# Course Information:
- **Time:** Friday 4:00-6:50pm
- **Classroom:** ERB 670
- **Class number:** 29401
- **Homepage:** [http://crystal.uta.edu/~cli/cse6339](http://crystal.uta.edu/~cli/cse6339)

# Instructor: Chengkai Li
- **Office hours:** Thursday 10:30-12:30pm
- **Office:** ERB 628
- **Phone:** (817) 272-0162
- **E-mail:** cli [AT] uta [DOT] edu
- **Homepage:** [http://ranger.uta.edu/~cli](http://ranger.uta.edu/~cli)

---

**Course Description:** Claims of "fact" are constantly made from data--by journalists, politicians, lobbyists, public relations specialists, sports fans, etc. Wherever numbers and data are involved, they can be laden with "lies, d--ed lies, and statistics." Database research has in the past focused on how to answer queries, but has not devoted much attention to discerning the quality of the resulting claims, or to formulating good queries from the outset. There is demand for research that fills this void in many domains where decisions are increasingly driven by data, particularly in journalism. Data-driven fact-checking and lead-finding are growing in importance, as more data become publicly available in the movement of "democratizing data." This course is project-driven. We will build Websites, applications, systems to support public interest journalism. We plan to make our systems available for public use after the course concludes. In this process, we will learn, apply, and invent techniques for database systems, text mining, data mining, Web applications, data visualization, social media and social computational systems, cloud computing, and so on. We will also be exposed to the modern practice of journalism.

**Student Learning Outcomes:** A solid understanding of the basic concepts, principles, and techniques of data management and data mining in computational journalism; an ability to analyze real-world applications, to model data management and data mining problems, and to assess different solutions; an ability to design, implement, and evaluate data management and data mining methods and systems for computational journalism.

**Prerequisites:** CSE 3330/5330 Database Systems I or consent of instructor

**Grades**
- Paper Review 15%
- Paper Presentation 15%
- Tutorial Presentation 15%
- In-class discussion 10%
- Project Presentation, Demo, Website and Paper 45%

The final letter grades will be based on the curve of students' performance.

**Attendance:** At The University of Texas at Arlington, taking attendance is not required. Rather, each faculty member is free to develop his or her own methods of evaluating students’ academic performance, which includes establishing course-specific policies on attendance. As the instructor of this section, I require all students to attend lectures.

**Announcements:** Stay tuned and make sure to check Blackboard frequently. Important announcements will be posted there.
Assignments and Deadlines

- All the assignments must be submitted through Blackboard. We will NOT take hardcopy or email submission, unless the university verifies that Blackboard was malfunctioning or unavailable. If you are not able to submit through Blackboard due to its technical failure, you can email your assignment to us, together with a screenshot showing the technical failure. We will verify with the university.
- Everything is due by 11:59pm on the due date. The deadline is automatically managed by Blackboard. You can still turn in assignment after the deadline. However, you automatically lose 5 points per hour after the due time, till you get 0. (Each individual assignment is 100 points.) We cannot waive the penalty, unless there was a case of illness or other substantial impediment beyond your control, with proof in documents.

Regrading: Regrading request must be made within 7 days after we post scores on Blackboard. TA will handle regrade requests. If student is not satisfied with the regarding results, you get 7 days to request again. The instructor will regrade, and the decision is final.

Drop Policy: Students may drop or swap (adding and dropping a class concurrently) classes through self-service in MyMav from the beginning of the registration period through the late registration period. After the late registration period, students must see their academic advisor to drop a class or withdraw. Undeclared students must see an advisor in the University Advising Center. Drops can continue through a point two-thirds of the way through the term or session. It is the student's responsibility to officially withdraw if they do not plan to attend after registering. Students will not be automatically dropped for non-attendance. Repayment of certain types of financial aid administered through the University may be required as the result of dropping classes or withdrawing. For more information, contact the Office of Financial Aid and Scholarships (http://www.uta.edu/ses/fao).

