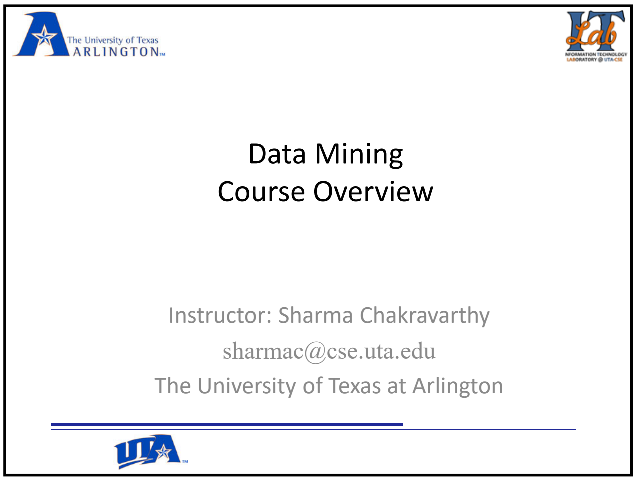
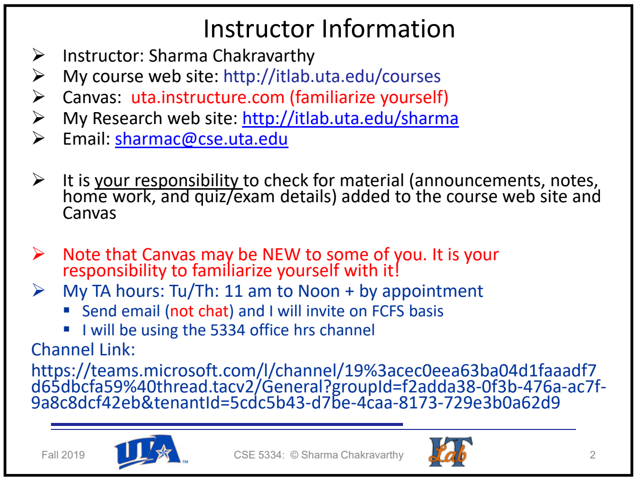


Data Mining
Course Overview

Instructor: Sharma Chakravarthy
sharmac@cse.uta.edu
The University of Texas at Arlington



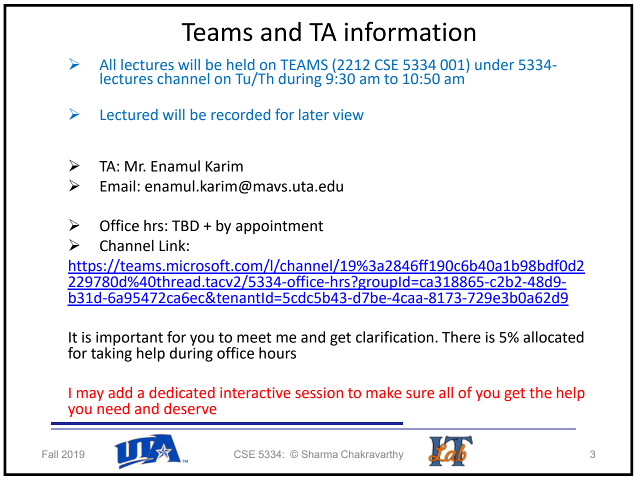
Instructor Information

- Instructor: Sharma Chakravarthy
- My course web site: http://itlab.uta.edu/courses
- Canvas: uta.instructure.com (familiarize yourself)
- My Research web site: http://itlab.uta.edu/sharma
- Email: sharmac@cse.uta.edu

- It is your responsibility to check for material (announcements, notes, home work, and quiz/exam details) added to the course web site and Canvas

- Note that Canvas may be NEW to some of you. It is your responsibility to familiarize yourself with it!
- My TA hours: Tu/Th: 11 am to Noon + by appointment
  - Send email (not chat) and I will invite on FCFS basis
  - I will be using the 5334 office hrs channel

