



FRONTIERS IN MASSIVE DATA ANALYSIS

Resources, Trade-offs, and Limitations (chapter #6)

INTRODUCTION

Massive data computation uses many types of resources. At a high level, they can be partitioned into the following categories:

- ***Computational resources*** (space, time, number of processing units, and the amount of communication between them)
- ***Statistical or information-theoretic resources*** (number of data samples and their type)
- ***Physical resources*** (amount of energy used during the computation)



RELEVANT ASPECTS OF THEORETICAL COMPUTER SCIENCE

1. Tractability and Intractability
2. Sublinear, Sketching, and Streaming Algorithms
3. Communication Complexity
4. External Memory
5. Parallel Algorithms
6. Computational Learning Theory



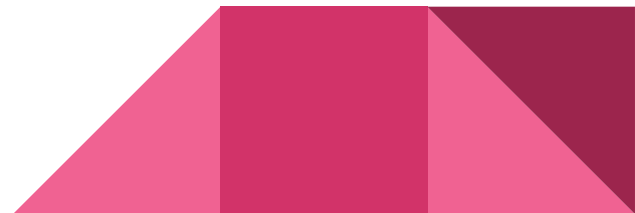
Tractability and Intractability

Most of the natural problems are known to fall into one of these two categories:

1. **Tractable** - problems that have polynomial-time algorithms are often referred.
2. **Intractable** - problems for which such algorithms are conjectured to not exist.

Main Problem:

NP-complete - although any given solution to an NP-complete problem can be verified in polynomial time, there is no known efficient way to locate a solution in the first place.



Sublinear, Sketching, and Streaming Algorithms

Sublinear algorithms are characterized as using an amount of resources (e.g., time or space) that is much smaller than the input size, often exponentially smaller.

Popular computational model:

Data-stream computing - the data need to be processed “on the fly”—i.e., the algorithm can make only a single pass over the data, and the storage used by the algorithm can be much smaller than the input size.



Communication Complexity

1. The amount of information that needs to be extracted from the input, or communicated between two or more parties sharing parts of the input, to accomplish a given task.
2. In contrast to NP-completeness, communication complexity techniques make it possible to prove that some tasks cannot be accomplished using limited communication.



External Memory

1. Focus on the cost of transferring data between the fast local memory and slow external memory (e.g., a disk).
2. The complexity of an algorithm is then measured by the total number of I/O operations that the algorithm performs.
3. External memory algorithms are “cache-aware”; that is, they must be supplied with the amount of available main memory before they can proceed.



Parallel Algorithms

Which problems one can obtain a speedup using parallelism?

Perhaps the one that has attracted the greatest amount of attention is the class of problems having polynomial time sequential algorithms for which one can obtain exponential speedups by using parallelism.



Computational Learning Theory

Computational aspects of extracting knowledge from data.

The question is:

How much data and computational resources are needed in order to “learn” a concept of interest with a given accuracy and confidence?



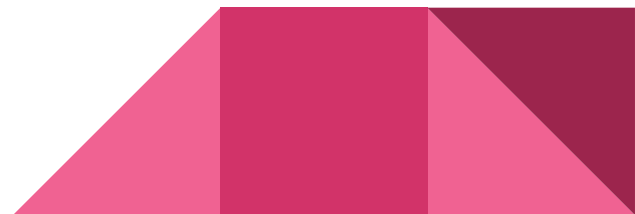
GAPS AND OPPORTUNITIES

1. Challenges for Computer Science:

- Computational Hardness of Massive Data Set Problems
- The Role of Constants
- New Models for Massive Data Computation

2. Challenges for Other Disciplines:

- Statistics
- Physical Resources



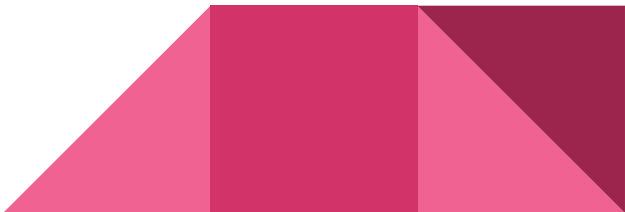
Computational Hardness of Massive Data Set Problems

It involves:

1. Defining more refined boundaries between the tractable and the intractable that model the massive-data computation more accurately.
2. Identifying new “hard” problems that are (conjectured to be) unsolvable within those boundaries.



The Role of Constants

1. Optimizing constant factors is often thought to belong to algorithm engineering rather than algorithm design.
 2. Exception to this trend includes the study of classic problems like median finding or sorting, especially in the context of average-case analysis.
 3. Ignoring constant factors can obscure the dependencies on implicit data parameters, such as the dimension of the underlying space, precision, etc.
 4. It is no longer the case that computers are getting faster: the increase in processing power in coming years is projected to come from increased parallelism rather than clock speed.
- 

New Models for Massive Data Computation

Constructing and investigating new models of computation:

- MapReduce
- Hadoop and variations
- Multicores
- Graphic processing units (GPUs)
- Parallel databases



Statistics

The data have some statistical properties – that is, they are a sequence of samples from some distribution or that the data have sparsity or other structural properties.

Open questions:

- In computational learning theory, the computational limitations are typically specified in terms of polynomial-time computability, and thus the limitations of that topic apply.
- Privacy (how much information about the data must be revealed in order to perform some computation or answer some queries about the data?)

Physical Resources

1. The large amount of consumed and dissipated energy is the key reason why the steady increase in processor clock rates has slowed in recent years.
2. The ubiquity of energy-limited mobile computing devices (smart phones, sensors, and so on) has put a premium on optimizing energy use.
3. The impact of computation and data storage on the environment has motivated the development of green computing (Hölzle and Weihl, 2006).



Thank you for your attention

More information available in ["FRONTIERS IN MASSIVE DATA ANALYSIS"](#) book.