Americans with Disabilities Act: The University of Texas at Arlington is on record as being committed to both the spirit and letter of all federal equal opportunity legislation, including the Americans with Disabilities Act (ADA). All instructors at UT Arlington are required by law to provide "reasonable accommodations" to students with disabilities, so as not to discriminate on the basis of that disability. Any student requiring an accommodation for this course must provide the instructor with official documentation in the form of a letter certified by the staff in the Office for Students with Disabilities, University Hall 102. Only those students who have officially documented a need for an accommodation will have their request honored. Information regarding diagnostic criteria and policies for obtaining disability-based academic accommodations can be found at www.uta.edu/disability or by calling the Office for Students with Disabilities at (817) 272-3364.

Title IX: The University of Texas at Arlington is committed to upholding U.S. Federal Law “Title IX” such that no member of the UT Arlington community shall, on the basis of sex, be excluded from participation in, be denied the benefits of, or be subjected to discrimination under any education program or activity. For more information, visit www.uta.edu/titleIX.

Academic Integrity: All students enrolled in this course are expected to adhere to the UT Arlington Honor Code:

&quote;I pledge, on my honor, to uphold UT Arlington’s tradition of academic integrity, a tradition that values hard work and honest effort in the pursuit of academic excellence. I promise that I will submit only work that I personally create or contribute to group collaborations, and I will appropriately reference any work from other sources. I will follow the highest standards of integrity and uphold the spirit of the Honor Code.&quot;

Instructors may employ the Honor Code as they see fit in their courses, including (but not limited to) having students acknowledge the honor code as part of an examination or requiring students to incorporate the honor code into any work submitted. Per UT System Regents’ Rule 50101, §2.2, suspected violations of university’s standards for academic integrity (including the Honor Code) will be referred to the Office of Student Conduct. Violators will be disciplined in accordance with University policy, which may result in the student’s suspension or expulsion from the University.
Student Support Services: UT Arlington provides a variety of resources and programs designed to help students develop academic skills, deal with personal situations, and better understand concepts and information related to their courses. Resources include tutoring, major-based learning centers, developmental education, advising and mentoring, personal counseling, and federally funded programs. For individualized referrals, students may visit the reception desk at University College (Ransom Hall), call the Maverick Resource Hotline at 817-272-6107, send a message to resources@uta.edu, or view the information at www.uta.edu/resources.

Electronic Communication: UT Arlington has adopted MavMail as its official means to communicate with students about important deadlines and events, as well as to transact university-related business regarding financial aid, tuition, grades, graduation, etc. All students are assigned a MavMail account and are responsible for checking the inbox regularly. There is no additional charge to students for using this account, which remains active even after graduation. Information about activating and using MavMail is available at http://www.uta.edu/oit/cs/email/mavmail.php.

Student Feedback Survey: At the end of each term, students enrolled in classes categorized as lecture, seminar, or laboratory shall be directed to complete a Student Feedback Survey (SFS). Instructions on how to access the SFS for this course will be sent directly to each student through MavMail approximately 10 days before the end of the term. Each student’s feedback enters the SFS database anonymously and is aggregated with that of other students enrolled in the course. UT Arlington’s effort to solicit, gather, tabulate, and publish student feedback is required by state law; students are strongly urged to participate. For more information, visit http://www.uta.edu/sfs.

Final Review Week: A period of five class days prior to the first day of final examinations in the long sessions shall be designated as Final Review Week. The purpose of this week is to allow students sufficient time to prepare for final examinations. During this week, there shall be no scheduled activities such as required field trips or performances; and no instructor shall assign any themes, research problems or exercises of similar scope that have a completion date during or following this week unless specified in the class syllabus. During Final Review Week, an instructor shall not give any examinations constituting 10% or more of the final grade, except makeup tests and laboratory examinations. In addition, no instructor shall give any portion of the final examination during Final Review Week. During this week, classes are held as scheduled. In addition, instructors are not required to limit content to topics that have been previously covered; they may introduce new concepts as appropriate.

Emergency Exit Procedures: Should we experience an emergency event that requires us to vacate the building, students should exit the room and move toward the nearest exit. When exiting the building during an emergency, one should never take an elevator but should use the stairwells. Faculty members and instructional staff will assist students in selecting the safest route for evacuation and will make arrangements to assist individuals with disabilities.

Student Support Services: UT Arlington provides a variety of resources and programs designed to help students develop academic skills, deal with personal situations, and better understand concepts and information related to their courses. Resources include tutoring, major-based learning centers, developmental education, advising and mentoring, personal counseling, and federally funded programs. For individualized referrals, students may visit the reception desk at University College (Ransom Hall), call the Maverick Resource Hotline at 817-272-6107, send a message to resources@uta.edu, or view the information at www.uta.edu/resources.