Channel Link:
https://teams.microsoft.com/l/channel/19%3acec0eea63ba04d1faaadf7d65dbcfa59%40thread.tacv2/General?groupId=f2adda38-0f3b-476a-ac7f-9a8c8dcf42eb&tenantId=5cdc5b43-d7be-4caa-8173-729e3b0a62d9



Teams and TA information

- All lectures will be held on TEAMS (2212 CSE 5334 001) under 5334-lectures channel on Tu/Th during 9:30 am to 10:50 am

- Lectured will be recorded for later view

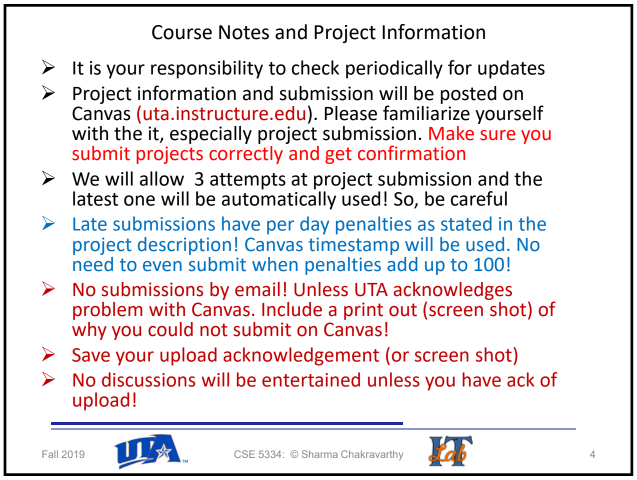
- TA: Mr. Enamul Karim
- Email: enamul.karim@mavs.uta.edu

- Office hrs: TBD + by appointment
- Channel Link:

https://teams.microsoft.com/l/channel/19%3a2846ff190c6b40a1b98bdf0d2229780d%40thread.tacv2/5334-office-hrs?groupId=ca318865-c2b2-48d9-b31d-6a95472ca6ec&tenantId=5cdc5b43-d7be-4caa-8173-729e3b0a62d9

It is important for you to meet me and get clarification. There is 5% allocated for taking help during office hours

I may add a dedicated interactive session to make sure all of you get the help you need and deserve



Course Notes and Project Information

- It is your responsibility to check periodically for updates
- Project information and submission will be posted on Canvas (uta.instructure.edu). Please familiarize yourself with the it, especially project submission. Make sure you submit projects correctly and get confirmation
- We will allow 3 attempts at project submission and the latest one will be automatically used! So, be careful
- Late submissions have per day penalties as stated in the project description! Canvas timestamp will be used. No need to even submit when penalties add up to 100!
- No submissions by email! Unless UTA acknowledges problem with Canvas. Include a print out (screen shot) of why you could not submit on Canvas!
- Save your upload acknowledgement (or screen shot)
- No discussions will be entertained unless you have ack of upload!

TEAMS usage norms (please follow)

- DO Not start the meeting. I will start the meeting and then you can join. DO NOT initiate recording
- Keep your audio off except when you are interacting (questions, clarifications etc.)
  - Raise your hand to make it easier for me to see and call on you
- Keep the video on if your bandwidth allows. I would like to get to know you by face as it is online throughout
  - TEAMS now shows more people and has different modes. Check out the together mode
- I would like to meet with EVERYONE over the first week during my office hours so I get to know your background and your motivation for taking this course
- Lectures will conducted on “General” channel and will be recorded.

