CSCI 699: Machine Learning in Databases

Units: 4.0
Spring 2024 – Mondays and Wednesdays – 5:00-6:50PM

Location: TBA https://infolab.usc.edu/csci699/

Instructor: Cyrus Shahabi
Office: PHE-306a
Office Hours: MW 4-5 pm

Contact Info: shahabi@usc.edu

Catalogue Description
Utilization of Machine Learning approaches (e.g., Deep Learning, NLP, reinforcement learning) used for indexing, query optimization, approximate query processing, and system tuning in Databases.

Course Description
The purpose of this course is to bring the student up to date in recent advancements and utilization of Machine Learning (ML) approaches in the field of Databases. The course will first cover various Machine Learning techniques (e.g., Deep Learning, NLP, reinforcement learning) used for indexing, query optimization, approximate query processing, and system tuning; and proceed with techniques used to potentially replace a database with an ML model. Students in the course will be guided through the recent database and ML literature from the VLDB, SIGMOD, ICDE, ICLR, and SIGKDD conferences.

The course assumes that students are familiar with database conceptual data modeling tools such as Entity-Relationship (ER) data model, logical data models such as the relational and object-relational data model, physical design of a database using persistent data structures such as B+-tree and Hash indexes, transactions, concurrency control, and crash recovery techniques. Knowledge of ML and neural networks is also beneficial.

Learning Objectives
The objective of the course is to familiarize the students with the ML techniques used to improve databases. Each student will select a topic from the list of assigned papers and then read all related technical papers from that area. Students will then present the results of their reading in class and either come up with a new and/or improved approach or summarize their readings as a literature survey, resulting in a research paper (ideally submitted to a conference).

Recommended Preparation: knowledge at the level of CSCI-585 or CSCI-587

Credit Restrictions: None.

Technological Proficiency and Hardware/Software Required
Students in this course will utilize a personal laptop or desktop and be familiar with the usage of the ACM Digital Library and Google Search.

Required Readings and Supplementary Materials
Papers, slides, URLS and chapters handed out by the instructor

Grading Breakdown
- Each student is expected to present two presentations in class covering technical papers related to their selected topic. Those presentations are graded.
    - 20 points each presentation, 40 points total
- Additionally, each student will develop a research or survey paper at the end of the course. Those written papers are also graded.
    - 50 points as final deliverable
- Class Participation
    - 10 points as final deliverable.
- Total points - 100

Academic Integrity
This course will follow the expectations for academic integrity as stated in the USC Student Handbook. The general USC guidelines on Academic Integrity and Course Content Distribution are provided in the subsequent "Statement on Academic Conduct and Support Systems" section.

Please ask the instructor or TA if you are unsure about what constitutes unauthorized assistance on an exam or assignment or what information requires citation and/or attribution.

You may not record this class without the express permission of the instructor and all other students in the class. Distribution of any notes, recordings, exams, or other materials from a university class or lectures — other than for individual or class group study — is prohibited without the express permission of the instructor.

Use of Generative AI in this Course

Generative AI is not permitted: Since creating, analytical, and critical thinking skills are part of the learning outcomes of this course, all assignments should be prepared by the student working individually or in groups as described on each assignment. Students may not have another person or entity complete any portion of the assignment. Developing strong competencies in these areas will prepare you for a competitive workplace. Therefore, using AI-generated tools is prohibited in this course, will be identified as plagiarism, and will be reported to the Office of Academic Integrity.

