MIS 611D "TOPICS IN DATA AND WEB MINING" - Spring 2023
Dr. Hsinchun Chen, Professor, Department of MIS

- **Instructor:** Hsinchun Chen, Ph.D., Professor, Management Information Systems Dept, Eller College of Management, University of Arizona; Director, AI Lab & AZSecure Program
- **Classroom Time/Room:** TH 12:30AM-3:15PM in-person, MCCL 430E
- **Instructor's Office Hours:** T/TH 10:45-11:45AM in-person or via Zoom
- **Office/Phone:** MCCL 430X, (520) 621-4153
- **Email/web site:** hsinchun@arizona.edu (email is the best way to reach me!);
- https://eller.arizona.edu/people/hsinchun-chen
- **Class Web site: https://ailab-ua.github.io/courses/ (IMPORTANT!) All class slides, papers, and readings are hosted on this permanent and open Github site. Class communications, assignments, submissions, and gradings will be supported by the UA/Eller D2L system (by TAs).**
- **Teaching Assistants (TAs):** Ben Ampel bampel@arizona.edu and Steven Ullman stevenullman@arizona.edu, 4th-year Ph.D. students (office: MCCL 430 Cubical #34-35); TA Office Hours: TBD

CLASS MATERIAL (Optional)

- *Machine Mining with PyTorch and Scikit-Learn,* by Raschka, Liu and Mirjalili, Packt Publishing, February 2022.

Additional readings/handouts will be distributed in class and made available via the class web site.

COURSE OBJECTIVES

Business intelligence and analytics and the related field of big data analytics and data science have become increasingly important in both the academic and the business communities over the past two decades. The *IBM Tech Trends Report* identified business analytics as one of the four major technology trends in the 2010s and beyond. A report by the McKinsey Global Institute predicted that by 2018, the United States alone will face a shortage of 140,000 to 190,000 people with deep data analytical skills, as well as a shortfall of 1.5 million data-savvy managers with the know-how to analyze big data to make effective decisions. Big data and data science have begun to transform different facets of the society, from e-commerce and global logistics, to smart health and cyber security.

This Ph.D./MS level course (Ph.D. core course) will cover the important concepts and techniques related to data analytics and data science, including: statistical foundation, data mining methods, data visualization, AI, deep learning, and web mining techniques that are applicable to emerging e-commerce, government, and health and security applications. **The course will be conducted in a graduate-level format, containing lectures, discussions, readings, lab sessions, and hands-on research projects. The course will support several diverse human and AI learning strategies: rote learning, learning by rules, learning from examples, learning by analogy, and learning by exploration.** The course will require some basic computing (Python, Java) and database (SQL) background. The course will prepare students to become a data scientist or a data-savvy manager for different businesses. The Course Objectives include the following:

- Students will become familiar with important data analytics, business intelligence, data mining, machine learning, and deep learning **concepts, terminologies, techniques, and algorithms. (rote)**
- Students will learn to use selected **data analytics and visualization tools** such as Tableau, Scikit-Learn, and Python for relevant data analytics applications. **(rules, analogy)**

- Students will learn through **team-based research projects** to adopt and leverage state-of-the-art data extraction, analytics, and visualization methods in important applications and domains, including: business, e-commerce, finance, security and health. **(examples, exploration)**
- Students will learn to turn data into actionable business intelligence and explain and communicate results via **professional presentation and paper** in a scientific, business and managerial context. **(analogy, exploration)**
- Students will be introduced to an **intellectual road map for growth in data analytics and data science**, including: future courses and graduate programs, key publications (conferences, journals, news media), major research groups, federal funding agencies, major companies and their underlying technologies, future applications, etc. **(exploration)**

PREREQUISITE FOR THE COURSE

Programming experience in selected modern computing languages (e.g., Python, Java) and DBMS (SQL) is required. This course is hands-on (but not heavy hand-holding), with support from knowledgeable TAs. The workload will be somewhat heavy (10-15 hours per week on average); so only students who are interested in pursuing a career in data analytics or data science should register for this course. The instructor will allow for sit-in or audit for selected students based on their background and interest.