Channel link: https://teams.microsoft.com/l/channel/19%3ab4403d0d23a949429f323d74abc19a08%40thread.tacv2/office-hrs?groupId=f2adda38-0f3b-476a-ac7f-9a8c8dcf42eb&tenantId=5cdc5b43-d7be-4caa-8173-729e3b0a62d9

Tests, uploads, and submissions

- We will use the lockdown browser with Respondus for tests or in-person tests in class rooms; Section 900 will be completely online!
- If you are not familiar, please familiarize before the first test
- There will be 3 tests – Test 1, Test 2, and Test3/Final exam. See schedule for details
- All submission will be uploaded to Canvas. Late submissions may not be accepted, or if accepted carries heavy penalty. See schedule and/or project description for details.
  - **Keep your receipt of submission (or a screen shot showing date and time). Without that NO discussions will be entertained on submission**

Tests, projects, and HW Breakdown

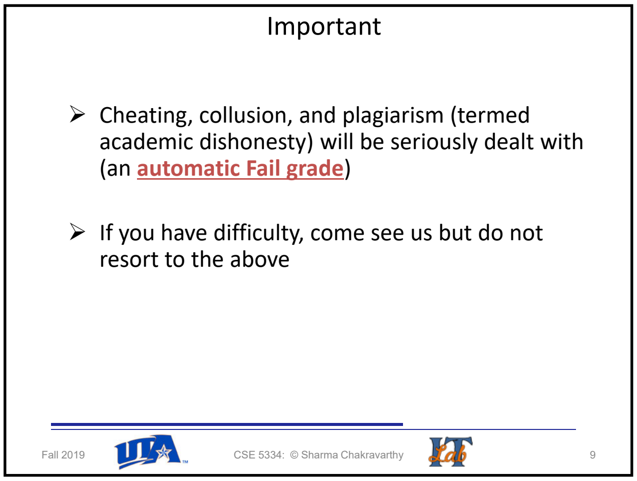
See Schedule for dates and additional details

➢ We will have 3 tests	45%
➢ We will have 3 projects	45%
➢ Asking questions during class/office hrs	5%
➢ Asking meaningful questions on Canvas discussion board	5%
TOTAL	100%

- Typically, class average and ½ to 1 std deviation around it is a B
  - 1 std deviation above class average is a guaranteed A
  - 1 std deviation below class average is likely to be a C
  - And so on

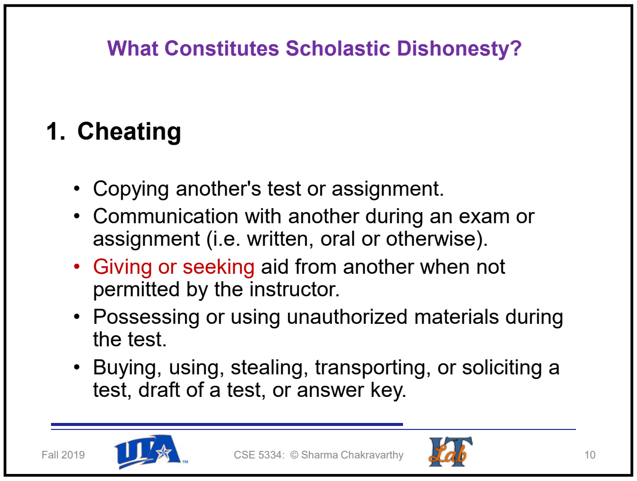
Distance education contacts

- If this course is also offered as a distance education course
- Website address:
  www.uta.edu/engineering/distance
- For technical problems email:
  login.problems@engineering.uta.edu



Important

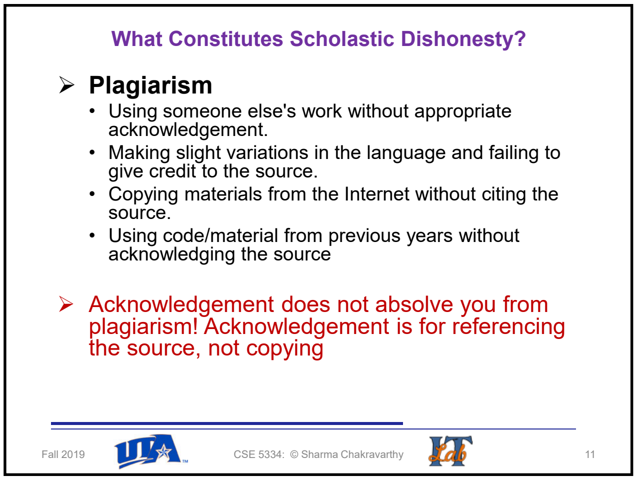
- ➢ Cheating, collusion, and plagiarism (termed academic dishonesty) will be seriously dealt with (an **automatic Fail grade**)
- ➢ If you have difficulty, come see us but do not resort to the above



What Constitutes Scholastic Dishonesty?

