University of Texas at El Paso
Department of Computer Science & Department of Geological Sciences
CS 4390/5390:Web-based Data Integration , GEOL 5315: Selected Topics – Geological Sciences

Logistics

Instructors:
Natalia Villanueva Rosales, e-mail: nvillanuevarosales [at] utep.edu, office: CCSB Room 3.0508, phone: (915) 747-8643.
Hugo Gutierrez Jurado, email: hagutierrez [at] utep.edu, office: GEO Room 227A, phone: (915) 747-5159.

Class time: Tuesdays and Thursdays, 10:30- 11:50am
Location: Classroom Building 402 – Visualization Wall Room
Office hours:
Villanueva Rosales: Tu-Th, 8:30-10:30am and 3:00-4:00pm (Instructor) and by appointment outside this time.
Gutierrez Hurtado: Tu-Th, 8:30-10:30 am and by appointment outside this time.
Email is the preferred way for contacting the instructors.

NOTE: When contacting the instructors by email, please use in the subject the prefix [CS5390].

Course Description

This is an interdisciplinary active-learning course in which advanced concepts and algorithms in data science will be introduced and applied to real-life issues within the earth and environmental sciences domains, a fast-growing area of research and development. The value and need for interdisciplinary work have increased as technology and new and more complex societal and environmental challenges continue to develop. One of the main objectives of this course is to expose the students through the class activities to the most important operating concepts of Computer and Earth Sciences. From this exercise, students will develop an appreciation and awareness of the opportunities created when combining data, theory and methods from both fields, improving and acquiring tools and soft skills highly regarded by industry and employers. Throughout the course, the students will be introduced to the foundations and applications of integrating heterogeneous data from decoupled sources. Data integration approaches covered in the class include high-level data models (e.g., ontologies) with a focus on semantic web technologies and current research trends in data retrieval and processing. Additionally, the course delves into the challenges in scientific discoveries related to data management, with a focus in the area of Earth Sciences. Earth research relies on environmental data collection that is captured in different formats and temporal and spatial resolution, from ground-based sensor data to satellite and other forms of remote sensing data. Hands-on activities and a course-long project will be used to put in practice the different aspects of data science in the context of Earth Science data management including the acquisition, retrieval, processing, analysis, and sharing of data using the FAIR principles – **F**indable, **A**ccessible, **I**nteroperable and **R**eusable resources.

This course is intended for senior and graduate students in science and engineering. CS2302 is a pre-requisite for students that are Computer Science majors. In addition, computer science students are

expected to have some experience in programming and databases. Students from other programs interested (or in need of) data science to publish and integrate resources (e.g. data, information, and methods) from their research are highly encouraged to register in this course and are not expected to have any previous knowledge or experience in programming.

Learning Outcomes

Course Outcomes

Divided into the following three broad levels of Bloom's taxonomy:

Level 1: Knowledge and Comprehension.

Level 1 outcomes are those in which the student has been exposed to the terms and concepts at a basic level and can supply basic definitions. The material has been presented only at a superficial level. Upon successful completion of the course, students will be able to:

1a. Describe and compare data models (e.g., Entity-Relationship model, ontologies, UML model) and how they are currently used for data management.

1b. Identify current challenges in data management, in particular for the integration and exchange of decoupled data sets. The challenges in data management will be motivated by Earth Sciences research challenges.

1c. Identify current trends in data science research, both fundamental and applied in the Earth Sciences domain.

1d. Describe technical solutions to the challenges in data sharing, integration and reuse of data, in particular for the Earth Sciences domain.

1e. Identify the careers/roles associated with data scientists, in particular for scientific research.

1f. Describe the elements of the scientific method.

Level 2: Application and Analysis.

Level 2 outcomes are those in which the student can apply the material in familiar situations, e.g., can work a problem of familiar structure with minor changes in the details. Upon successful completion of the course, students will be able to:

2a. Demonstrate uses of explicitly stored metadata/schema associated with data on the web.

2b. Use current cyberinfrastructure for the sharing, integration or exchange of data with a focus on reproducible workflows in the Earth Sciences domain.

2c. Demonstrate the ability to work in teams.

2d. Demonstrate the ability to traverse the Computer Science and Earth Sciences domains to develop in teams, applications that solve real-world problems, with a focus on addressing Earth Science research challenges.

2e. Demonstrate the ability to write a technical report following the scientific method.

2f. Demonstrate technical communication abilities, including the capacity to synthesize the most important aspects of a work, to give context and highlight the relevance of the work performed and the results obtained, and to point to the next steps or new directions for improving current performances.

2g. Demonstrate the use of expanded vocabulary and the capacity to translate and combine complex concepts from and across the two knowledge domains (I.e. Computer and Earth Sciences).

2h. Demonstrate the ability to independently and as teams perform literature and methodological searches with scientific rigor.

Level 3: Synthesis and Evaluation.
Level 3 outcomes are those in which the student can apply the material in new situations. This is the highest level of mastery. Upon successful completion of the course, students will be able to: 3a. Design a relational database schema from a problem statement to conceptual/logical/physical database design.

3a. Design a high-level data integration model from a problem statement using standard notation and modeling principles.

3b. Design and implement a tool that enables the integration of data from heterogeneous, decoupled data.

3c. Design and implement an interface for a data-integration tool applying best practices for usability, privacy, security, and reproducibility of results.

3d. Assess the functionality of the tool and data integration model developed in light of the original purpose and objectives stated by the Earth Science issue addressed.

Grading

1. Project 30%
   Use cyberinfrastructure and data science methodologies and tools to access, integrate, query, manipulate, analyze and share data in a meaningful way
2. Presentation and assignments 30%
   Once your project has started, each week you will present project's progress reports and topics related to your project, addressing all of the following:
   - Specific objectives pursued for the project for the reporting period;
   - Description of challenges encountered, and solutions devised/proposed to overcome them;
   - Main results and achievements;
   - New or revised objectives for the following week.

   Assignments will be related with your project and will be discussed in class 1-2 weeks before the deadline. You will present papers about data science.
3. Exam(s) and Quiz(zes) 40%
   Written evaluation of your understanding of the topics reviewed.

Resources

Recommended readings and tools
- A list of readings included scientific papers and book chapters will be provided by the instructors.
- Protégé, a free open source ontology editor and knowledge base framework. Depar

Special Accommodations: If you have a disability and need classroom accommodations, please contact the Center for Accommodations and Support Services (CASS) at 747-5148 or by email to cass@utep.edu, or visit their office located in UTEP Union East, Room 106. For additional information, please visit the CASS

website at www.sa.utep.edu/cass. CASS' staff are the only individuals who can validate and if need be, authorize accommodations for students.

Scholastic Dishonesty: Any student who commits an act of scholastic dishonesty is subject to discipline. Scholastic dishonesty includes, but not limited to cheating, plagiarism, collusion, submission for credit of any work or materials that are attributable to another person.

Cheating is copying from the test paper of another student. Communicating with another student during a test to be taken individually. Giving or seeking aid from another student during a test to be taken individually. Possession and/or use of unauthorized materials during tests (i.e. crib notes, class notes, books, etc.). Substituting for another person to take a test. Falsifying research data, reports, academic work offered for credit.

Plagiarism is using someone's work in your assignments without the proper citations. Submitting the same paper or assignment from a different course, without direct permission of instructors. To avoid plagiarism see: https://www.utep.edu/student-affairs/osccr/_Files/docs/Avoiding-Plagiarism.pdf.

Collusion is unauthorized collaboration with another person in preparing academic assignments.

NOTE: When in doubt on any of the above, please contact your instructor to check if you are following authorized procedure.