COURSE TOPICS (Selected topics will be covered)

Topic 1: Introduction (the field of MIS, CS; data analyst, data engineer, data scientist)

- From computational design science in MIS to applied data science
- Business intelligence and analytics, opportunities & techniques
- Big data: the 3 Vs (volume, velocity, variety)
- Emerging AI applications, from face recognition to autonomous vehicle
- Data, text and web mining overview: AI, ML, deep learning
- Data mining and web computing tools (by TAs): Tableau, Scikit-Learn, Hadoop, SPARK

Topic 2: Web Computing/Mining (the changing "information/data" world; critical applications and underlying technologies)

- Web 1.0, Surface Web, 1995-: WWW, search engines, spidering, indexing/searching, graph search, genetic algorithms
- Web 2.0, Social Web, 2005-: deep web, social media, crowdsourcing systems, network sciences, recommender systems
- Web 3.0, Mobile Web, 2010-: IoTs, mobile & cloud computing, big data analytics, dark web, cybersecurity
- Web 4.0, AI/Smart Web, 2015-: AI-empowered society, image recognition, machine translation, smart home/city/health, cybersecurity, privacy, political disinformation, deepfake

Topic 3: Data Mining & Machine Learning (the analytics techniques; traditional statistical and machine learning algorithms)

- Statistical analysis: Regression, Naïve Bayes, Principal Component Analysis (PCA)
- Symbolic learning: Decision Trees, Random Forest
- Statistical machine learning: Support Vector Machines (SVM), Matrix Factorization
- Network Analysis: Social Network Analysis (SNA), graph models

Topic 4: Text Mining (handling unstructured text; a multilingual world)

- Information retrieval & extraction, search engines: vector space model, tfidf, entity & topic extraction, search engines design
- Sentiment and affect analysis: lexicon-based, machine learning based
- Topic modeling; word embedding
- Information/data visualization: scientific, text and web visualization; Tableau

Topic 5: AI & Deep Learning (basic and advanced DL algorithms and strategies; an AI-empowered world)

- Artificial neural networks: Multi-Layered Perceptron, Feedforward-Backpropagation Networks (MLP, FFBP NN), Self-Organizing Maps (SOM)
- Basic Deep Learning: Convolutional NN, Recurrent NN, Long Short-Term Memory
- Advanced Deep Learning: Transformers, BERT, GPT, GALL-E, Graph Neural Networks (GNN), GCN, GAT; Transfer Learning, Contrastive Learning

GRADING POLICY (ABSOLUTE GRADE SCALE, A: 90+; B: 80+; C: 70+; D 70-)

Team project proposal	5%
Team lab assignment 1 (Tableau)	10%
Midterm team	30%
Team review paper	15%
Team lab assignment 2 (Scikit-Learn)	10%
Team research project	30%
Class attendance and participation	10%
TOTAL:	110%

TEAM PROJECT PROPOSAL (5%)

Each student will be required to form **a team of 2 members** with complementary skills (e.g., application knowledge, Python, SQL, analytics, presentation). A team proposal (3 pages, Word document) including plan for both review paper (see below) and research project (see below) will be submitted by each team in the third week of the semester. The proposal needs to justify the selection of application area and includes preliminary ideas or plan for execution.

TEAM LAB ASSIGNMENTS (20%)

In order to improve students' hands-on data analytics knowledge and to facilitate final project execution, there will be two Team Lab Assignments: **Lab 1 Tableau (visualization) and Lab 2 Scikit-Learn (analytics, any ML or DL methods)**, both are popular data analytics/visualization tools used by data analysts/scientists. Each team is required to identify 2-3 public or open data sources (e.g., data.gov, Kaggle, UCI) in the application area of their final Research Project (e.g., security, health, finance, e-commerce) and execute selected meaningful data exploration/visualization or analytics (3-4 types) functions. Each assignment is worth 10% of final grade. A team report summarizing results with meaningful screen shots (5 pages, IEEE format) needs **to be submitted in two weeks for each assignment via D2L.** Students are expected to become familiar with selected data extraction, analytics and visualization tools and software.

MIDTERM EXAM (30%)

The midterm exam will be **closed book, closed notes and in the short-essay format**. The questions will be based mostly on classroom lectures. There will be NO Final Exam for this class.

