- Count number of distinct IP addresses passing through a server.
- Streaming model.
- Problem: 128^4 IP addresses

- We want only an estimate – FM sketch.

Locality Sensitive Hashing

- 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 seach.

Syllabus

- **Software paradigms:**
  - **Big Data Processing:** Motivation and Fundamentals. Map-reduce 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

- **Optimization and Machine learning algorithms:**
  - **Optimization algorithms**: Stochastic gradient descent, Variance reduction, Momentum algorithms, ADAM. Dual-coordinate descent algorithms.
  - **Algorithms for distributed optimization:** Stochastic gradient descent and related methods. ADMM and decomposition methods, Federated Learning.

Syllabus

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

COURSE DETAILS

Venue

- Classroom: CSE - 119
- Slots:
  - Monday (8:00 - 9:55)
  - Tuesday (12:00 – 12:55)
- Website: http://cse.iitkgp.ac.in/~sourangshu/coursefiles/cs60021_2022a.html
- Moodle (for assignment submission): https://moodlecse.iitkgp.ac.in/moodle/

Teaching Assistants

- Soumi Das
- Kiran Purohit

Evaluation

- Grades:
  - Tests: 50
  - Term Project: 10
  - Class Test: 20
  - Assignment: 20

- Number of Assignments: 2 – 4
- Both Term Project and assignment will require you to write code.

THANKS !