Course Schedule: A Weekly Breakdown

Week	Session	Topics	Readings	Deliverables
1	1	Introduction		
1	2	Learned one dimensional indexing	Indexing	Summary of presented papers
2	3	**Martin Luther King's Birthday**	Indexing	
2	4	Learned multidimensional Indexing	Indexing	Summary of presented papers
3	5	Learned updatable indexing	Indexing	Summary of presented papers
3	6	Sorting	Sorting	Summary of presented papers
4	7	Bloom Filters	Bloom filters	Summary of presented papers
4	8	Cardinality Estimation (data models)	Cardinality Estimation	Summary of presented papers
5	9	Cardinality Estimation (query models)	Cardinality Estimation	Summary of presented papers
5	10	Reinforcement learning for combinatorial Optimization	Combinatorial Optimization	Summary of presented papers
6	11	Query optimization with reinforcement learning	Query Optimization	Summary of presented papers

Week	Session	Topics	Readings	Deliverables
6	12	Query optimization without demonstration. Query rewriting	Query Optimization	Summary of presented papers
7	13	**President's Day**		
7	14	Approximate query processing (data model)	Approximate query processing	Summary of presented papers
8	15	Approximate query processing (query model)	Approximate query processing	Summary of presented papers
8	16	Approximate query processing (privacy/missing data)	Approximate query processing	Summary of presented papers
9	17	Database Tuning with NLP and RL	Database Tuning	Summary of presented papers
9	18	Database Tuning through cost modeling and optimization	Database Tuning	Summary of presented papers
10	19	**Spring Recess**		
10	20	**Spring Recess**		
11	21	DB ML systems (Training data collection)	DB ML systems	Summary of presented papers
11	22	DB ML systems (SQL Understanding and column annotation)	DB ML systems	Summary of presented papers
12	23	Optimized data layout	Data layout	Summary of presented papers
12	24	Caching	Caching	Summary of presented papers
13	25	Learned (LSH for) similarity search	Nearest Neighbour/Similarity Search	Summary of presented papers
13	26	Using LSH to improve ML model efficiency	Nearest Neighbour/Similarity Search	Summary of presented papers

Week	Session	Topics	Readings	Deliverables
14	27	Neural network memorization and generalization	Neural Network Memory	Summary of presented papers
14	28	Neural networks with memory cells	Neural Network Memory	Summary of presented papers
15	29	Students Paper Reviews		Presentation
15	30	Students Paper Reviews		Presentation
FINAL		Final Paper submission	Refer to the final exam schedule in the USC *Schedule of Classes* at classes.usc.edu.	

Course Topics and Reading Lists:

Indexing
The Case for Learned Index Structures
Tim Kraska, Alex Beutel, Ed H. Chi, Jeffrey Dean, Neoklis Polyzotis
SIGMOD 18

ALEX: An Updatable Adaptive Learned Index
Jialin Ding, Umar Farooq Minhas, Jia Yu, Chi Wang, Jaeyoung Do, Yinan Li, Hantian Zhang, Badrish Chandramouli, Johannes Gehrke, Donald Kossmann, David Lomet, Tim Kraska
SIGMOD 20

Updatable Learned Index with Precise Positions
Jiacheng Wu, Yong Zhang, Shimin Chen, Jin Wang, Yu Chen, Chunxiao Xing
VLDB 21

Learning Multi-dimensional Indexes
Vikram Nathan, Jialin Ding, Mohammad Alizadeh, Tim Kraska
SIGMOD 20

Tsunami: A Learned Multi-dimensional Index for Correlated Data and Skewed Workloads. Jialin Ding, Vikram Nathan, Mohammad Alizadeh, Tim Kraska.
VLDB 21.

Effectively learning spatial indices. Jianzhong Qi, Guanli Liu, Christian S Jensen, and Lars Kulik. VLDB 20.

Sorting
The Case for a Learned Sorting Algorithm.
Ani Kristo, Kapil Vaidya, Ugur Çetintemel, Sanchit Misra, Tim Kraska.
SIGMOD 2020.

Defeating duplicates: A re-design of the LearnedSort algorithm.
Ani Kristo, Kapil Vaidya, Tim Kraska.

AIDB 2021.

Bloom filters

A Model for Learned Bloom Filters and Optimizing by Sandwiching.
Michael Mitzenmacher.
NeurIPS 2018

Meta-Learning Neural Bloom Filters.
Jack W Rae, Sergey Bartunov, Timothy P Lillicrap.
ICML 2019.