TEAM REVIEW PAPER (15%)

Each team will select an emerging, specific data analytics application area of interest (e.g., health, finance, e-commerce, security) and develop a comprehensive review paper (5 pages, IEEE format) for the topic. **Secondary literature review (10-20 references)** will be needed based on recent papers published in major news media, magazines, conferences, and journals. **The paper will be submitted via D2L.**

TEAM RESEARCH PROJECT PRESENTATION/PAPER (30%)

Each team will be required to propose and execute an interesting and **meaningful data analytics (selected DL methods) research project** for applications of interest to the students. The instructor will suggest suitable data and algorithms for consideration. The class TAs will also provide assistance in data preparation and analytics using selected open source tools. **Each team (ALL students) will present at the end of the semester (20 minutes with 15 PPT slides) and a final research paper (8 pages, IEEE format) will be submitted via D2L after all presentation sessions.** The instructor will provide details about the final paper format and structure. Students are expected to gain significant hands-on data analytics skills and knowledge and professional project communication and presentation experiences.

ATTENDANCE, PARTICIPATION AND ACADEMIC INTEGRITY (10%)

Students are required to attend all lectures on time and honor academic integrity. Missing classes will result in loss of points or administrative drop by the instructor. Students are required to send excuse notes (via email) to the instructor before missing classes. Students are permitted to bring laptop to classroom for note taking purposes, but not for checking email or web surfing. Professional attitude and strong work ethics are needed for this class. Students are encouraged to consult the instructor for advice and help.

LAB SESSIONS and GUEST SPEAKERS

Selected lab sessions will be provided by the class TAs during the semester on the following topics: Python, Tableau, Scikit-Learn, etc. Selected guest speakers may be invited to present in the class.

D2L CLASS SUPPORT

The class will be supported by D2L in the following areas: (1) class announcements, assignments, and email to the entire class, (2) students submitting assignments, papers, and presentation slides online, and (3) grade postings and notifications for all students.

COURSE OUTLINE (tentative)

DATE	TOPIC	CONTENT/NOTES
Jan 12 (Th)	Syllabus & registration; AI tools	Class roster, syllabus
Jan 19 (Th)	**MIS, CS, Design Science Overview**	Overview, discussions
	Big data, applications, research template	Readings, discussions
Jan 20 (Fr)	**Python review**	**TA session/lab**
Jan 26 (Th)	BI, data analytics, data mining, ML, DL, AI	Readings, discussions
	Web Computing	Overview, readings
	PROPOSAL DUE (REVIEW & RESEARCH, 5%)	
Feb 2 (Th)	**Tableau review**, info viz	**TA session/lab**
	Web 1.0, surface web, search algorithms	Readings, lectures
Feb 9 (Th)	Web 2.0, social web, network science	Readings
	Web 3.0, mobile web, cybersecurity, SFS	Readings
Feb 16 (Th)	Web 4.0, smart/AI web, SPARK	Readings
	LAB 1 DUE (TABLEAU, 10%)	
	Data Mining/Machine Learning/Text Mining	Overview, discussions
Feb 22 (Th)	Statiscal analysis, regression	Slides
	Naïve Bayes, decision trees	Slides, readings
Mar 2 (Th)	Classification, KNN, Random Forest	Slides
	Support Vector Machines (SVM)	Slides
Mar 4-12	**SPRING RECESS**	NO CLASS
Mar 16 (Th)	Clustering, k-means, hierarchical clustering	Slides
	SatScan, text mining	Readings
Mar 23 (Th)	Topic modeling, info viz	Slides
	Scikit-Learn review	**TA session/lab**
Mar 30 (Th)	**MIDTERM (30%)**	
	Deep Learning/AI	Overview, readings
Apr 6 (Th)	AI/DL history	Slides, readings
	Multi-Layer Perceptron (MLP, FFBP), SOM	Slides
Apr 13 (Th)	DL overview	Slides, readings
	REVIEW PAPER DUE (15%)	
	Convolutional NN for images	Slides
Apr 20 (Th)	Recurrent NN/LSTM for sequences	Slides
	LAB 2 DUE (SciKit-Learn, 10%)	
	Transofrmers, BERT, GPT	Slides, readings
Apr 27 (Th)	Graph NN, GCN, GAT	Slides, readings
	Future of AI	Readings
May 2 (Tu)	**RESEARCH PROJECT PRESENTATION (15%)**	Teams 1-5
May 5 (Fr)	**RESEARCH PROJECT PAPER DUE (15%)**	Hand in Dr. Chen's office 430X
May 5-11	FINAL EXAM WEEK	NO EXAM FOR MIS 611D