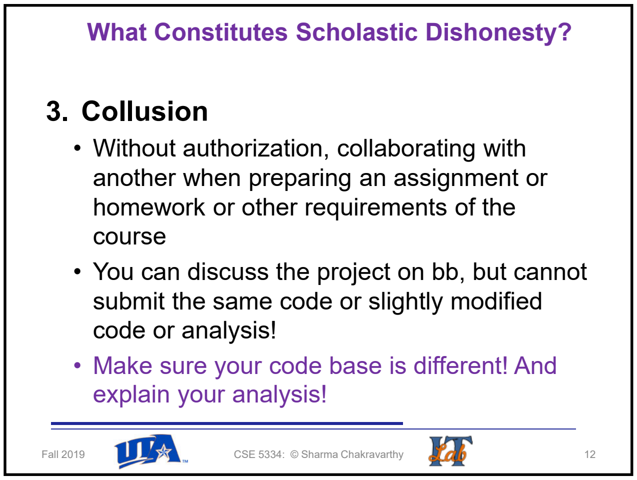
1. Cheating

- Copying another's test or assignment.
- Communication with another during an exam or assignment (i.e. written, oral or otherwise).
- Giving or seeking aid from another when not permitted by the instructor.
- Possessing or using unauthorized materials during the test.
- Buying, using, stealing, transporting, or soliciting a test, draft of a test, or answer key.



What Constitutes Scholastic Dishonesty?

- ➢ **Plagiarism**
  - Using someone else's work without appropriate acknowledgement.
  - Making slight variations in the language and failing to give credit to the source.
  - Copying materials from the Internet without citing the source.
  - Using code/material from previous years without acknowledging the source
- ➢ Acknowledgement does not absolve you from plagiarism! Acknowledgement is for referencing the source, not copying



What Constitutes Scholastic Dishonesty?

3. Collusion

- Without authorization, collaborating with another when preparing an assignment or homework or other requirements of the course
- You can discuss the project on bb, but cannot submit the same code or slightly modified code or analysis!
- Make sure your code base is different! And explain your analysis!

Overview

- This is a first course on Data mining at the CSE department @ UTA. Many a times undergraduate/graduate courses are combined and taught as one course!
- The emphasis of this course is on understanding:
  - underpinnings of mining
  - the plethora of mining techniques,
  - data sets and how to prepare them,
  - how to choose an appropriate technique, and
  - meaningful analysis of results!
  - Visualization is becoming important and will be part of the projects

Organization of the course

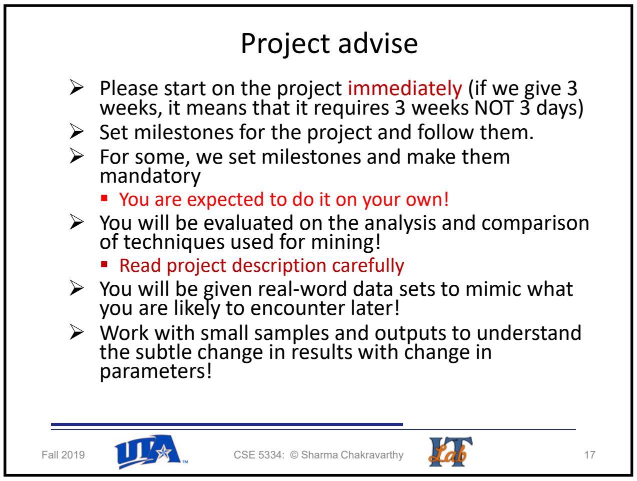
- 4 modules
  1. Intro to mining; what is not mining, overview of supervised and unsupervised learning, predictive modeling
  2. Cluster analysis (k-means, DBSCAN)
  3. Association rules (Apriori and FP tree)
  4. Graph-based approaches to mining, FSG or frequent subgraph mining
- 3 implementation/analysis/visualization projects using R and R Studio (may use Pandas as well)
- 3 tests /quizzes (either in-person or lockdown browser+respondus)
- Practice problems are assigned (and checked if submitted) to help prepare for quizzes/exams (no grade)

