EPPS6323 Knowledge Mining

Spring 2026 Syllabus

Karl Ho, PhD.

January 27, 2026

1 Course Information

Schedule	Tuesday 4:00 PM - 6:45 PM
Location	FO 1.502
Semester	Spring 2026

1.1 Instructor Information

Instructor	Karl Ho, Ph.D.
Office	GR 3.203
Phone	972-883-2017
Email	kyho@utdallas.edu
Office Hours	Monday & Wednesday, 1:30 PM – 3:30 PM (by appointment)

1.2 Teaching Assistant

TA	Ariel Wang
Office	GR 3.318
Email	Ariel.Wang@UTDallas.edu
Office Hours	Wednesday & Thursday 1:30 PM – 3:30 PM (by appointment)

2 Course Overview

This course explores the interdisciplinary field of **knowledge mining**, integrating theoretical concepts from data mining, statistical learning, and machine learning, with a focus on leveraging AI, large language models (LLMs), and generative AI for knowledge discovery.

The 2026 edition of this course reflects the significant paradigm shift toward **agentic AI systems**—autonomous agents that can reason, plan, use tools, and execute multi-step tasks with minimal human intervention. Students will learn to work with AI as a **research partner**, developing skills in:

- Retrieval-Augmented Generation (RAG) architectures
- Multimodal knowledge extraction and knowledge graphs
- AI-assisted coding for scientific research
- Ethical AI deployment in research contexts

Students will gain hands-on experience with tools and techniques for analyzing and visualizing data, extracting insights, and applying ethical considerations in real-world research scenarios.

> **AI-Native Learning Approach**
>
> This course adopts an "AI as research partner" pedagogy. Students will progressively develop skills in human-AI collaboration, moving from AI-assisted work to AI-delegated research tasks with appropriate oversight.

3 Prerequisites

There is no official prerequisite for this course. The key requirements are:

1. **Willingness to learn** materials that may be outside your current field/discipline. This course includes content from Computer Science, Statistics, and other disciplines.
2. **Passion for learning** new technologies without apprehension, particularly given the rapid evolution of AI tools.
3. **Commitment to apply** what is learned in your own research.
4. **True to yourself** since all things can be exposed in AI age.

4 Learning Objectives

Upon completion of this course, students will be able to:

1. **Understand** the foundational principles of knowledge mining and its role in contemporary research, including the emerging AI for Science (AI4S) paradigm.
2. **Apply** advanced techniques in data mining, machine learning, and AI, including reasoning models, RAG architectures, and agentic AI workflows.

3. **Develop** practical skills in AI-assisted tools for data analysis, text mining, and visualization, with emphasis on maintaining scientific rigor.
4. **Design and implement** RAG pipelines and knowledge graph systems for research applications.
5. **Evaluate** AI-generated outputs critically, understanding limitations and potential for hallucination.
6. **Investigate** ethical implications of knowledge mining and AI applications in research.
7. **Collaborate effectively** with AI systems as research partners across scaffolded interaction levels.

5 Required Texts and Readings

5.1 Primary Texts

Text	Notes
Bowen, J. A., & Watson, C. E. (2024). *Teaching with AI: A Practical Guide to a New Era of Human Learning*. Johns Hopkins University Press.	Available electronically at McDermott Library
Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, 16(3), 199-231.	Available electronically
Daoud, A. & Dubhashi, D. (2020). Statistical modeling: The three cultures. arXiv:2012.04570	Link

5.2 New Required Readings (2026)

Text	Topic	Notes
Poldrack, R. A. et al. (2025). Ten Simple Rules for AI-Assisted Coding in Science. arXiv:2510.22254	AI-Assisted Coding	Link

Text	Topic	Notes
Kamalov, F. et al. (2025). Evolution of AI in Education: Agentic Workflows. arXiv:2504.20082	Agentic AI	Link
Nature (2025). AI for Science 2025 Collection.	AI4S Paradigm	Available at UTD Library

5.3 Recommended Reading

5.3.1 Data Mining

- Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3rd ed.). O'Reilly. (HKP)

5.3.2 Statistical Learning and Machine Learning

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *Introduction to Statistical Learning with Applications in R* (2nd ed.). Springer. (ISLR) Website
- Kuhn, M. & Silge, J. (2021). *Tidy Modeling with R*. Online (KS)

5.3.3 Text Mining and NLP

- Silge, J. & Robinson, D. (2022). *Text Mining with R: A Tidy Approach*. Online (SR)

5.3.4 AI and Data Science

- Ho, K. (2024). *Data Programming with GenAI*. Online
- Molnar, C. (2019). *Interpretable Machine Learning*. Online

5.3.5 For Social Scientists

- Grimmer, J., Roberts, M. E., & Stewart, B. M. (2021). Machine Learning for Social Science: An Agnostic Approach. *Annual Review of Political Science*, 24(1), 395–419.
- Attewell, P. A. & Monaghan, D. B. (2015). *Data Mining for the Social Sciences: An Introduction*. UC Press.

