**GGS 787: Scientific Data Mining for Geoinformatics**

**Syllabus is not final on the course schedule**

**Instructor:** [**Ruixin Yang**](http://mason.gmu.edu/~ryang)

# Exploratory Hall 2409, Tel: 993-3615, E-mail: ryang@gmu.edu

**Time & Place**: Thursdays 4:30-7:10pm, Exploratory Hall 2310
**Office Hours**: by appointment (Zoom Meeting: <https://gmu.zoom.us/j/4655943637>).

**Textbook**:

# **Required (Primary): Tan**, **Pang-Ning, Michael Steinbach, [**Anuj Karpatne] and **Vipin Kumar**, “Introduction to Data Mining,” 2006 (1st Edition), Addison-Wesley; 2019 (2nd Edition), Pearson. [Online Information](https://www-users.cse.umn.edu/~kumar001/dmbook/index.php)

# Required (*Secondary*): Han, Jiawei, Micheline Kamber, and [Jian Pei], “Data Mining: Concepts and Techniques,” 2006 (2nd Edition); 2011 (3rd Edition) Morgan Kaufmann.

**GMU Catalog Entry:**

[**GGS 787**](http://catalog.gmu.edu/content.php?catoid=15&navoid=1040&filter%5B27%5D=-1&filter%5B29%5D=&filter%5Bcourse_type%5D=-1&filter%5Bkeyword%5D=GGS&filter%5B32%5D=1&filter%5Bcpage%5D=2&filter%5Bitem_type%5D=3&filter%5Bonly_active%5D=1&filter%5B3%5D=1)**- Scientific Data Mining for Geoinformatics (**Credits: 3)

Covers specialized data mining algorithms, geoscience data models, and data information systems. Emphasis on domain-specific data mining algorithms suitable for spatial data and spatio-temporal data with geoscience and geoinformatics applications. Introduces real geoscience data mining applications in detailed applications.

**Prerequisites**: Competency in programming at the level of CSI 601-607 or permission of instructor.

**Goals and Objectives:**

To introduce basic data mining concepts and algorithms, software implementations for data mining, and applications in geoscience and geoinformatics. Both understanding and the implementation of the certain mining methods will be required.

**Learning Outcomes:**

After successful completion of this course,

1. Students will understand basic data mining concepts and major algorithms;
2. Students will be able to articulate and effectively communicate concepts and ideas related to Data Mining to experts, non-experts, and other professionals in a work environment;
3. Students will have the ability to appropriately apply the knowledge acquired in the course for various hypothetical and real-world data mining tasks, thus being able to mine new and useful information;
4. Students will be able to properly interpret data mining results.

**Course Web Site**: Canvas, the University’s enterprise learning management system at [https://canvas.gmu.edu/](https://canvas.gmu.edu/courses/26927) (or <https://lms.gmu.edu/>). You must use the system for accessing course materials/assignments and for the final project submission. Note: this semester is the first semester for the university to use it. I am learning it.

**Email Communication:** *Based on the university policy,*students must use their MasonLive email account to receive important University information, including communications related to this class. I will not respond to messages sent from or send messages to a non-Mason email address. See <http://masonlive.gmu.edu> for more information on Mason Email System. Please use “GGS 787” in your subject line when initiate an email.

**Computing Requirements:** No specific statistical package/tool/programming will be required for general assignments and the final project in this course. The instructor will use Matlab in most cases. Specific assignments may be given associated with specific tool(s).

**References:**

References will be added during the semester. Almost all of the reference materials (or links) will be available through the Mason Blackboard System in the reference folder of this course.