How to do well?

- Attend/view all lectures
- Do follow up reading before and immediately after the lecture (not 1 day before the exam)
- Come prepared and ask questions in the class
- Make the class interactive
- There are NO dumb or trivial questions; all questions are important
- Solve all practice problems yourself and submit it
- Make use of my (and TA's) office hours
  - Come to office hours and ask questions
- Challenge yourself and me!

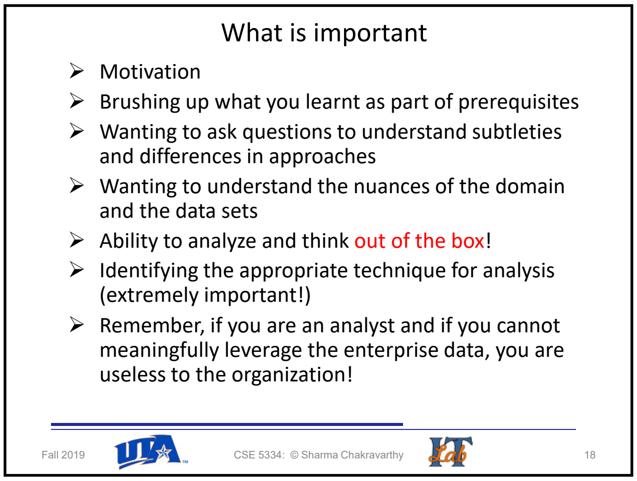
Project Teams Rules

- You can form teams of at most 2 students for doing the project (self-subscribe on Canvas)
- You are responsible for choosing your partner (you can also do the project alone)
- Once you choose, you CANNOT change the partner (you cannot have a different partner for each project)
- **Both members of the team will get the SAME grade**
- **I will not entertain any complaints against each other EXCEPT plagiarism**
- **If one is caught cheating, both will be reported to UTA or take any other judicial action I need to**
- **So, choose your partner wisely!**



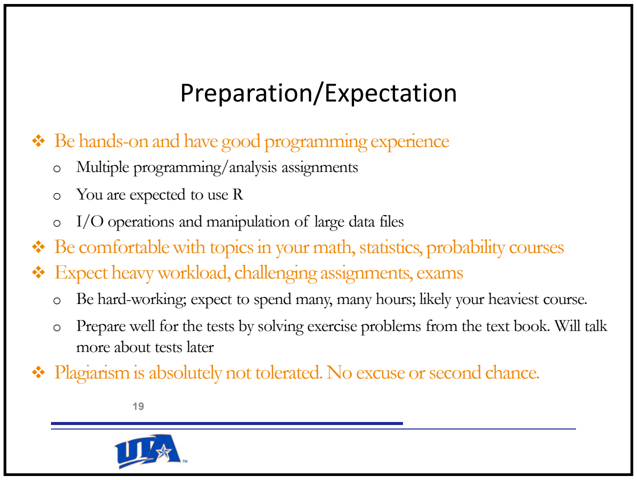
Project advise

- Please start on the project immediately (if we give 3 weeks, it means that it requires 3 weeks NOT 3 days)
- Set milestones for the project and follow them.
- For some, we set milestones and make them mandatory
  - You are expected to do it on your own!
- You will be evaluated on the analysis and comparison of techniques used for mining!
  - Read project description carefully
- You will be given real-word data sets to mimic what you are likely to encounter later!
- Work with small samples and outputs to understand the subtle change in results with change in parameters!



