# 2024 CS412 DSO: Introduction to Data Mining

# **CS 412: Introduction to Data Mining Course Introduction**

## Important Links

Coursera: <https://www.coursera.org/learn/cs-412/home/week/1>

Campuswire: <https://campuswire.com/c/G081E50D9/feed>, Code: 0163

Course Material: <https://drive.google.com/drive/folders/1H5sdzu2HbZl_ooTx07IXucstEW76xaRs>

**TA Office Hours**

Time: 9:30-10:30 am CT, Tuesday (Priyanka); Zoom:<https://illinois.zoom.us/j/86595923212?pwd=bUY0OXZHTUlrMzc2SWhsZjliREdJUT09>

Time: 8:30-9:30 pm CT Thursday (Tanay);

Zoom:<https://illinois.zoom.us/j/82807587302?pwd=WnZDZFNSQnN4U3c5cGMvL21VbVprUT09>

**Instructor Office Hour**

Time: 8:30-9:30 pm CT, Friday (Prof. Han); Zoom:<https://illinois.zoom.us/j/2173336903?pwd=ckFxNnFWWW5kSXlqRDlIbndWOUViQT09>

## Course Description

This course is an introductory course on data mining. It introduces the basic concepts, principles, methods, implementation techniques, and applications of data mining, with a focus on three major data mining functions: (1) pattern discovery and (2) cluster analysis, and (3) classification.

In the first part of the course, which focuses on pattern discovery, you will learn why pattern discovery is important, what the major tricks are for efficient pattern mining, and how to apply pattern discovery in some interesting applications. The course provides you the opportunity to learn concepts, principles, and skills to practice and engage in scalable pattern discovery methods on massive data; discuss pattern evaluation measures; study methods for mining diverse kinds of frequent patterns, sequential patterns, and sub-graph patterns; and study constraint-based pattern mining and explore their applications.

In the second part of the course, which focuses on cluster analysis, you will learn concepts and methodologies for cluster analysis, which is also known as clustering, data segmentation, or unsupervised learning. We will introduce the basic concepts of cluster analysis and then study a set of typical clustering methodologies, algorithms, and applications. This includes partitioning methods, such as k-means, hierarchical methods, such as BIRCH, density-based methods, such as DBSCAN, and grid-based methods, such as CLIQUE. We will also discuss methods for clustering validation. The learning will be enhanced by clustering software and programming assignments.

In the third part of the course, which focuses on classification methods, you will learn concepts and methodologies for classification, which is also known as supervised learning. We will introduce the basic concepts of classification and then study a set of typical classification methods, algorithms, and applications. This includes decision tree induction, Bayes classification methods, linear classifier, model evaluation and selection, ensemble methods, and additional concepts on classification such as multi-class classification, semi-supervised classification, active learning, and transfer learning. Moreover, we introduce a few advanced classification methods, including Bayesian belief network, support vector machine (SVM), neural network and deep learning, pattern-based classification, k-nearest neighbors (KNN), and genetic algorithm. The learning will be enhanced by programming assignments.

The technical contents of the course are based on the upcoming textbook Data Mining: Concepts and Techniques (4th ed), as well as the on-campus course CS 412–Introduction to Data Mining, which is offered in the Department of Computer Science at the University of Illinois. Please note several themes covered in the textbook are not covered in this online course, including (1) data preprocessing and preparation, and (2) data warehouse and data cube technology. This is because these themes have been covered or will be covered, with possible in-depth treatment, in several other courses offered in the Data Science Online Master program. Therefore, this course will focus on the in-depth study of the three major data mining functions illustrated above.

## Course Goals and Objectives

Upon successful completion of this course, for pattern discovery, you will be able to:

* Recall important pattern discovery concepts, methods, and applications, in particular, the basic concepts of pattern discovery, such as frequent pattern, closed pattern, max-pattern, and association rules.
* Identify efficient pattern mining methods, such as Apriori, ECLAT, and FPgrowth.
* Compare pattern evaluation issues, especially several popularly used measures, such as lift, chi-square, cosine, Jaccard, and Kulczynski, and their comparative strengths.
* Compare mining diverse patterns, including methods for mining multi-level, multi-dimensional patterns, qualitative patterns, negative correlations, compressed and redundancy-aware top-k patterns, and mining long (colossal) patterns.
* Learn well-known sequential pattern mining methods, including methods for mining sequential patterns, such as GSP, SPADE, PrefixSpan, and CloSpan.
* Learn graph pattern mining, including methods for subgraph pattern mining, such as gSpan, CloseGraph, graph indexing methods, mining top-k large structural patterns in a single large network, and graph mining applications, such as graph indexing and similarity search in graph databases.
* Learn constraint-based pattern mining, including methods for pushing different kinds of constraints, such as data and pattern-based constraints, anti-monotone, monotone, succinct, convertible, and multiple constraints.
* Enjoy various pattern mining applications, such as software bug mining and mining quality phrases.