5.3.6 For Economists

- Athey, S. (2018). The impact of machine learning on economics. In *The Economics of Artificial Intelligence* (pp. 507-547). University of Chicago Press.

- Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2), 3-28.

6 Course Structure

6.1 Thematic Organization

The course is organized around four major themes:

Phase	Weeks	Focus Area
I. Foundations	1–4	Data science principles, AI paradigms, two/three cultures
II. Methods	5–8	ML techniques, NLP, LLMs, RAG architectures
III. AI Integration	9–13	AI-assisted coding, synthetic data, evaluation metrics
IV. Applications & Ethics	14–16	Ethics, future directions, project presentations

6.2 AI Interaction Scaffolding

Students will progress through three levels of AI collaboration:

Phase	Weeks	AI Role	Student Responsibility
Foundation	1–5	AI as Assistant	Human initiates; AI responds to specific queries
Collaboration	6–10	AI as Collaborator	Iterative dialogue; shared agency in problem-solving
Delegation	11–16	AI as Delegatee	Human sets goals; AI executes with oversight

7 Grading and Requirements

7.1 Grade Structure

Component	Weight	Description
Participation & AI Collaboration Log	10%	Active participation + documented AI interactions

Component	Weight	Description
Assignments (posted on website)	10%	Applied exercises demonstrating techniques
Project Proposal	20%	Original research design with AI integration plan
Progress Reports (2)	30%	Iterative development with documented AI collaboration
Final Project & Presentation	30%	Original research with methodology documentation

7.2 Personal Website Requirement

Each student must create and maintain a website hosted on **GitHub**. All assignments and class materials should be posted on your website before the next class.

> Recommended Method
>
> Use **Quarto** for building your GitHub website.

7.3 Project Requirements

The final project must be **20–25 pages** in length. Requirements include:

- **Originality**: Topic, data, or method(s) must be originally designed and unpublished
- **Consultation**: Proposal must be discussed with instructor and approved
- **Replication**: Acceptable with strong justification (e.g., dissertation research)
- **AI Documentation**: Full documentation of AI tool usage required

> Important
>
> No Kaggle project/data replications. Period. Projects must demonstrate original contribution. Python can be used but only to supplemnt R-based analysis. If you insist to use Python, do extra steps to convert to R (e.g., reticulate, rpy2, etc.) and document the process. Notify the instructor.

7.4 Assignment Guidelines

Exercises apply data programming techniques in:

- Data modeling and visualization
- Exploratory data analysis

- AI-assisted code development and validation

Students are encouraged to develop techniques beyond class materials while maintaining scientific rigor.

8 Tentative Schedule

> **Note**
>
> This schedule is subject to change. Updates will be announced in class and on Teams.

8.1 Part I: Foundations (Weeks 1–4)

Week	Date	Topic	Due
1	Jan 20	Introduction to Knowledge Mining & AI4S Paradigm	
2	Jan 27	Data Science Foundations: Two/Three Cultures	
3	Feb 3	Latest Trends: RAG, Agentic AI, Knowledge Graphs	
4	Feb 10	Machine Learning for Knowledge Mining: Reasoning Models	

8.2 Part II: Methods & Techniques (Weeks 5–8)

Week	Date	Topic	Due
5	Feb 17	Causal and Predictive Modeling	
6	Feb 24	Text Mining, NLP, and Multimodal Knowledge Extraction	
7	Mar 3	LLMs, Reasoning Models, and Agentic AI Systems	**Proposal**
8	Mar 10	AI for Research and Knowledge Discovery: RAG Pipelines	

8.3 Part III: AI Integration (Weeks 9–13)

Week	Date	Topic	Due
9	Mar 17	*Spring Break - No Class*	
10	Mar 24	AI-Assisted Data Collection and Coding	
11	Mar 31	Synthetic Data and Silicon Samples	**Progress 1**
12	Apr 7	Evaluation Metrics, Hallucination Detection, Model Selection	
13	Apr 14	Advanced Applications: Multi-Agent Systems for Research	

8.4 Part IV: Ethics & Applications (Weeks 14–16)

Week	Date	Topic	Due
14	Apr 21	Data Ethics in Knowledge Mining	
15	Apr 28	Future Directions in Knowledge Mining	**Progress 2**
16	May 5	Final Project Presentations	**Final**

9 Course Policies

9.1 Participation

Full attendance is **required and imperative**. Participation includes:

- Full preparation for class including completion of assigned readings
- Active involvement in class discussion
- Documented AI collaboration (AI Collaboration Log)

Only medical emergencies (with documentation) constitute excused absences. Contact instructor in first week for special medical accommodations.