Partitioned Learned Bloom Filters.
Kapil Vaidya, Eric Knorr, Tim Kraska, Michael Mitzenmacher.
ICLR 2021.

Stacked Filters: Learning to Filter by Structure.
Kyle Deeds, Brian Hentschel, and Stratos Idreos.
VLDB 2021.

Stable Learned Bloom Filters for Data Streams. Qiyu Liu, Libin Zheng, Yanyan Shen, Lei Chen.
VLDB 2020.

Adaptive Learned Bloom Filter (Ada-BF): Efficient Utilization of the Classifier with Application to Real-Time Information Filtering on the Web. Zhenwei Dai, Anshumali Shrivastava.
NeurIPS 2020.

Cardinality Estimation

Monotonic Cardinality Estimation of Similarity Selection: A Deep Learning Approach.
Yaoshu Wang, Chuan Xiao, Jianbin Qin, Xin Cao, Yifang Sun, Wei Wang, Makoto Onizuka.
SIGMOD 2020.

NeuroCard: One Cardinality Estimator for All Tables.
Zongheng Yang, Amog Kamsetty, Sifei Luan, Eric Liang, Yan Duan, Xi Chen, Ion Stoica.
VLDB 2021

Learning-Based Frequency Estimation Algorithms.
Chen-Yu Hsu, Piotr Indyk, Dina Katabi, Ali Vakilian.
ICLR 2019.

Composable Sketches for Functions of Frequencies: Beyond the Worst Case
Edith Cohen Ofir Geri Rasmus Pagh
ICML 20

Flow-Loss: Learning Cardinality Estimates That Matter.
Parimarjan Negi, Ryan Marcus, Andreas Kipf, Hongzi Mao, Nesime Tatbul, Tim Kraska, Mohammad Alizadeh
VLDB 21

Estimating Cardinalities with Deep Sketches
Andreas Kipf, Dimitri Vorona, Jonas Müller, Thomas Kipf, Bernhard Radke, Viktor Leis, Peter Boncz, Thomas Neumann, Alfons Kemper
SIGMOD 19

Selectivity Functions of Range Queries are Learnable
Xiao Hu, Yuxi Liu, Haibo Xiu, Pankaj K Agarwal, Debmalya Panigrahi, Sudeepa Roy, Jun Yang
SIGMOD 22

Combinatorial Optimization

Learning Combinatorial Optimization Algorithms over Graphs
Hanjun Dai, Elias B. Khalil, Yuyu Zhang, Bistra Dilkina, Le Song

Query Optimization

SkinnerDB: regret-bounded query evaluation via reinforcement learning
Immanuel Trummer , Samuel Moseley, Deepak Maram, Saehan Jo, Joseph Antonakakis
VLDB 2018

Neo: A Learned Query Optimizer.
Ryan Marcus, Parimarjan Negi, Hongzi Mao, Chi Zhang, Mohammad Alizadeh, Tim Kraska, Olga Papaemmanouil, Nesime Tatbul.
VLDB 2019.

A Learned Query Rewrite System using Monte Carlo Tree Search.
Xuanhe Zhou, Guoliang Li, Chengliang Chai, and Jianhua Feng.
VLDB 2022.

Balsa: Learning a Query Optimizer Without Expert Demonstrations
Zongheng Yang ; Wei-Lin Chiang ; Sifei Luan; Gautam Mittal; Michael Luo ; Ion Stoica
SIGMOD 2202

Bao: Making Learned Query Optimizers Practical.
Ryan Marcus, Parimarjan Negi, Hongzi Mao, Nesime Tatbul, Mohammad Alizadeh, Tim Kraska.
SIGMOD 2021.

Phoebe: A Learning-based Checkpoint Optimizer.
Yiwen Zhu, Matteo Interlandi, Abhishek Roy, Krishnadhan Das, Hiren Patel, Malay Bag, Hitesh Sharma, Alekh Jindal.
VLDB 2021.

