CSC 696J: Advanced Topics in Data Systems

Gould-Simpson, Rm 701 TuTh 8:00AM - 9:15AM

Course Description

The goal of this graduate seminar course is to learn more about research in the general field of data systems. In this course, we will read and review research papers on data systems. We will also learn how to do research in computer science by reading, evaluating, presenting, and conducting a research project in data systems. Topics include big data systems, cloud databases, AI systems, natural language-based querying systems, machine learning for systems, high dimensional data management, data preparation, and serverless computing, etc.

Instructor and Contact Information

Lei Cao
712 Gould-Simpson
caolei@arizona.edu
520-621-4632

Office hours: Wednesday 2:00PM – 3:00PM at my office or by email appointment
Instructor home page: https://www.cs.arizona.edu/~caolei/
Piazza link: https://piazza.com/arizona/fall2025/csc696j2025fall, access code: csc696j
D2L: https://d2l.arizona.edu/d2l/home/1655929
Grade Scope: https://www.gradescope.com/courses/1087383 entry code: EED2WJ
Course home page: https://www.cs.arizona.edu/~caolei/teach.html

Course Format and Teaching Methods

Classes will consist of lectures and discussions based on readings from the data systems literature. There will be a semester long project, 12 assignments, and 7 in class quizzes.

Course Objectives

The goal of this graduate seminar course is to learn more about data systems research, specific to the domain chosen by the instructor and students.

Students will learn the state-of-the-art in research within the chosen domain and with the following objectives:

- Learn to read and review research papers in data systems or related domains.
- Learn how to synthesize papers that attempt to solve similar research problems, make comparisons between such papers, and identify remaining limitations that could identify to future research.
- Understand, implement concepts of domain-specific research by conducting a semester-long research project.
- Develop computer science research skills that include effectively communicating research verbally and in writing through reports and presentations.

Expected Learning Outcomes

● Learn to identify and summarize the research problem, context, proposed approach, evaluation, and limitations in research papers.

● Learn to organize and plan the execution of a research project.

● Learn to effectively communicate the problem being solved, context, proposed approach, evaluation, and limitations of their own research project.

● Learn to provide constructive suggestions to others in the course about their research projects.

For a more granular description of the learning objectives, see the week-by-week schedule and the description of the assignments below.

Data Systems is a big field, and there is no way we can cover all of it in one course. With that said, this course covers a large amount of material, and the assignments are a central part of the course. **Students are expected to dedicate a significant amount of time on the course outside of the classroom, especially if they have background deficiencies to make up**.

Transferable Career Skills

National Association of Colleges and Employers (NACE) Career Readiness:

Career readiness is a foundation from which to demonstrate requisite core competencies that broadly prepare the college-educated for success in the workplace and lifelong career management. For new college graduates, career readiness is key to ensuring successful entrance into the workforce.

There are eight career readiness competencies, each of which can be demonstrated in a variety of ways." (NACE, 2025)

● Career & Self Development

● Communication

● Critical Thinking

● Equity & Inclusion

● Leadership

● Professionalism

● Teamwork

● Technology

In this course, we will focus on the following competencies:

● Technology: Students will learn the cutting-edge techniques to design and develop distributed database systems, AI systems, and other data intensive systems.

● Teamwork: The course will include lectures, in-class discussions, and activities, with a strong emphasis on teamwork through group work that requires collaboration with other students both during and outside of class.

● Communication: Students are encouraged to emphasize communication by interacting with teaching assistants and using various channels, such as D2L, email, class discussions, and Piazza, to stay updated on course materials.

Makeup Policy for Students Who Register Late

Students who register after the first class meeting may make up missed assignments/projects at a deadline set in consultation with the instructor.

Course Communications

We will use official UA email and Piazza as the primary mode of contact. D2L will be used to provide grading and feedback.

Required Texts or Readings

The primary texts for this course will be research papers and related materials, distributed by the instructor throughout the semester.