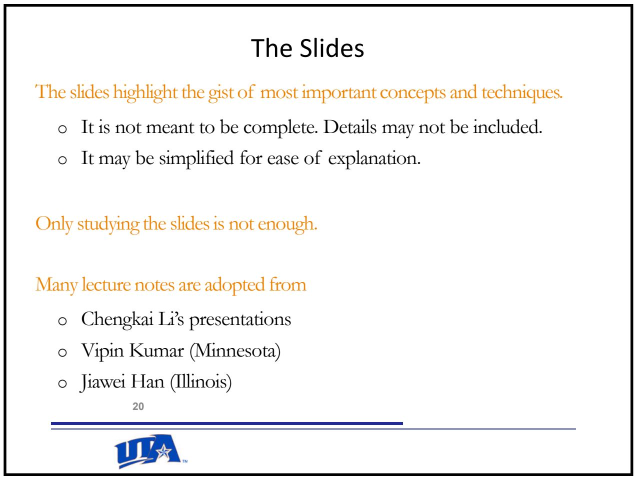
What is important

- Motivation
- Brushing up what you learnt as part of prerequisites
- Wanting to ask questions to understand subtleties and differences in approaches
- Wanting to understand the nuances of the domain and the data sets
- Ability to analyze and think out of the box!
- Identifying the appropriate technique for analysis (extremely important!)
- Remember, if you are an analyst and if you cannot meaningfully leverage the enterprise data, you are useless to the organization!



Preparation/Expectation

- Be hands-on and have good programming experience
  - Multiple programming/analysis assignments
  - You are expected to use R
  - I/O operations and manipulation of large data files
- Be comfortable with topics in your math, statistics, probability courses
- Expect heavy workload, challenging assignments, exams
  - Be hard-working; expect to spend many, many hours; likely your heaviest course.
  - Prepare well for the tests by solving exercise problems from the text book. Will talk more about tests later
- Plagiarism is absolutely not tolerated. No excuse or second chance.



The Slides

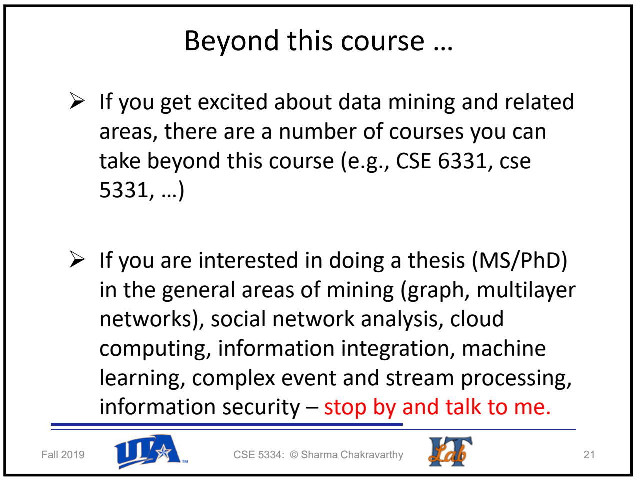
The slides highlight the gist of most important concepts and techniques.

- It is not meant to be complete. Details may not be included.
- It may be simplified for ease of explanation.

Only studying the slides is not enough.

