MIS 464 "DATA ANALYTICS" - Spring 2020

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
- **Time/Classroom:** T/TH 9:30AM-10:45AM MCCL 130
- **Instructor's Office Hours:** T/TH 10:45-11:45AM or by appointment
- **Office/Phone:** MCCL 430X, (520) 621-4153
- **Email/web site:** hchen@eller.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 TA).**
- **Teaching Assistant (TA):** (Reza) Mohammadreza Ebrahimi ebrahimi@email.arizona.edu, 4th-year Ph.D. student (office: MCCL 430 Cubical #34-35); TA Office Hours: TA hours (before assignment and project deadlines) will be announced via email and D2L.

CLASS MATERIAL (Optional)

- *Data Mining: Practical Machine Learning Tools and Techniques*, by Witten, Frank, Hall & Pal, 4th Edition, 2017, Morgan Kaufmann (also with a 5-week MOOC course). See more at: http://www.cs.waikato.ac.nz/ml/weka/
- *Artificial Intelligence: A Modern Approach*, by Russel & Norvig, 3rd Edition, 2000, Prentice Hall
- *Deep Learning*, by Goodfellow, Bengio & Courville, 2016, MIT Press

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 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 undergraduate senior level course (elective) will cover the important concepts and techniques related to data analytics, 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.** Most business school seniors with proper background and interest are welcome. 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, Weka, 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**, 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)**

The course will introduce students to a possible career as a data analyst (BS level) and a potential path to become a data engineer (MS level) or even a data scientist (mostly Ph.D. level) in the future.

PREREQUISITE FOR THE COURSE

Programming experience in selected modern computing languages (e.g., Python, Java, C++) and DBMS (SQL). This course is hands-on (but not heavy hand-holding), with support from a knowledgeable TA. 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 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 in CS
- Business intelligence and analytics, opportunities & techniques
- 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, Weka, Hadoop, SPARK

Topic 2: Web Mining/Computing (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, web services & mesh-ups, social media, crowdsourcing systems, network sciences, recommender systems
- Web 3.0, Mobile Web, 2010-: IoTs, mobile & cloud computing, big data analytics, dark web, mobile analytics, cybersecurity
- Web 4.0, AI Web, 2015-: AI-empowered society, 5G, image recognition, machine translation, smart home/city/health, cybersecurity, privacy, political disinformation, deepfake

Topic 3: Data Mining (the analytics techniques; machine learning, deep learning)

- Symbolic learning: decision trees, random forest
- Statistical analysis: regression, Principal Component Analysis (PCA), Naïve Bayes
- Statistical machine learning: Support Vector Machines (SVM), Hidden Markov Models (HMM), Conditional Random Fields (CRF), Matrix Factorization

- Neural networks and soft computing: Feedforward-Backpropagation networks (FFBP NN), Self-Organizing Maps (SOM), Genetic Algorithms
- Network Analysis: social network analysis (SNA), graph models
- Deep learning: Convolutional NN, Recurrent NN, Long Short-Term Memory
- Representation Learning: Transfer Learning, Deep Generative Models

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

- Digital library and search engines
- Information retrieval & extraction: vector space model, entity & topic extraction
- Authorship analysis: lexical, syntactic, structural, and semantic analysis
- Sentiment and affect analysis: lexicon-based, machine learning based
- Topic modeling; word embeddings
- Information visualization: scientific, text and web visualization

Topic 5: Future Directions in Data Analytics (major courses, conferences, groups, and opportunities)

- Other relevant UA MS/Ph.D. courses/programs: business intelligence (MIS587), DM/ML (MIS545, MATH574M, ECE523), web mining/computing (MIS510), big data (MIS584, MIS586), SNA (SOC526), statistical NLP (LINQ539), optimization (SIE545), econometrics (ECON418), etc.
- Important news and scientific media: Science, Nature; The Economist, NYT, WSJ
- Emerging research in major data and web mining conferences: NIPS, ICLR, ICML; AAAI, IJCAI; ACM KDD, IEEE ICDM, WWW; ACM SIGIR, ACM CHI
- Key journals: MISQ, ISR, JMIS; IEEE TKDE, ACM TOIS; JAMIA, JBI, JASIST
- Emerging research in major academic institutions: Stanford, Berkeley, CMU, UW
- Emerging research in major industry research labs: Google, Facebook, Amazon, Netflix, Microsoft
- Emerging data and web mining applications: smart health, smart city, e-commerce, AV, drones, robotics, 5G, privacy, political disinformation