**Grading Policy:**

Homework Assignments: 60%

Final Project 40%

-

Total 100% (**Letter grades based on both absolute and relative numbers)**

**Notes on Assignments:**

* If multiple files are involved, the assignments will be distributed in .7z (zipped). If you need, you can check Mason ITS site at <https://its.gmu.edu/service/software-listing-7-zip/> for installing the software on your computer.
* Assignments should be submitted only through the Assignment submission section of the Canvas system - DO NOT email assignments directly to the instructor.
* It is expected that your submission will be in either PDF or Word format.
* Please make sure you have a backup of all the materials you submit.
* Please make sure to put your name with your assignment, and use your name or other identification information for your file names.
* If more than one file is submitted, you should submit a single **ZIP** file (such as the .7z) containing all the assignment files. In that case, it is strongly suggested that you put all the files into a folder and name the folder with your identity.
* The grace time is the noon of the following day after the due day. Submission after the grace time may result in losing of points, 10% per day for the first two days. No grading for submission later more than 2 days.
* Different weights may be applied to assignments in the final points calculation.

**The followings are university wide required information from Office of the Provost:**

**UNIVERSITY POLICIES**

* **University Catalog:** The University Catalog, <http://catalog.gmu.edu>, is the central resource for university policies affecting student, faculty, and staff conduct in university academic affairs. Other policies are available at <http://universitypolicy.gmu.edu/>. All members of the university community are responsible for knowing and following established policies.
* **Sexual Harassment*:*** As a faculty member and designated “Responsible Employee,” I am required to report all disclosures of sexual assault, interpersonal violence, and stalking to Mason’s [*Title IX Coordinator*](https://diversity.gmu.edu/sexual-misconduct) per [*university policy 1412*](https://universitypolicy.gmu.edu/policies/reporting-of-clery-act-crimes-andor-prohibited-sexual-conduct/). If you wish to speak with someone confidentially, please contact the [*Student Support and Advocacy Center*](http://ssac.gmu.edu/) (703-380-1434) or [*Counseling and Psychological Services*](https://caps.gmu.edu/) (703-993-2380). You may also seek assistance from [*Mason’s Title IX Coordinator*](https://diversity.gmu.edu/sexual-misconduct) (703-993-8730; *titleix@gmu.edu*).
* **Academic Integrity** (from Mason[Stearns Center for Teaching and Learning](https://stearnscenter.gmu.edu/)):Mason is an Honor Code university; please see the [Office for Academic Integrity](https://oai.gmu.edu/full-honor-code-document/) for a full description of the code and the honor committee process. Three fundamental principles to follow at all times are that: (1) all work submitted be your own, as defined by the assignment; (2) when you use the work, the words, or the ideas of others, including fellow students or online sites, you give full credit through accurate citations; and (3) if you are uncertain about the ground rules on a particular assignment or exam, ask for clarification. No grade is important enough to justify academic misconduct.
* **Generative-AI (GenAI) Tools:**Use of GenAI tools will sometimes be in alignment with the learning outcomes for this course. It is expected that the GenAI for this course is very limited. If used, one should follow the fundamental principles of the Honor Code. This includes being honest about the use of these tools for submitted work and including citations when using the work of others, whether individual people or Generative-AI tools. When meeting the outcome requires original human action, creativity or knowledge, AI tool use would not align with the stated course goals.
* **Office of Disability Services:** If you are a student with a disability and you need academic accommodations, please see me and contact the Office of Disability Services (ODS) at 993-2474. All academic accommodations must be arranged through the ODS, <http://ods.gmu.edu>.
* **Diversity and Inclusion**: Mason, an intentionally inclusive community, promotes and maintains an equitable and just work and learning environment. We welcome and value individuals and their differences including race, economic status, gender expression and identity, sex, sexual orientation, ethnicity, national origin, first language, religion, age, and disability.
* **Name and Pronoun Use:** If you wish, please share your name and pronouns with me and how best to address you in class and via email. I use he/him/his for myself and you may address me as “Dr./Prof. Yang."
* **Full Mason** [Common Course Policies](https://stearnscenter.gmu.edu/home/gmu-common-course-policies/)**.**