Approximate query processing

DBEST: Revisiting approximate query processing engines with machine learning models. Qingzhi Ma, Peter Triantafillou.
SIGMOD 2019.

Approximate Query Processing for Data Exploration using Deep Generative Models.
Saravanan Thirumuruganathan, Shohedul Hasan, Nick Koudas, Gautam Das.
ICDE 2020.

DeepDB: learn from data, not from queries!.
Benjamin Hilprecht, Andreas Schmidt, Moritz Kulessa, Alejandro Molina, Kristian Kersting, Carsten Binnig.
VLDB 2020.

A Neural Database for Differentially Private Spatial Range Queries
S Zeighami, R Ahuja, G Ghinita, C Shahabi
VLDB 22

A Neural Approach to Spatio-Temporal Data Release with User-Level Differential Privacy
R Ahuja, S Zeighami, G Ghinita, C Shahabi
SIGMOD 23

NeuroDB: A Neural Network Framework for Answering Range Aggregate Queries and Beyond
S Zeighami, C Shahabi

NeuroComplete: A Neural Database for Answering Analytical Queries on Incomplete Relational Data

Sepanta Zeighami, Raghav Seshadri, Cyrus Shahabi

Database Tuning

DB-BERT: A Database Tuning Tool that "Reads the Manual"
Immanuel Trummer
SIGMOD 2022

Budget-aware Index Tuning with Reinforcement Learning
Wentao Wu; Chi Wang; Tarique Siddiqui ; Junxiong Wang; Vivek Narasayya ; Surajit Chaudhuri ; Philip A Bernstein
SIGMOD 2022

The Data Calculator: Data Structure Design and Cost Synthesis From First Principles, and Learned Cost Models
S. Idreos, K. Zoumpatianos, B. Hentschel, M. S. Kester, and D. Guo
SIGMOD 2018

Monkey: Optimal navigable key-value store
N Dayan, M Athanassoulis, S Idreos
SIGMOD 2017

DB ML systems

Active Learning for ML Enhanced Database Systems.
Lin Ma, Bailu Ding, Sudipto Das, Adith Swaminathan.
SIGMOD 2020.

Tastes Great! Less Filling! High Performance and Accurate Training Data Collection for Self-Driving Database Management Systems
Matthew Butrovich, Wan Shen Lim, Lin Ma, John Rollinson, William Zhang, Yu Xia, Andrew Pavlo
SIGMOD 22

PreQR: Pre-training Representation for SQL Understanding
Xiu Tang, Sai Wu, Mingli Song, Shanshan Ying,Feifei Li, Gang Chen
SIGMOD 22

Annotating Columns with Pre-trained Language Models
Yoshihiko Suhara, Jinfeng Li, Yuliang Li, Dan Zhang, Çağatay Demiralp, Chen Chen, Wang-Chiew Tan
SIGMOD 22

Data layout

Qd-tree: Learning Data Layouts for Big Data Analytics. Zongheng Yang, Badrish Chandramouli, Chi Wang, Johannes Gehrke, Yinan Li, Umar Farooq Minhas, Per-Åke Larson, Donald Kossmann, Rajeev Acharya. SIGMOD 2020.

Instance-Optimized Data Layouts for Cloud Analytics Workloads. Jialin Ding, Umar Farooq Minhas, Badrish Chandramouli, Chi Wang, Yinan Li, Ying Li, Donald Kossmann, Johannes Gehrke and Tim Kraska. SIGMOD 2021.

Caching

Competitive Caching with Machine Learned Advice.
Thodoris Lykouris, Sergei Vassilvitskii.
ICML 2018.

Learning Caching Policies with Subsampling.
Haonan Wang, Hao He, Mohammad Alizadeh, Hongzi Mao.

NeurIPS 2019.