Assignments and Examinations: Schedule/Due Dates
12 assignments and 7 in class quizzes. See scheduled topics and activities for due dates.

Final Examination
No final Exam.

Grading Scale and Policies
Grades are assigned based on assignments, final project, and class participation. The grading breakdown is as follows:

- Class Participation / Group discussion of related readings (10%)
- Quizzes throughout the semester (20)
- Written reviews and evaluation of research papers (20%)
- Presentations throughout the semester of research topics (10%)
- Final project (40%)

The final grade in the course is determined by the better of a per-class grading curve and overall performance:
- 90% or better: A;
- 80% or better: B;
- 70% or better: C;
- 60% or better: D;
- below 60%: E.

University policy regarding grades and grading systems is available at
https://catalog.arizona.edu/policy/courses-credit/grading/grading-system

Incomplete (I) or Withdrawal (W):
Requests for incomplete (I) or withdrawal (W) must be made in accordance with University policies, which are available at https://catalog.arizona.edu/policy/courses-credit/grading/grading-system.

Dispute of Grade Policy
If you wish to dispute your grade for an assignment, you have two weeks after the grade has been turned in. In addition, even if you only dispute one portion of the grading for that unit, I reserve the right to revisit the entire unit (assignment or project).

Scheduled Topic and Activities

Week	Date	Description
1	8/26	Lecture: course introduction and logistics Homework1: read and write critiques for papers: J. Dean, and S. Ghemawat, **MapReduce: simplified data processing on large clusters**, Communications of the ACM 51 (1): 107--113 (January 2008) Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, Ankur Dave, Justin Ma, Murphy McCauley, Michael J. Franklin, Scott Shenker, and Ion Stoica, **Resilient distributed datasets: a fault-tolerant abstraction for in-memory cluster computing**. In Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation (NSDI'12).
1	8/28	Lecture: MapReduce
2	9/2	**In class Quiz 1;** Lecture: Spark Homework 1 due: the critique of MapReduce and Spark (Resilient distributed datasets) papers Homework2: read and write critiques for papers: Zihui Gu, Ju Fan, Nan Tang, Lei Cao, Bowen Jia, Sam Madden, Xiaoyong Du:**Few-shot Text-to-SQL Translation using Structure and Content Prompt Learning.** Proc. ACM Manag. Data 1(2): 147:1-147:28 (2023) Yue Gong, Chuan Lei, Xiao Qin, Kapil Vaidya, Balakrishnan Narayanaswamy, Tim Kraska: **SQLens: An End-to-End Framework for Error Detection and Correction in Text-to-SQL**. CoRR abs/2506.04494 (2025)

2	9/4	Student presentation: Few-shot Text-to-SQL Homework 2 due: the critics of Few-shot Text-to-SQL
3	9/9	Guest Lecture: text2SQL (**Dr. Chuan Lei, AWS Science**) Homework 2 due: the critique of the SQLens paper Homework 3: read and write critiques for papers: Bart Samwel, John Cieslewicz, et al, **F1 query: declarative querying at scale**. Proc. VLDB Endow. 11, 12 (August 2018), 1835–1848. Zhaoze Sun, Deng Qiyan, Chengliang Chai, Kaisen Jin, Xinyu Guo, Han Han, Ye Yuan, Guoren Wang, Lei Cao: **QUEST: Query Optimization in Unstructured Document Analysis**. CoRR abs/2507.06515 (2025) Optional reading: Yifei Yang, Matt Youill, Matthew Woicik, Yizhou Liu, Xiangyao Yu, Marco Serafini, Ashraf Aboulnaga, and Michael Stonebraker, **FlexPushdownDB: hybrid pushdown and caching in a cloud DBMS.** Proc. VLDB Endow. 14, 11 (July 2021), 2101–2113.
3	9/11	**In class Quiz 2** Student presentation: F1 query Project option (1) discussion Homework 3 due: the critics of the F1 query paper
4	9/16	Guest Lecture: Unstructured Data Analysis (**Prof. Chengliang Chai, BIT, China)** Homework 3 due: the critique of the QUEST paper Homework 4: read and write critiques for papers: Bobbi W. Yogatama, Weiwei Gong, Xiangyao Yu: **Orchestrating Data Placement and Query Execution in Heterogeneous CPU-GPU DBMS.** Proc. VLDB Endow. 15(11): 2491-2503 (2022) Ferdi Kossmann, Ziniu Wu, Alex Turk, Nesime Tatbul, Lei Cao, Samuel Madden: **CascadeServe: Unlocking Model Cascades for Inference Serving.** CoRR abs/2406.14424 (2024) **(Tentative)** Optional reading:

		https://arxiv.org/abs/2508.04701 https://dl.acm.org/doi/10.1145/3318464.3380595 https://dl.acm.org/doi/abs/10.1145/3514221.3526132 https://dl.acm.org/doi/10.14778/3704965.3704977
4	9/18	Guest Lecture: GPU DBMS (**Dr. Bobbie Yogatama, NVIDIA**) Homework 4 due: the critique of the CPU-GPU DBMS paper
5	9/23	Guest Lecture: Agentic Workflow Optimization **(Ferdi Kossmman, PhD MIT)** Project option (2) discussion Homework 4 due: the critique of the CascadeServe paper (Tentative) Homework 5: read and write critiques for papers: **Palimpzest: Optimizing AI-Powered Analytics with Declarative** **Query Processing (**https://www.vldb.org/cidrdb/papers/2025/p12-liu.pdf) Tim Kraska, Alex Beutel, Ed H. Chi, Jeffrey Dean, and Neoklis Polyzotis. **The Case for Learned Index Structures**. In Proceedings of the 2018 International Conference on Management of Data (SIGMOD '18). Association for Computing Machinery, New York, NY, USA, 489–504. Optional reading: Vikram Nathan, Jialin Ding, Mohammad Alizadeh, and Tim Kraska. 2020. **Learning Multi-Dimensional Indexes**. In Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data (SIGMOD '20). Association for Computing Machinery, New York, NY, USA, 985–1000. Chunwei Liu, Gerardo Vitagliano, Brandon Rose, Matthew Printz, David Andrew Samson, Michael J. Cafarella: **PalimpChat: Declarative and Interactive AI analytics**. SIGMOD Conference Companion 2025: 183-186
5	9/25	Guest Lecture: Declarative AI Systems **(Prof. Chunwei Liu, Purdue)** Homework 5 due: the critique of the Palimpzest paper
6	9/30	**In class Quiz 3**

		Student presentation: learned indexes Project option (3) discussion Homework 5 due: the critique of the learned indexes paper Homework 6: read and write critiques for papers: Zeyuan Shang, Emanuel Zgraggen, Benedetto Buratti, Ferdinand Kossmann, Philipp Eichmann, Yeounoh Chung, Carsten Binnig, Eli Upfal, and Tim Kraska. **Democratizing Data Science through Interactive Curation of ML Pipelines.** In Proceedings of the 2019 International Conference on Management of Data (SIGMOD '19). Association for Computing Machinery, New York, NY, USA, 1171–1188. Grace Fan, Jin Wang, Yuliang Li, Dan Zhang, Renée J. Miller: **Semantics-aware Dataset Discovery from Data Lakes with Contextualized Column-based Representation Learning.** Proc. VLDB Endow. 16(7): 1726-1739 (2023) https://www.vldb.org/pvldb/vol16/p1726-fan.pdf Optional reading: Yihao Hu, Jin Wang, Sajjadur Rahman: LakeVisage: **Towards Scalable, Flexible and Interactive Visualization Recommendation for Data Discovery over Data Lakes**. CoRR abs/2504.02150 (2025)
6	10/2	Student presentation: the ML pipeline paper Homework 6 due: the critique of the ML pipeline paper
7	10/7	Guest Lecture: Data Discovery (**Prof. Jin Wang, ASU**) Project option (4) discussion Homework 6 due: the critique of the Starmie paper Homework 7: read and write critiques for papers: Guest Lecture's paper (TBD) Optional reading: Nan Tang, Ju Fan, Fangyi Li, Jianhong Tu, Xiaoyong Du, Guoliang Li, Samuel Madden, Mourad Ouzzani, **RPT: Relational Pre-trained Transformer Is Almost All You Need towards Democratizing Data Preparation.** Proc. VLDB Endow. 14(8): 1254-1261 (2021)