**OTHER USEFUL CAMPUS RESOURCES:**

* WRITING CENTER: Johnson Center, Room 227E;
Phone: 703-993-1200; Email: wcenter@gmu.edu; <http://writingcenter.gmu.edu>
* UNIVERSITY LIBRARIES “Ask a Librarian.” <http://library.gmu.edu/ask>
* Counseling and Psychological Services (CAPS): (703) 993-2380;
<http://caps.gmu.edu>
* **University Calendar:** Details regarding the current Academic Calendar. [Calendars | Office of the University Registrar | George Mason University (gmu.edu)](https://registrar.gmu.edu/calendars/)

**Tentative Course Schedule (Contents):**

**The course contents and schedule are “under construction.” Therefore, the list below should be considered as a table of course contents instead of schedule. The assignment given and due dates will be adjusted accordingly. All efforts will be made to cover as much topics below as possible.**

**(Last modified on Monday, September 02, 2024)**

1. Introduction (Data Science and Data Mining)
	* Syllabus
	* Introduction to Data Science
	* Introduction to Data Mining
* Reading Assignment: Chapter 1
* HW1 given
1. Data Issues
	* Attribute Types
	* Data Models in Geoscience
	* Data Quality
	* Reading Assignment: Sections: 2.1, 2.2
2. Data Preprocessing
	* Data Preprocessing
	* Measures of Similarity and Dissimilarity
	* Project Topic due (9/12)
	* Reading Assignment: Sections: 2.3, 2.4
3. Association Analysis (Association Rules) Part 1
	* Basic Concepts
	* Compact Presentation
	* Rule Generation
	* Evaluation of Association Patterns
	* Special Topics
		+ Skewed support patterns
		+ Continuous attributes
		+ Sequential patterns
* Reading Assignment: Sections: Chapter 5 except for section 5.6
1. Association Analysis (Association Rules) Part 2
	* Special Topics
		+ Skewed support patterns
		+ Continuous attributes
		+ Sequential patterns
	* Geoscience Data Mining Applications I: Tropical Cyclone Intensity
	* SHIPS Model and Database
	* Intensity Change of Tropical Cyclones
	* Rapid Intensification of Tropical Cyclones
	* Future Research: RI Prediction Based on Data Mining Results
* Reading Assignment: Sections: 6.2, 6.4; Supplementary Materials
1. Cluster Analysis
	* Basic Concepts
	* K-means
	* Agglomerative Hierarchical Clustering
	* Other Clustering Algorithms (ENVI Built-in)
	* Cluster Evaluation
* Reading Assignment: Sections: 8.1-8.3, 8.5
1. Geoscience Data Mining Applications II: Content-Based Geoscience Data Search
	* Concepts of Content-Based Search
* Methods and Applications
* Reading Assignment: Supplementary Materials
1. Advanced Clustering Algorithms and Applications
	* Density-Based Clustering
* Geoscience Data Mining Applications III: Mining Climate Indices (Spatio-Temporal)
* Reading Assignment: Sections: 9.1, 9.3, 9.4
1. Classification: Part 1: Basics
	* Basic Concepts
	* Decision Trees Induction
		+ Hunt’s Algorithm
		+ Optimal split values, Measures of impurity (Gini index, entropy, etc.)
	* Project outline due (10/24)
	* Reading Assignment: Sections 3.1-3.3
2. Classification: Part 2: Issues and Evaluation
	* Issues in classifications
	* Basic Evaluation
	* Reading Assignment: Sections 3.4-3.9
3. Classification: Part 3: Alternative Techniques
* Rule-Based Classifier
* Nearest-Neighbor classifiers
* Bayesian Classifiers
* Reading Assignment: Sections: 4.1-4.5
* Reading Assignment: Sections 5.1-5.3
1. Classification: Part 4: Alternative Techniques
* Support Vector Machine (SVM)
	+ Classification evaluation measures, more
* Reading Assignment: Sections 4.6, 5.5-5.7
1. Classification: Part 5: Applications
* Support Vector Machine (SVM)
	+ Classification evaluation measures, more
* Reading Assignment: Sections 4.6, 5.5-5.7
1. Deep Learning Application
2. Final Project Due (Exam Day, December 12, 2024)