CS 63018/73018 PROBABILISTIC DATA MANAGEMENT 3 Credit Hours

Instructor’s Name: Dr. Xiang Lian

Reference Books


Course Content:
(Slashed with CS 73018) This course addresses the fundamental concepts and techniques for probabilistic data management in the area of databases. Probabilistic data are pervasive in many real-world applications, such as sensor networks, GPS system, location-based services, mobile computing, multimedia databases, data extraction and integration, trajectory data analysis, semantic web, privacy preserving, and so on. This class also covers major research topics such as probabilistic or uncertain data models, probabilistic queries, probabilistic query answering techniques, and data quality issues in databases.

Prerequisites or co-requisites: Graduate standing
Required, elective, or selected elective

Goals:
1. Learn the basics of probabilistic data management
2. Understand the data model of probabilistic data
3. Learn important queries over probabilistic data
4. Learn probabilistic query answering techniques for probabilistic data
5. Learn the processing of probabilistic graph databases
6. Understand the data quality issues in probabilistic databases

Outcomes:
7. Explain real applications of probabilistic and uncertain data management in databases.
8. Know the classifications of data uncertainties according to different criteria.
9. Explain the causes and importance of studying probabilistic data management.
10. Describe data uncertainty models, possible worlds semantics, correlations in probabilistic data, and probabilistic graph models.
11. Know various types of probabilistic queries in probabilistic/uncertain databases.
12. Describe the models, problem definitions, and the proposed techniques for each probabilistic query type in the literature.
13. Learn to read/write research papers, and understand the general trend of the research in probabilistic data management.
14. Summarize and analyze the pros and cons of existing works in probabilistic/uncertain databases.
15. Identify one or two future directions in probabilistic databases, which have not been studied before, or not been extensively studied before, to work on.
16. Write a survey on related works of probabilistic data management.
17. Propose new solutions to existing problems or novel solutions to new problems in probabilistic and uncertain data management.
18. Write a research report or research project/paper on the proposed problems or solutions.
19. Do experiments on the proposed ideas in probabilistic data management.
20. Give a presentation on the project report to show off the outcome of the research project. Optionally, give a presentation on the research papers in the literature (with bonus points).
21. Work in a team (each with 2-3 members) to collaboratively write the survey and research papers.

**Topics to be Covered:**

1. **An Overview of Probabilistic Data Management**
   1.1. Examples of probabilistic data in real applications
   1.2. Causes of uncertainties in probabilistic data in applications
   1.3. Importance of studying probabilistic data management
   1.4. Classifications of the data uncertainty
2. **Data Uncertainty Model**
   2.1. Definition of probabilistic data
   2.2. Granularity of data uncertainties
   2.3. Representation of probabilistic data
   2.4. Possible worlds semantics
   2.5. Correlated probabilistic data
   2.6. Probabilistic graph data model
3. **Probabilistic Query Answering Over Probabilistic/Uncertain Databases**
   3.1. Probabilistic range and nearest neighbor queries
   3.2. Probabilistic group nearest neighbor queries
   3.3. Probabilistic reverse nearest neighbor queries
   3.4. Probabilistic top-k queries
   3.5. Probabilistic (reverse) skyline queries
   3.6. Probabilistic join queries
4. **Probabilistic Graph Databases**
   4.1. Probabilistic XML/RDF databases
   4.2. Queries over probabilistic graph databases
5. Data Quality in Probabilistic Databases
   5.1. Inconsistencies in probabilistic databases
   5.2. Cleaning over uncertain data
   5.3. Probabilistic data repairs and consistent query answering

Abet Learning Outcomes:

- Analyze a complex computing problem and to apply principles of computing and other relevant disciplines to identify solutions.
- Design, implement, and evaluate a computing-based solution to meet a given set of computing requirements in the context of the program’s discipline.
- Apply computer science theory and software development fundamentals to produce computing-based solutions.
- Communicate effectively in a variety of professional contexts.
- Recognize professional responsibilities and make informed judgments in computing practice based on legal and ethical principles.
- Function effectively as a member or leader of a team engaged in activities appropriate to the program’s discipline.