Leaper: A Learned Prefetcher for Cache Invalidation in LSM-tree based Storage Engines.
Lei Yang, Hong Wu, Tieying Zhang, Xuntao Cheng, Feifei Li, Lei Zou, Yujie Wang, Rongyao Chen, Jianying Wang, Gui Huang.
VLDB 2020.

Nearest Neighbor/Similarity Search

HAP: An Efficient Hamming Space Index Based on Augmented Pigeonhole Principle
Qiyu Liu, Yanyan Shen, Lei Chen
SIMOD 22

Deep Visual-Semantic Quantization for Efficient Image Retrieval
Yue Cao, Mingsheng Long, Jianmin Wang, Shichen Liu
CVPR 2017

A Learning to Tune Framework for LSH
Xiu Tang; Sai Wu; Gang Chen; Jinyang Gao; Wei Cao; Zhifei Pang
ICDE 21

Reformer: The Efficient Transformer
Nikita Kitaev, Łukasz Kaiser, Anselm Levskaya
ICLR 2020

MONGOOSE: A learnable LSH framework for efficient neural network training
B Chen, Z Liu, B Peng, Z Xu, JL Li, T Dao, Z Song, A Shrivastava, C Re
ICLR 2021

Neural Network Memory

Scaling Memory-Augmented Neural Networks with Sparse Reads and Writes
Jack W Rae, Jonathan J Hunt, Tim Harley, Ivo Danihelka, Andrew Senior, Greg Wayne, Alex Graves, Timothy P Lillicrap

End-To-End Memory Networks
Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, Rob Fergus

HiPPO: Recurrent Memory with Optimal Polynomial Projections
Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, Christopher Re

Small ReLU networks are powerful memorizers: a tight analysis of memorization capacity
Chulhee Yun, Suvrit Sra, Ali Jadbabaie

Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks
Peter L. Bartlett, Nick Harvey, Chris Liaw, Abbas Mehrabian

Statement on Academic Conduct and Support Systems

Academic Integrity:
The University of Southern California is a learning community committed to developing successful scholars and researchers dedicated to the pursuit of knowledge and the dissemination of ideas. Academic misconduct, which includes any act of dishonesty in the production or submission of academic work, comprises the integrity of the person who commits the act and can impugn the perceived integrity of the entire university community. It stands in opposition to the university's mission to research, educate, and contribute productively to our community and the world.

All students are expected to submit assignments that represent their own original work, and that have been prepared specifically for the course or section for which they have been submitted. You may not submit work written by others or "recycle" work prepared for other courses without obtaining written permission from the instructor(s).

Other violations of academic integrity include, but are not limited to, cheating, plagiarism, fabrication (e.g., falsifying data), collusion, knowingly assisting others in acts of academic dishonesty, and any act that gains or is intended to gain an unfair academic advantage.

The impact of academic dishonesty is far-reaching and is considered a serious offense against the university. All incidences of academic misconduct will be reported to the Office of Academic Integrity and could result in outcomes such as failure on the assignment, failure in the course, suspension, or even expulsion from the university.

For more information about academic integrity see the student handbook or the Office of Academic Integrity's website, and university policies on Research and Scholarship Misconduct.

Please ask your instructor if you are unsure what constitutes unauthorized assistance on an exam or assignment, or what information requires citation and/or attribution.

Course Content Distribution and Synchronous Session Recordings Policies
USC has policies that prohibit recording and distribution of any synchronous and asynchronous course content outside of the learning environment.

Recording a university class without the express permission of the instructor and announcement to the class, or unless conducted pursuant to an Office of Student Accessibility Services (OSAS) accommodation. Recording can inhibit free discussion in the future, and thus infringe on the academic freedom of other students as well as the instructor. (Living our Unifying Values: The USC Student Handbook, page 13).