For cluster analysis, you will be able to:

* Recall basic concepts, methods, and applications of cluster analysis, including the concept of clustering, the requirements and challenges of cluster analysis, a multi-dimensional categorization of cluster analysis, and an overview of typical clustering methodologies.
* Learn multiple distances or similarity measures for cluster analysis, including Euclidean and Minkowski distances; proximity measures for symmetric and asymmetric binary variables; distance measures between categorical attributes, ordinal attributes, and mixed types; proximity measures between two vectors – cosine similarity; and correlation measures between two variables – covariance and correlation coefficient.
* Learn popular distance-based partitioning algorithms for cluster analysis, including K-Means, K-Medians, K-Medoids, and the Kernel K-Means algorithms.
* Learn hierarchical clustering algorithms, including basic agglomerative and divisive clustering algorithms, BIRCH, a micro-clustering-based approach, CURE, which explores well-scattered representative points, CHAMELEON, which explores graph partitioning on the KNN Graph of the data, and a probabilistic hierarchical clustering approach.
* Learn the density-based approach to cluster analysis, which can group dense regions of arbitrary shapes, such as DBScan and OPTICS.
* Learn the grid-based approach, which organizes individual regions of the data space into a grid-like structure, such as STING and CLIQUE.
* Study concepts and methods for clustering evaluation and validation by introducing clustering validation using external measures and internal measures, and the measures for evaluating cluster stability and clustering tendency.

For classification, you will be able to:

* Learn the basic concepts, methods, and applications of classification, including the concept of classification, the basic setting of model learning, and the major challenges of classification analysis.
* Learn the concepts of the decision tree, entropy, and information gain, attribute selection with information gain, and decision tree induction algorithm. We will also discuss how to handle continuous-valued attributes, introduce other measures such as gain ratio, Gini-index, and minimal description length (MDL) principle, and discuss overfitting and tree pruning methods. We also introduce the RainForest algorithm for scalable classification and introduce an interactive perception-based classification method
* Learn the concepts of linear classifier; model evaluation and selection methods; techniques to improve classification accuracy: ensemble methods.
* Learn additional concepts on classification, such as multiclass classification, semi-supervised learning, active learning, Bayesian belief networks, and support vector machines.
* Learn concepts and methods for neural networks and deep learning, pattern-based classification, several other classification methods such as K-nearest neighbor and genetic algorithms.

## Textbook and Readings

The lectures are designed to be self-contained, with a reference to the just published 4th edition of the textbook: *Data mining: Concepts and techniques*(4th ed.). Waltham: Morgan Kaufmann.

Chapters of the 4th edition textbook are available under the Course Resources tab of this Coursera course. Note that these are all the chapters related to the topics covered in this course, so the free PDF version of the chapters is sufficient for this course. Please do not distribute the book chapters made available for this course.

## Course Forum

Join discussions and ask questions on the Campuswire forum. The link to join is <https://campuswire.com/c/G081E50D9/feed>**,** the access code is **0163**.

## Course Files

The course slides, book reference and syllabus can be found in the google drive: <https://drive.google.com/drive/folders/1H5sdzu2HbZl_ooTx07IXucstEW76xaRs>

## Course Outline

This 4-credit hour course is 16 weeks long. You should invest **6–8** hours every week in this course.

The course is composed of three parts. Part 1 of the course, Week 1 to Week 5, focuses on pattern discovery. Part 2 of the course, Week 6 to Week 11, focuses on cluster analysis. Part 3 of the course, Week 12 to Week 16, focuses on classification. All of the course content will be released on the first day of class, with the exception of the 3 proctored exams, which will not be released until the day of each exam (for more information on the proctored exams, read the section Elements of This Course below). Although all content (except for exams) is made available to the entire class on the first day, the course follows a schedule (see the table below).