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

• Team project proposal (1/30)	5%
• Team lab assignments (2/25; 4/7)	20%
• Midterm exam (3/17)	30%
• Team review paper (4/7)	15%
• Team research project (4/28)	30%
• Class attendance and participation	10%
TOTAL	110%

TEAM PROJECT PROPOSAL (5%)

Each student will be required to form a two-person team 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: Tableau (visualization) and Weka (analytics), both are popular data analytics/visualization tools used by data analysts/scientists. Each team is required to identify 1-2

public or open data sources (e.g., data.gov, Kaggle; will be introduced in class) 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 functions. Each assignment is worth 10% of final grade. A team report summarizing results with meaningful screen shots (5 pages, IEEE format; will be explained in class) needs to be submitted in two weeks for each assignment. 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. Academic integrity will be strictly enforced. Consequences for cheating will be severe.

REVIEW PAPER AND PRESENTATION (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. Each team (both students) will be required to present their review in the second half of the semester (10 minutes with 8 slides). The instructor will suggest selected emerging application areas for consideration.

TEAM RESEARCH PROJECT PRESENTATION/PAPER (30%)

Each team will be required to propose and execute an interesting and meaningful data analytics research project for applications of interest to the students. The instructor will suggest suitable data and algorithms for consideration. The class TA will also provide assistance in data preparation and analytics using selected open source tools. Each team (both students) will present at the end of the semester (15 minutes with 12 slides) and a final research paper (8 pages, IEEE format) will be submitted 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 TA during the semester on the following topics: Python, Tableau, Weka, etc. Selected guest speakers may be invited to present in the class.

D2L CLASS SUPPORT

The class will be supported by D2L (by class TA) in the following areas: (1) class announcements, assignments, and email to the entire class, (2) students submitting assignments, papers, and presentation slides online, (3) grade postings and notifications for all students, (4) optional periodic quizzes to gauge students' progress and understanding.

COURSE OUTLINE (tentative)

DATE	TOPIC	CONTENT/NOTES
Jan 16	Syllabus & registration	Class roster, syllabus
Jan 17 (F)	**Python** review I	**TA session/lab**
Jan 21 (T)	**MIS, CS, Design Science Overview**	Readings, discussions
Jan 23	Big data, applications	Readings, discussions
Jan 24 (F)	**Python** review II	**TA session/lab**
Jan 28 (T)	BI, data analytics, data mining, ML	Readings, discussions
Jan 30	AI, deep learning	Readings, discussions
	PROPOSAL DUE (REVIEW & RESEARCH, 5%)	
Feb 4 (T)	**Web Computing & Mining**	Overview, discussions
Feb 6	**Tableau**, Hadoop, SPARK	**TA session/lab**
Feb 11 (T)	Web 1.0, Surface Web	Overview, discussions
Feb 13	Search engine, graph search	Readings, lecture
Feb 18 (T)	Web 2.0, Social Web	Overview, discussions
Feb 20	Deep web, social media, SNA	Readings, lecture
Feb 25 (T)	Web 3.0, Mobile Web, IoT, dark web	Overview, discussions
	LAB 1 DUE (TABLEAU, 10%)	
Feb 27	Web 4.0, AI Web, 5G, privacy	Overview, discussions
Mar 3 (T)	**Data Mining**	Overview, discussions
Mar 5	Symbolic learning, AI, decision trees	ID3, RF
Mar 9-13	SPRING RECESS	NO CLASS
Mar 17 (T)	**MIDTERM EXAM (30%)**	
Mar 19	Statistical analysis, regression, Bayes	Overview, discussions
Mar 24 (T)	**Weka,** DM tools	**TA session/lab**
Mar 26	Statistical ML, SVM, CRF	Readings, lecture
Mar 31 (T)	Neural networks, Backprop, SOM	Readings, lecture
Apr 2	Deep learning, CNN	Readings, lecture
Apr 7 (T)	**REVIEW PAPER/PRESENTATION (15%)**	
	LAB 2 DUE (WEKA, 10%)	
Apr 9	**REVIEW PAPER/PRESENTATION**	
Apr 14 (T)	Deep learning, LSTM	Readings, lecture
Apr 16	**Text Mining**	Overview, discussions
Apr 21 (T)	IE, Sentiment analysis, Topic modeling	Readings, lecture
Apr 23	Information Visualization	Readings, lecture
Apr 28 (T)	**RESEARCH PROJECT PRESENTATION (15%)**	
Apr 30	**RESEARCH PROJECT PRESENTATION**	
May 5 (T)	**RRESEARCH PROJECT PRESENTATION**	
May 8-14	FINAL EXAM WEEK	NO EXAM FOR MIS 464
May 8 (F)	**RESEARCH PROJECT PAPER DUE (15%)**	