Big Data and Its Technical Challenges

Exploring the inherent technical challenges in realizing the potential of Big Data
What is Big Data?

- Big data is a large volume of unstructured data which cannot be handled by standard database management systems.
Introduction

- Data–driven mathematical models
- Big data affects many aspects of society
- Popular news media now appreciates the value of Big Data

- Sloan Digital Sky Survey
- Biological sciences
- Machine translation
- Facebook, Twitter, Microsoft
- Education, homeland security
Creating value from Big Data is a multi-step process

Most people only focus on analysis/modeling step
Data acquisition

- A record of some underlying activity of interest
- Large volumes of data have been produced everyday
- The planned square kilometer array telescope
- The data collected by sensors often are spatially and temporally correlated
- Loading of large datasets is often a challenge
The information collected will not be in a format ready for analysis
Electronic health records
Information extraction process
Highly application-dependent
Effective large-scale analysis often requires the collection of heterogeneous data from multiple sources.

360-degrees health view of a patient.
Big data is often dynamic, heterogeneous, inter-related, and untrustworthy

one can use suitable statistical care to get good results without being overwhelmed by the volume
Interpretation

- Examine all the assumptions made and retracing the analysis
- Many possible sources of error
- Better support the human thought process and social circumstances
Case Study – TransDec

- enables real-time visualization, querying, and analysis of dynamic transportation systems
- particularly address the fundamental data management and visualization challenges
Case Study – TransDec

Acquisition → Cleaning → Aggregation/Representation → Interpretation → Analysis
TransDec – Acquisition

- Traffic-loop detectors
- Bus and rail
- Ramp meters and CMS
- Special events
Challenges in Big Data Analysis

- Heterogeneity
- Inconsistency and incompleteness
- Scale
- Timeliness
- Privacy and data ownership
- Visualization and collaboration
Big data increasingly includes information provided by diverse sources
Uncertainty, errors, and missing values must be managed
This incompleteness and these errors must be managed during data analysis
Managing large and rapidly increasing volumes of data has been a challenging issue for many decades. Cloud computing enables organizations to outsource Big Data processing.
Real-time techniques is to summarize and filter what is to be stored
Fraudulent credit card
Index structures are created in advance to find qualifying elements quickly
Privacy and data ownership

- Both a technical and a sociological problem
- Inappropriate use of personal data
- Many online services require us to share private information (facebook)

- Electronic health records
- Location-based services
- Religious preferences
- The data owned by an organization
Visualization and collaboration

- The human perspective
- A CAPTCHA is a type of challenge–response test used in computing to determine whether or not the user is human.
- A Big Data analysis system must support input from multiple human experts
- Crowdsourcing – Wikipedia
Conclusion

- An era of Big Data
- This data-driven world has the potential to improve the efficiencies of enterprises and improve the quality of our lives
- A number of challenges must be addressed
- The solutions to a challenge may not be the same in all situations
- Huge rewards waiting for those who use Big Data correctly
The End

Thank you for your listening