| **Week** | **Duration** | **Part 1: Pattern Discovery** | **Slides** |
| --- | --- | --- | --- |
| 1 | 1/16 - 1/21 | Course Orientation; Course Part 1 Pattern Discovery Overview; Pattern Discovery Basic Concepts; Efficient Pattern Mining Methods; Pattern Discovery Programming Assignment 1 | [link](https://docs.google.com/presentation/d/1sieaG7Yrwit8qt_REbQOgIHkdJwb29EK), [link](https://docs.google.com/presentation/d/1H0FN2HjSGKejVPPxRIj8t6UxmAxOwXEw/) |
| 2 | 1/22 - 1/28 | Pattern Evaluation; Mining Diverse Frequent Patterns | [link](https://docs.google.com/presentation/d/1nxAsOnf9YyzhJrHgeMzI75PFL9EVt9Es/edit?usp=drive_link&ouid=102272340632634607571&rtpof=true&sd=true), [link](https://docs.google.com/presentation/d/1i67TQwoz5sr09-qBVCBr57nWxVgdKPIF/edit?usp=drive_link&ouid=102272340632634607571&rtpof=true&sd=true) |
| 3 | 1/29 - 2/4 | Sequential Pattern Mining; Graph Pattern Mining; Pattern Mining Applications: Software Bug Mining | [link](https://docs.google.com/presentation/d/111gWrBajDN8D8OehsWtKvgTWTL3ZSQFz/edit?usp=drive_link&ouid=102272340632634607571&rtpof=true&sd=true), [link](https://docs.google.com/presentation/d/1X7MLR8N42gxxLtlO3X8ONnp7rBwKD7ew/edit?usp=drive_link&ouid=102272340632634607571&rtpof=true&sd=true) |
| 4 | 2/5 - 2/11 | Constraint-Based Mining; Pattern Mining Applications: Phrase Mining; Pattern Discovery Programming Assignment 2 | [link](https://docs.google.com/presentation/d/1Dd9Je2GeiamUlCAtRn7YX5O-GF5cpxk-/edit?usp=drive_link&ouid=102272340632634607571&rtpof=true&sd=true), [link](https://docs.google.com/presentation/d/1TiLvzw1YR48U1EGk6jJcKmKrd1GwknR9/edit?usp=drive_link&ouid=102272340632634607571&rtpof=true&sd=true), [link](https://docs.google.com/presentation/d/1Fj0eXZ4lYvbJUQQKqv1FCAQFRPP7mSEv/edit?usp=drive_link&ouid=102272340632634607571&rtpof=true&sd=true) |
| 5 | 2/12 - 2/18 | Part 1 Practice Exam; Part 1 Exam on Pattern Discovery |  |
| **Week** | **Duration** | **Part 2: Cluster Analysis** | **Slides** |
| 6 | 2/19 - 2/25 | Course Part 2 Cluster Analysis Overview; Cluster Analysis Introduction; Similarity Measures for Cluster Analysis | [link](https://docs.google.com/presentation/d/1tNY9A6G-SXFrnCZIYT8Oqqxc9akj8WCm/edit?usp=drive_link&ouid=102272340632634607571&rtpof=true&sd=true), [link](https://docs.google.com/presentation/d/1gCRKHBs3rmnBRHHHO36z-7gs_-7YT3-F/edit?usp=drive_link&ouid=102272340632634607571&rtpof=true&sd=true), [link](https://docs.google.com/presentation/d/1CFvzehtXUWY5yutYVPLToPkg0ivwY_f5/edit?usp=drive_link&ouid=102272340632634607571&rtpof=true&sd=true) |
| 7 | 2/26 - 3/3 | Partitioning-Based Clustering Methods; Hierarchical Clustering Methods | [link](https://docs.google.com/presentation/d/1caBcthzrhI99CnBac3DvJCyjIiTdpeIz/edit?usp=drive_link&ouid=102272340632634607571&rtpof=true&sd=true), [link](https://docs.google.com/presentation/d/1G2Xv17gcj6yImhKIDvCbdiQB2zP8zZ0V/edit?usp=drive_link&ouid=102272340632634607571&rtpof=true&sd=true) |