9.2 Generative AI Policy

AI Usage Policy

Generative AI is permitted in specific contexts and **requires acknowledgment**.

When using AI tools on assignments, include an appendix with:

a. **Full exchange**: The entire conversation, highlighting relevant sections
b. **Tool specification**: Which AI tools were used (e.g., "Claude 3.5 Sonnet via API")
c. **Usage explanation**: How AI was used (idea generation, code drafting, text refinement, etc.)
d. **Rationale**: Why AI tools were used (efficiency, overcoming blocks, experimentation, etc.)

9.2.1 AI Collaboration Log

In addition to assignment appendices, maintain a running **AI Collaboration Log** documenting:

- Patterns of effective AI interaction
- Failed approaches and lessons learned
- Evolution of your AI collaboration skills

Warning

AI tools should be used **wisely and reflectively** to deepen understanding, not to bypass learning. The goal is developing adaptive collaboration skills for an AI-augmented research future.

9.3 Document Guidelines

All documents must adhere to:

- **Format**: PDF only (no Word or Google Docs)
- **Naming**: `lastname_documenttype.pdf` (e.g., `smith_proposal.pdf`)
- **Paper**: US Letter size
- **Margins**: 1 inch on all sides
- **Font**: 12-point
- **Style**: APSA format (Style Manual)
- **Submission**: Email attachment (no cloud drive links)

9.4 Class Q&A

Piazza is used for class discussion. Before emailing, check/post on Piazza.

- Enroll: https://piazza.com/utdallas/spring2026/epps6323

- Access code: **1502** (classroom number)

9.5 Software and Tools

9.5.1 Required Tools

- R and RStudio
- Python 3.x
- Git and GitHub account
- Quarto

9.5.2 AI Tools

Students should have access to at least one of the following:

- Claude (Anthropic)
- ChatGPT (OpenAI)
- GitHub Copilot (recommended for coding)

9.5.3 Style Guide

Follow the Tidyverse Style Guide for R code.

10 University Policies

10.1 Comet Creed

> *"As a Comet, I pledge honesty, integrity, and service in all that I do."*

This creed was voted on by the UT Dallas student body in 2014.

10.2 Academic Support Resources

See http://go.utdallas.edu/academic-support-resources for university academic support resources.

10.3 UT Dallas Syllabus Policies and Procedures

The information at http://go.utdallas.edu/syllabus-policies constitutes the University's policies and procedures segment of this syllabus.

Please review catalog sections regarding credit/no credit grading and withdrawal policies.

> **Disclaimer**
>
> The descriptions and timelines in this syllabus are subject to change at the discretion of the instructor. Changes will be communicated via class announcements and Piazza.

11 Appendix: Key Concepts in 2026 Knowledge Mining

11.1 Agentic AI

AI systems that can reason, plan, use tools, and execute multi-step tasks autonomously. Key patterns include:

- **Tool Use**: AI accessing external tools and APIs
- **Reflection**: AI evaluating and improving its own outputs
- **ReAct**: Reasoning and Acting in interleaved steps
- **Planning**: Breaking complex tasks into sub-goals
- **Multi-Agent Collaboration**: Multiple AI agents working together

11.2 Retrieval-Augmented Generation (RAG)

Architecture that enhances LLM outputs by retrieving relevant information from external knowledge bases. Components:

- **Indexing**: Processing and storing documents as embeddings
- **Retrieval**: Finding relevant documents for a query
- **Generation**: Producing outputs grounded in retrieved context

11.3 Knowledge Graphs

Structured representations of knowledge using nodes (entities) and edges (relationships). In 2026, increasingly integrated with:

- Multimodal data (text, images, video)
- Vector databases for semantic search
- LLMs for knowledge extraction and reasoning

11.4 AI for Science (AI4S)

The convergence of AI innovation with scientific research, representing a transformative research paradigm that integrates:

- Experimental science (empirical induction)
- Theoretical science (theoretical modeling)

- Computational science (simulation)
- Data-intensive science (pattern recognition)
- AI-driven discovery (autonomous hypothesis generation)