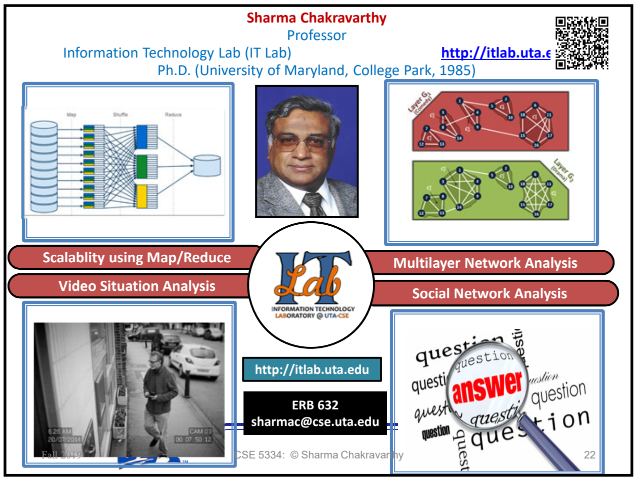
Many lecture notes are adopted from

- Chengkai Li's presentations
- Vipin Kumar (Minnesota)
- Jiawei Han (Illinois)



Beyond this course …

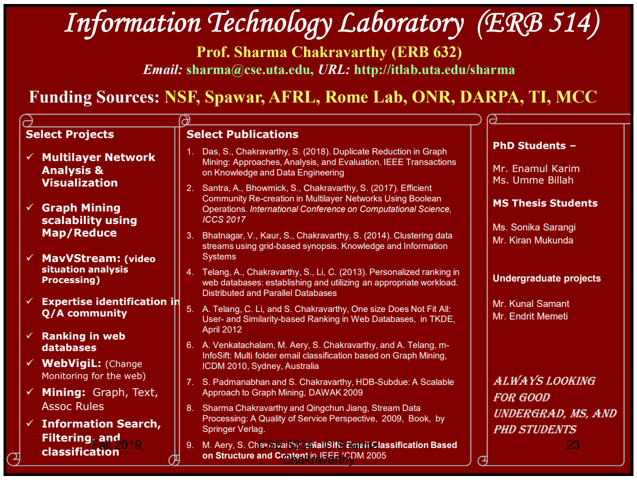
- ➢ If you get excited about data mining and related areas, there are a number of courses you can take beyond this course (e.g., CSE 6331, cse 5331, …)
- ➢ If you are interested in doing a thesis (MS/PhD) in the general areas of mining (graph, multilayer networks), social network analysis, cloud computing, information integration, machine learning, complex event and stream processing, information security – stop by and talk to me.



Sharma Chakravarthy
Professor
Information Technology Lab (IT Lab) http://itlab.uta.e
Ph.D. (University of Maryland, College Park, 1985)

http://itlab.uta.edu

ERB 632
sharmac@cse.uta.edu



Information Technology Laboratory (ERB 514)

Prof. Sharma Chakravarthy (ERB 632)
Email: sharma@cse.uta.edu, *URL:* http://itlab.uta.edu/sharma

Funding Sources: NSF, Spawar, AFRL, Rome Lab, ONR, DARPA, TI, MCC

Select Projects

- ✓ **Multilayer Network Analysis & Visualization**
- ✓ **Graph Mining scalability using Map/Reduce**
- ✓ **MavVStream:** (video situation analysis Processing)
- ✓ **Expertise identification in Q/A community**
- ✓ **Ranking in web databases**
- ✓ **WebVigiL:** (Change Monitoring for the web)
- ✓ **Mining:** Graph, Text, Assoc Rules
- ✓ **Information Search, Filtering, and classification**