| 8 | 3/4 - 3/10 | Hierarchical Clustering Methods (continued); Density-Based and Grid-Based Clustering Methods; Cluster Analysis Programming Assignment 1 | [link](https://docs.google.com/presentation/d/1JeeSSof1nTqdl8OhBVAHuWwu-YIOG4q6/edit?usp=drive_link&ouid=102272340632634607571&rtpof=true&sd=true), [link](https://docs.google.com/presentation/d/129OBmMA45jYW-xdiIN1FPj3bJ9TXLxJz/edit?usp=drive_link&ouid=102272340632634607571&rtpof=true&sd=true) |
| 9 | 3/11 - 3/17 | Spring Break |  |
| 10 | 3/18 - 3/24 | Methods for Clustering Validation; Cluster Analysis Programming Assignment 2 | [link](https://docs.google.com/presentation/d/14jd9x7TMrOFEXcGSV7OSwFoUQk4IaEw5/edit?usp=drive_link&ouid=102272340632634607571&rtpof=true&sd=true), [link](https://docs.google.com/presentation/d/1xxVVEoH7-r9tAGBYO1_svJliYHWsYl1g/edit?usp=drive_link&ouid=102272340632634607571&rtpof=true&sd=true) |
| 11 | 3/25 - 3/31 | Part 2 Practice Exam; Part 2 Exam on Cluster Analysis |  |
| **Week** | **Duration** | **Part 3: Classification** | **Slides** |
| 12 | 4/1 – 4/7 | Course Part 3 Classification Overview; Lesson 0: Classification in Data Mining: An Introduction; Lesson 1: Decision Tree Induction; Lesson 2-1: Bayes Classifier | [link](https://docs.google.com/presentation/d/1si_7HsefnAulMIVHcPr_Ne4Qbc1Vrdfe/edit?usp=drive_link&ouid=102272340632634607571&rtpof=true&sd=true), [link](https://docs.google.com/presentation/d/1odeB5vKy-Pyv50c1ODy3nSsvmy6VsILE/edit?usp=drive_link&ouid=102272340632634607571&rtpof=true&sd=true), [link](https://docs.google.com/presentation/d/1bH9FVSayWuuaqeIE5H9dVwtTI_2E1A23/edit?usp=drive_link&ouid=102272340632634607571&rtpof=true&sd=true) |
| 13 | 4/8 –4/14 | Lesson 2-2: Bayesian Networks; Lesson 3: Model Evaluation, Selection and Improvements | [link](https://docs.google.com/presentation/d/1tnaKZV-KqeoYJlEgHeTu-hfierAlQ-Un/edit?usp=drive_link&ouid=102272340632634607571&rtpof=true&sd=true), [link](https://docs.google.com/presentation/d/1UgohadMBCmw6n-i_iOBXt4aP1akv3CAh/edit?usp=drive_link&ouid=102272340632634607571&rtpof=true&sd=true) |
| 14 | 4/15 –4/21 | Lesson 4: Linear Classifier and Support Vector Machines; Lesson 5-1: Neural Networks | [link](https://docs.google.com/presentation/d/1rjvhpGSxOyqubwEdP1TDqYg0-WVKiJkJ/edit?usp=drive_link&ouid=102272340632634607571&rtpof=true&sd=true), [link](https://docs.google.com/presentation/d/12DnEqQcDDDUDcdr4T9UYxci6qKdaukbE/edit?usp=drive_link&ouid=102272340632634607571&rtpof=true&sd=true) |
| 15 | 4/22 – 4/28 | Lesson 5-2: Deep Learning; Lesson 6: Pattern-based Classification and K-Nearest Neighbors Algorithm | [link](https://docs.google.com/presentation/d/1Rb4OvBeCQ8OjzqqZNmdLddQdbOvxiA4y/edit?usp=drive_link&ouid=102272340632634607571&rtpof=true&sd=true) |
| 16 | 4/29 –5/5 | Part 3 Practice Exam; Part 3 Exam on Classification |  |