Distribution or use of notes, recordings, exams, or other intellectual property, based on university classes or lectures without the express permission of the instructor for purposes other than individual or group study. This includes but is not limited to providing materials for distribution by services publishing course materials. This restriction on unauthorized use also applies to all information, which had been distributed to students or in any way had been displayed for use in relationship to the class, whether obtained in class, via email, on the internet, or via any other media. (Living our Unifying Values: The USC Student Handbook, page 13).

Students and Disability Accommodations:
USC welcomes students with disabilities into all of the University's educational programs. The Office of Student Accessibility Services (OSAS) is responsible for the determination of appropriate accommodations for students who encounter disability-related barriers. Once a student has

completed the OSAS process (registration, initial appointment, and submitted documentation) and accommodations are determined to be reasonable and appropriate, a Letter of Accommodation (LOA) will be available to generate for each course. The LOA must be given to each course instructor by the student and followed up with a discussion. This should be done as early in the semester as possible as accommodations are not retroactive. More information can be found at osas.usc.edu. You may contact OSAS at (213) 740-0776 or via email at osasfrontdesk@usc.edu.

Support Systems:

Counseling and Mental Health - (213) 740-9355 – 24/7 on call
Free and confidential mental health treatment for students, including short-term psychotherapy, group counseling, stress fitness workshops, and crisis intervention.

988 Suicide and Crisis Lifeline - 988 for both calls and text messages – 24/7 on call
The 988 Suicide and Crisis Lifeline (formerly known as the National Suicide Prevention Lifeline) provides free and confidential emotional support to people in suicidal crisis or emotional distress 24 hours a day, 7 days a week, across the United States. The Lifeline is comprised of a national network of over 200 local crisis centers, combining custom local care and resources with national standards and best practices. The new, shorter phone number makes it easier for people to remember and access mental health crisis services (though the previous 1 (800) 273-8255 number will continue to function indefinitely) and represents a continued commitment to those in crisis.

Relationship and Sexual Violence Prevention Services (RSVP) - (213) 740-9355(WELL) – 24/7 on call
Free and confidential therapy services, workshops, and training for situations related to gender- and power-based harm (including sexual assault, intimate partner violence, and stalking).

Office for Equity, Equal Opportunity, and Title IX (EEO-TIX) - (213) 740-5086
Information about how to get help or help someone affected by harassment or discrimination, rights of protected classes, reporting options, and additional resources for students, faculty, staff, visitors, and applicants.

Reporting Incidents of Bias or Harassment - (213) 740-5086 or (213) 821-8298
Avenue to report incidents of bias, hate crimes, and microaggressions to the Office for Equity, Equal Opportunity, and Title for appropriate investigation, supportive measures, and response.

The Office of Student Accessibility Services (OSAS) - (213) 740-0776
OSAS ensures equal access for students with disabilities through providing academic accommodations and auxiliary aids in accordance with federal laws and university policy.

USC Campus Support and Intervention - (213) 740-0411
Assists students and families in resolving complex personal, financial, and academic issues adversely affecting their success as a student.

Diversity, Equity and Inclusion - (213) 740-2101
Information on events, programs and training, the Provost's Diversity and Inclusion Council, Diversity Liaisons for each academic school, chronology, participation, and various resources for students.

USC Emergency - UPC: (213) 740-4321, HSC: (323) 442-1000 – 24/7 on call

Emergency assistance and avenue to report a crime. Latest updates regarding safety, including ways in which instruction will be continued if an officially declared emergency makes travel to campus infeasible.

USC Department of Public Safety - *UPC: (213) 740-6000, HSC: (323) 442-1200 – 24/7 on call*
Non-emergency assistance or information.

Office of the Ombuds - *(213) 821-9556 (UPC) / (323-442-0382 (HSC)*
A safe and confidential place to share your USC-related issues with a University Ombuds who will work with you to explore options or paths to manage your concern.

Occupational Therapy Faculty Practice - *(323) 442-2850 or* otfp@med.usc.edu
Confidential Lifestyle Redesign services for USC students to support health promoting habits and routines that enhance quality of life and academic performance.