6.897: Algorithms for Massive Data Sets

Piotr Indyk

Preferred time: Tu/Thu 4:00-5:30 pm
Place: Room 34-302
Prerequisites: 6.046 or permission of instructor

Large data sets arise everywhere nowadays (the Web is a notable example). The large amount of information needs to be processed and analyzed. How to process data of this size ? Very often, algorithms that are typically considered efficient, are too slow for massive data (e.g., consider a quadratic-time algorithm applied to a few gigabytes of data). Thus, new algorithmic techniques and paradigms need to be developed.

This course will present a systematic overview of algorithms designed to process large amounts of data, with the emphasis on the most recent developments. Principal topics include:

- Computational models: main memory, external memory, streaming.
- Algorithms for:
  - numerical data (sorting, searching)
  - geometric data in low and high dimensions (including nearest neighbor search, dimensionality reduction, SVD, clustering)
  - text data (inverted indices, suffix trees)
  - graph data (web link analysis, graph partitioning techniques)
  - error correction
  - data compression