### MOOC Version and CS 412 Content Mapping

If you have taken the MOOC version of the course, namely Pattern Discovery and Cluster Analysis, below is how the content in those two MOOCs maps to this course.

| **MOOC Equivalent** | **CS 412** |
| --- | --- |
| Pattern Discovery MOOC | Week 1–5 |
| Cluster Analysis MOOC | Week 6-11 |
| *No MOOC equivalent (Classification)* | Week 11-16 |

## Assignment Deadlines

For all assignment deadlines, please refer to the **Course Assignment Deadlines, Late Policy, and Academic Calendar** page.

## Elements of This Course

The course is comprised of the following elements:

* **Lecture Videos.** In each week, the concepts you need to know will be presented through a collection of short video lectures. You may stream these videos for playback within the browser by clicking on their titles or download the videos. You may also download the slides that go along with the videos.
* **In-Video Questions**. Some lecture videos have questions associated with them to help verify your understanding of the topics. These questions will automatically appear while watching the video if you stream the video through your browser. These questions do not contribute toward your final score in the class.
* **Lesson Quizzes**. Each week may contain one or multiple lessons. A lesson is a series of videos on a certain topic, which concludes with a lesson quiz. Quiz attempts are unlimited and the deadline for all quizzes is the last day of class. There is no time limit on how long you take to complete each attempt at the quiz. Each attempt may present a different selection of questions to you. Your highest score will be used when calculating your final score in the class.
* **Programming Assignments.** There are 7 total programming assignments in this course – two are designed around the topic of pattern discovery, three on cluster analysis and two on classification. For more information about the programming assignments, please read the instructions on programming assignment in respective weeks.
* **Proctored Exams**. There are 3 proctored exams on each topic. All exams will be proctored via a proctoring service called ProctorU. For more information about ProctorU and the proctor exams, read the **Proctored Exam** page.

## Grading Distribution and Scale

Your final grade will be calculated based on the activities listed in the table below. Your official final course grade will be listed in [Enterprise](https://apps.uillinois.edu/selfservice/). The course grade you see displayed in Coursera may not match your official final course grade.

### Grading Distribution

| **Assignment** | **Occurrence** | **Percentage Weight of Final Grade** |
| --- | --- | --- |
| Lesson Quizzes | 21 | 20% in total |
| Programming Assignments (or MP) | 7 | 7 x 5% per MP = 35% |
| Course Part 1 Exam (Pattern Discovery) | 1 | 15% |
| Course Part 2 Exam (Cluster Analysis) | 1 | 15% |
| Course Part 3 Exam (Classification) | 1 | 15% |
| Grand Total |  | 100% |

Some extra credit opportunities will also be provided during the semester.

### Grading Scale

| **Letter Grade** | **Percent Needed** | **Letter Grade** | **Percent Needed** | **Letter Grade** | **Percent Needed** |
| --- | --- | --- | --- | --- | --- |
| **A+** | **95%** | **B+** | **80%** | **C** | **65%** |
| **A** | **90%** | **B** | **75%** | **D** | **60%** |
| **A-** | **85%** | **B-** | **70%** | **F** | **Below 60%** |

The above grading scale is for reference only. The letter grades will be finalized by the course instructor at the end of the semester. In previous offerings, the final grade usually does not diverge too far from the reference cutoff.

### View Grades

You can view your grade on each assignment by clicking the **Grades** tab on the left menu bar of Coursera.

## Campus Policies

The University of Illinois at Urbana-Champaign is committed to providing a safe and welcoming campus environment free from discrimination based on sex, which includes sexual assault, sexual exploitation, stalking, sexual harassment, dating violence, and domestic violence (collectively referred to as sexual misconduct). See [Campus Policies and Procedures](http://wecare.illinois.edu/policies/campus/) for more information.

## Student Code

A student at the University of Illinois at the Urbana-Champaign campus is a member of a University community of which all members have at least the rights and responsibilities common to all citizens, free from institutional censorship; affiliation with the University as a student does not diminish the rights or responsibilities held by a student or any other community member as a citizen of larger communities of the state, the nation, and the world. See the [University of Illinois Student Code](https://studentcode.illinois.edu/) for more information.

## Academic Integrity

It is the responsibility of each student to refrain from infractions of academic integrity, from conduct that may lead to suspicion of such infractions, and from conduct that aids others in such infractions. Students have been given notice of this Part by virtue of its publication. Regardless of whether a student has actually read this Part, a student is charged with knowledge of it. Ignorance is not a defense. For more information about academic integrity, see [Article 1 – Student Rights and Responsibilities](https://studentcode.illinois.edu/article1/part4/1-401/) or visit the [Students' Quick Reference Guide to Academic Integrity](https://provost.illinois.edu/policies/policies/academic-integrity/students-quick-reference-guide-to-academic-integrity/).

Academic integrity violations may result in dismissal from the University.

## Computer Science Graduate Student Policies Handbook

The CS department also maintains a policies handbook for graduate students. For more information, see the [Graduate Student Handbook](https://cs.illinois.edu/sites/default/files/images/CSGraduateStudentHandbook_web.15-16.pdf). NOTE: This is the version from last year. We will update this link once the new version is finalized.

## Coursera Policy

All Coursera learners are required to follow an [Honor Code](https://learner.coursera.help/hc/en-us/articles/209818863-Coursera-Honor-Code) and a [Code of Conduct](https://learner.coursera.help/hc/en-us/articles/208280036-Coursera-Code-of-Conduct). Please review both of these items before commencing your studies.