CS 848 (Summer 2026)

Advanced Topics in Databases

Data Lake and Model Lake Managerment

Professor: Renée J. Miller

Course Information

- **Room:** DC 2568
- **Seminar Time:** Monday 1-4pm

Overview

In data science, most data is typically not stored in well designed database management systems, rather it resides in massive repositories that are often called data lakes. A data lake stores vast amounts of structured, semi-structured, and unstructured data in its original, raw format. Unlike traditional database management systems that require data to be organized before storage, data lakes allow the ingestion of data from diverse sources without this preprocessing, documentation, and organization. Data lakes have surfaced a host of new data management research problems, In this seminar, we will explore these problems and current solutions. As one example, data discovery is the process of quickly and scalably finding data of interest. This is a challenging problem in data lakes due to the lack of metadata and a consistent data organization. We will also consider model lakes, large repositories of pre-trained AI models. Huggingface is perhaps the most known open model lake, but any data science organization will host its own internal model lake. This seminar explores the recent innovations in data lake management and how some of these innovations are being applied to the management of model lakes. The course is based on weekly paper readings, student presentations, discussions, and a term project.

Course Logistics

- **Weeks 1–3:** Background material presented by the professor or guest lecturers. Students are expected to complete assigned readings.
- **Weeks 4–11:** Two-three paper presentations and discussions per week. Number will depend on enrollment.
- **Week 12-13:** Project presentations.

Workload Breakdown

Presentation (15%)

Each student will present one (or two) papers, critique it, and answer questions. Presentations should be 25–40 minutes. If class size is small, each student may present twice, increasing this component to 20%. If the class is large, strudents may present in pairs. Slides must be submitted by 7pm on the Friday prior to presentation.

Paper Reviews (20%)

Students will review two-three papers per week (except their own presentation paper). Number of papers will depend on enrollment.

- Submission deadline: Friday 7pm before lecture
- Maximum length: 1000 words (approx. 2 pages)
- Format: single column, single-spaced, 12pt font, 1-inch margins

Class Participation (15%)

Active participation in discussions is required and expected. If you are not comfortable asking questions, there will be an option to email questions to Professor prior to class.

Project (50%)

- Group work (2 students per group - may change depending on enrollment)
- Research topic and written report required

.

Weekly Schedule

See course schedule (subject to change depending on class size).

Submission Guidelines

- All submissions must be in PDF format
- Submit via provided Dropbox folder
- File naming format:

  `<week_no>-<last_name>-<type><number>.pdf`

  where `type` is `review` or `slides`

Administrative Issues

Students are expected to review policies on academic integrity and honesty.

Preliminary Reading List

- *Data Lake Management: Challenges and Opportunities* (2019) Authors: F. Nargesian, E. Zhu, R. Miller, K. Pu, P. Arocena (PVLDB).
- *Data Lakes: A Survey of Functions and Systems* (2023) Authors: R. Hai, C. Quix, M. Jarke (IEEE Trans. Knowl. Data Eng.)
- *Model Lakes* (2025) Authors: Koyena Pal, David Bau, R. J. Miller: (EDBT)
- *ModelTables: A Corpus of Tables about Models* (2025) Authors: Zhengyuan Dong, Victor Zhong, R. J. Miller: (CoRR abs/2512.16106)
- *Semantics-aware Dataset Discovery from Data Lakes* (2023) Authors: Grace Fan, Jin Wang, Yuliang Li, Dan Zhang, R. J. Miller: (PVLDB)
- *Table Discovery in Data Lakes: State-of-the-art and Future Direction* (2023) Authors: Grace Fan, Jin Wang, Yuliang Li, R. J. Miller (SIGMOD)
- *Symphony: Towards Trustworthy Question Answering and Verification using RAG over Multimodal Data Lakes* (2024) Authors: Nan Tang, Chenyu Yang, Zhengxuan Zhang, Yuyu Luo, Ju Fan, Lei Cao, Sam Madden, Alon Y. Halevy: (IEEE Data Eng. Bull.)
- *DeepJoin: Joinable Table Discovery with Pre-trained Language Models* (2023) Authors: Yuyang Dong, Chuan Xiao, Takuma Nozawa, Masafumi Enomoto, and Masafumi Oyamada. 2023 (PVLDB)
- *Automatic Generation of Model and Data Cards: A Step Towards Responsible AI* (2024) Authors: Jiarui Liu, Wenkai Li, Zhijing Jin, and Mona T. Diab. (ACL)