Schedule

As the instructor for this course, I reserve the right to adjust this schedule in any way that serves the educational needs of the students enrolled in this course.

<table>
<thead>
<tr>
<th>Date</th>
<th>Lecture</th>
<th>Presentations</th>
<th>Tutorials</th>
<th>Lecture Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>01/23</td>
<td>Course Overview</td>
<td></td>
<td></td>
<td>[PPT]</td>
</tr>
<tr>
<td>01/30</td>
<td>Big Data, Tools, Datasets</td>
<td></td>
<td></td>
<td>[PPT]</td>
</tr>
<tr>
<td>02/06</td>
<td>Data Science and Computational Journalism</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
02/20 (12:30pm)

Data Wrangling/Munging, Cleaning, Integration

Sumesh Balan
[Wrangler]


Ismail Vandeliwala

- Pandas
- Open Refine
- Data Wrangler

02/20

Data Acquisition (API, social media access, web crawling, information extraction, crowdsourcing)

Minumol Joseph
[OIE]

- **Open information extraction from the web.** Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S. Weld. Commun. ACM 51, 12 (December 2008), 68-74.
- **Open information extraction: the second generation.** Oren Etzioni, Anthony Fader, Janara Christensen, Stephen

Fatma Dogan

- Orange
- Many Eyes
- Tableau Public

(Ref: Open IE Software)
<table>
<thead>
<tr>
<th>Date</th>
<th>Event</th>
<th>Coursework</th>
</tr>
</thead>
<tbody>
<tr>
<td>02/23</td>
<td>Project Proposal Due</td>
<td>Soderland, and Mausam Mausam. IJCAI 2011.</td>
</tr>
</tbody>
</table>
| 02/27 | Text Analysis: entity recognition and disambiguation - sentiment analysis - topic detection - text summarization - fact-checking | Pawan Puttaswamy  
- [TwitterCredibility] Information credibility on twitter. Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. WWW 2011.  
| 03/06 | Design Challenges and Misconceptions in Named Entity Recognition | Ismail Vandeliwala  
| 03/13 | Spring Break | Deepa Warrier  
- Illinois Named Entity Tagger  
- Illinois Wikifier  
(Ref: Open Calais) |
| 03/20 | | Deepa Warrier  
- [Sentiment] A sentimental education: sentiment analysis using subjectivity summarization based on               |
|       | | Rohit Bhoopalam  
- Python NLTK  
- Amazon Sentiment Analysis Tutorial |
<table>
<thead>
<tr>
<th>Date</th>
<th>Event</th>
<th>Presenters</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>03/27</td>
<td>Midterm Project Presentation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>04/03</td>
<td>Last day to drop class</td>
<td></td>
<td></td>
</tr>
<tr>
<td>04/17</td>
<td></td>
<td>Rohit Bhoopalam</td>
<td>• [GraphLab] Y. Low, J. Gonzalez, Sona Hasani</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(Ref: Stanford NLP Software, Open NLP)</td>
</tr>
</tbody>
</table>


- Spark

- Giraph

- GraphLab
<table>
<thead>
<tr>
<th>Date</th>
<th>Presentation</th>
<th>Authors</th>
<th>Title</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>05/08</td>
<td>Project Presentation and Demo</td>
<td>Sumesh Balan</td>
<td>Google Fusion Table</td>
<td>MapBox Studio, CartoDB</td>
</tr>
</tbody>
</table>

University calendar: Spring 2015

Readings and Videos

- Data Journalism Handbook
- Journalism in the Age of Data

Resources
• ProPublica
• Center for Investigative Reporting
• Investigative Reporters and Editors
• PolitiFact.com
• FactCheck.org
• Washington Post Fact Checker Blog
• A computational journalism reading list by Jonathan Stray
• Public Affairs Data Journalism course at Stanford
• Computation + Journalism Class at Georgia Tech
• Computational journalism research program at the DeWitt Wallace Center for Media and Democracy, Sanford School of Public Policy, Duke University.

Where to find papers:

• Google
• Google Scholar
• CiteSeer
• DBLP Bibliography
• ACM Digital Library
• IEEE Xplore
• Other Computer Science articles