7	10/9	Guest Lecture: Multi-modal Data Analytics (**Dr. Gerardo Vitagliano, MIT)** Homework 7 due: the critique of the guest lecture's paper Homework 8: read and write critiques for paper: **AgenticData: An Agentic Data Analytics System for Heterogeneous Data** **LLaVA-PruMerge: Adaptive Token Reduction for Efficient Large Multimodal Models** (ICCV 2025) Optional reading: **GaussDB-Vector: A Large-Scale Persistent Real-Time Vector Database for LLM Applications** **Post-training Quantization on Diffusion Models** (CVPR 2023) **Pb-llm: Partially binarized large language models** (ICLR 2024)
8	10/14	Guest Lecture: Data+AI (**Dr. Ji Sun, Tsinghua University)** Homework 8 due: the critique of the AgenticData paper
8	10/16	Guest Lecture: Efficient Models (**Prof. Yuzhang Shang, UCF)** Homework 8 due: the critique of the LLaVA-PruMerge paper
9	10/21	**In class Quiz 4** Project proposal/initial results presentation Homework 9: read and write critiques for papers: Zongheng Yang, Eric Liang, Amog Kamsetty, Chenggang Wu, Yan Duan, Xi Chen, Pieter Abbeel, Joseph M. Hellerstein, Sanjay Krishnan, Ion Stoica, **Deep Unsupervised Cardinality Estimation**. Proc. VLDB Endow. 13(3): 279-292 (2019) Boyan Li, Jiayi Zhang, Ju Fan, Yanwei Xu, Chong Chen, Nan Tang, Yuyu Luo: **Alpha-SQL: Zero-Shot Text-to-SQL using Monte Carlo Tree Search**. CoRR abs/2502.17248 (2025) Optional reading: Jiayi Wang, Chengliang Chai, Jiabin Liu, Guoliang Li:

		FACE: A Normalizing Flow based Cardinality Estimator. Proc. VLDB Endow. 15(1): 72-84 (2021) Xiaoying Wang, Changbo Qu, Weiyuan Wu, Jiannan Wang, Qingqing Zhou: **Are We Ready For Learned Cardinality Estimation?** Proc. VLDB Endow. 14(9): 1640-1654 (2021)
9	10/23	Student presentation: the deep unsupervised cardinality estimation paper Homework 9 due: the critique of the deep unsupervised cardinality estimation paper
10	10/28	Guest Lecture: text2SQL and Data Agent (**Prof. Yuyu Luo**, **HKUST**) Homework 9 due: the critique of the Alpha-SQL paper Homework 10: read and write critiques for papers: Tianyu Li, Badrish Chandramouli, Sebastian Burckhardt, Samuel Madden: **DARQ Matter Binds Everything: Performant and Composable Cloud Programming via Resilient Steps**. Proc. ACM Manag. Data 1(2): 117:1-117:27 (2023) TBD Optional reading: Chuangxian Wei, Bin Wu, Sheng Wang, Renjie Lou, Chaoqun Zhan, Feifei Li, and Yuanzhe Cai. **AnalyticDB-V: a hybrid analytical engine towards query fusion for structured and unstructured data**. Proc. VLDB Endow. 13, 12 (August 2020), 3152–3165. https://doi.org/10.14778/3415478.3415541
10	10/30	Guest Lecture: LLM + DB **(Prof. Immanuel Trummer)**
11	11/04	**In class Quiz 5** Early project status presentation
11	11/06	Guest Lecture: Cloud Programming (**Dr. Tianyu Li, Incoming Assistant Professor, WISC**) Homework 10 due: the critics of the DARQ paper
12	11/11	**Veterans Day**
12	11/13	Guest Lecture: (**Prof. Roee Shraga, WPI**) the TBD paper Homework 10 due: the critique of the TBD paper Homework 11: read and write critiques for papers:

		Lampros Flokas, Weiyuan Wu, Yejia Liu, Jiannan Wang, Nakul Verma, Eugene Wu: **Complaint-Driven Training Data Debugging at Interactive Speeds**. SIGMOD Conference 2022: 369-383 Huayi Zhang, Binwei Yan, Lei Cao, Samuel Madden, Elke A. Rundensteiner: **MetaStore: Analyzing Deep Learning Meta-Data at Scale**. Proc. VLDB Endow. 17(6): 1446-1459 (2024) Optional reading: Weiyuan Wu, Lampros Flokas, Eugene Wu, Jiannan Wang: **Complaint-driven Training Data Debugging for Query 2.0**. SIGMOD Conference 2020: 1317-1334
13	11/18	**In class Quiz 6** Student presentation: the training data debugging paper Homework 11 due: the critics of the training data debugging paper
13	11/20	Student presentation: the MetaStore paper Homework 11 due: the critique of the MetaStore paper Homework 12: read and write critiques for papers: Alexander Lee, et al.: **Semantic Integrity Constraints: Declarative Guardrails for AI-Augmented Data Processing Systems** VLDB 2025
14	11/25	Guest Lecture: Semantics Data Processing **(Alexander Lee, Brown University)** Homework 12 due: the critique of the sematic integrity constraints paper
14	11/27	**Thanksgiving**
15	12/2	Project Presentation/Demo
15	12/4	Project Presentation/Demo
16	12/9	Guest Lecture: Vector DB **(Sylvia Zhang, MIT)** **Project Report due: 12/9**

Classroom Behavior Policy

To foster a positive learning environment, students and instructors have a shared responsibility. We want a safe, welcoming, and inclusive environment where all of us feel comfortable with each other and where we can challenge ourselves to succeed. To that end, our focus is on the tasks at hand and not on extraneous activities (e.g., texting, chatting, reading a newspaper, making phone calls, web surfing, etc.).

Students are asked to refrain from disruptive conversations with people sitting around them during lecture. Students observed engaging in disruptive activity will be asked to cease this behavior. Those who continue to disrupt the class will be asked to leave lecture or discussion and may be reported to the Dean of Students.

Some learning styles are best served by using personal electronics, such as laptops and iPads. These devices can be distracting to other learners. Therefore, students who prefer to use electronic devices for note-taking during lecture should use one side of the classroom.

Safety on Campus and in the Classroom

For a list of emergency procedures for all types of incidents, please visit the website of the Critical Incident Response Team (CIRT): https://cirt.arizona.edu/case-emergency/overview

Also watch the video available at
https://arizona.sabacloud.com/Saba/Web_spf/NA7P1PRD161/app/me/ledetail;spf-url=common%2Flearningeventdetail%2Fcrtfy000000000003841

University-wide Policies link

Links to the following UA policies are provided here: https://catalog.arizona.edu/syllabus-policies

- Absence and Class Participation Policies
- Threatening Behavior Policy
- Accessibility and Accommodations Policy
- Code of Academic Integrity
- Nondiscrimination and Anti-Harassment Policy
- Class Recordings
- Additional Resources
- Preferred Names and Pronouns

Department-wide Syllabus Policies and Resources link

Links to the following departmental syllabus policies and resources are provided here, https://www.cs.arizona.edu/cs-course-syllabus-policies :

- Department Code of Conduct
- Illnesses and Emergencies
- Obtaining Help
- Confidentiality of Student Records
- Land Acknowledgement Statement

Subject to Change Statement

Information contained in the course syllabus, other than the grade and absence policy, may be subject to change with advance notice, as deemed appropriate by the instructor.