Select Publications

1. Das, S., Chakravarthy, S. (2018). Duplicate Reduction in Graph Mining: Approaches, Analysis, and Evaluation. IEEE Transactions on Knowledge and Data Engineering
2. Santra, A., Bhowmick, S., Chakravarthy, S. (2017). Efficient Community Re-creation in Multilayer Networks Using Boolean Operations. *International Conference on Computational Science, ICCS 2017*
3. Bhatnagar, V., Kaur, S., Chakravarthy, S. (2014). Clustering data streams using grid-based synopsis. Knowledge and Information Systems
4. Telang, A., Chakravarthy, S., Li, C. (2013). Personalized ranking in web databases: establishing and utilizing an appropriate workload. Distributed and Parallel Databases
5. A. Telang, C. Li, and S. Chakravarthy, One size Does Not Fit All: User- and Similarity-based Ranking in Web Databases, in TKDE, April 2012
6. A. Venkatachalam, M. Aery, S. Chakravarthy, and A. Telang, m-InfoSift: Multi folder email classification based on Graph Mining, ICDM 2010, Sydney, Australia
7. S. Padmanabhan and S. Chakravarthy, HDB-Subdue: A Scalable Approach to Graph Mining, DAWAK 2009
8. Sharma Chakravarthy and Qingchun Jiang, Stream Data Processing: A Quality of Service Perspective, 2009, Book, by Springer Verlag.
9. **M. Aery, S. Chakravarthy, eMailSift: Email Classification Based on Structure and Content in IEEE ICDM 2005**

PhD Students –

Mr. Enamul Karim
Ms. Umme Billah

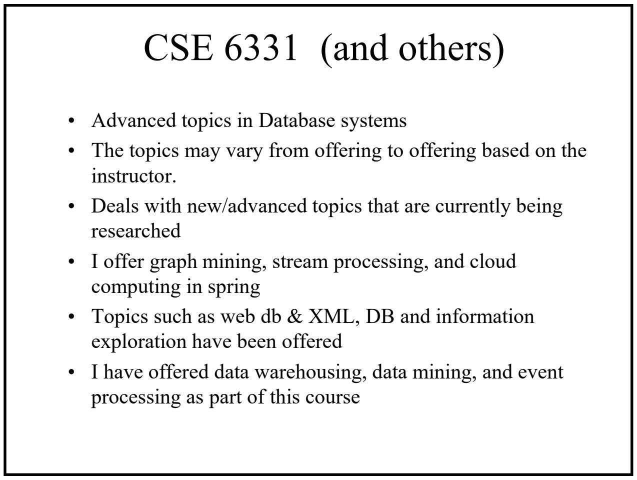
MS Thesis Students

Ms. Sonika Sarangi
Mr. Kiran Mukunda

Undergraduate projects

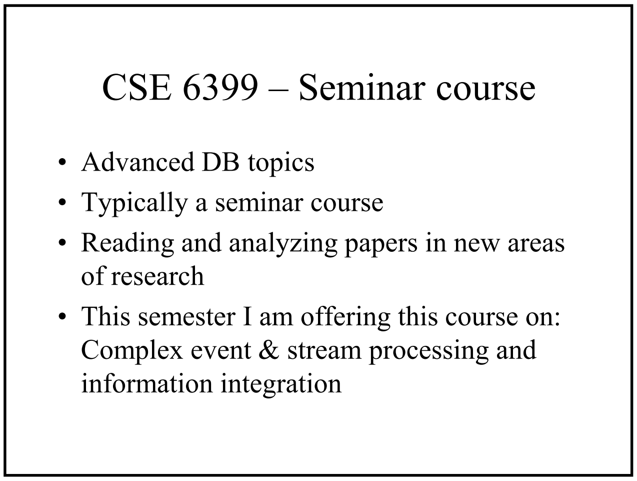
Mr. Kunal Samant
Mr. Endrit Memeti

ALWAYS LOOKING FOR GOOD UNDERGRAD, MS, AND PHD STUDENTS



CSE 6331 (and others)

- Advanced topics in Database systems
- The topics may vary from offering to offering based on the instructor.
- Deals with new/advanced topics that are currently being researched
- I offer graph mining, stream processing, and cloud computing in spring
- Topics such as web db & XML, DB and information exploration have been offered
- I have offered data warehousing, data mining, and event processing as part of this course



CSE 6399 – Seminar course

- Advanced DB topics
- Typically a seminar course
- Reading and analyzing papers in new areas of research
- This semester I am offering this course on: Complex event & stream processing and information integration

Discussion

Thank You !!!

For more information

http://itlab.uta.edu