CSE6242 / CX4242, Spring 2014

Data and Visual Analytics

Georgia Tech (http://www.gatech.edu), College of Computing (http://www.cc.gatech.edu)

1:30 - 3pm, Instructional Center (http://goo.gl/maps/o3b8B) 105, Tue & Thu

Prof. Duen Horng (Polo) Chau (http://www.cc.gatech.edu/~dchau)

This course will introduce you to broad classes of techniques and tools for analyzing and visualizing data at scale. It emphasizes on how to *combine* computation and visualization to perform effective analysis. We will cover methods from each side, and hybrid ones that combine the best of both worlds. Students will work small teams to complete a research project exploring novel approaches for interactive data & visual analytics.

Piazza Discussion Forum

We will use Piazza (http://piazza.com/gatech/spring2014/cse6242cs4242/home) for discussion (e.g., homework, project). Post your questions there, and the teaching staff and your fellow classmates will be able to help answer them quickly. You can also use Pizza to find project teammates.

T-square will only be used for submission of assignments and projects.

Office Hours

Instructor	Polo Chau (http://www.cc.gatech.edu/~dchau)	Thu, 3-4pm, Klaus 1324
TA	Robert Pienta (http://www.linkedin.com/pub/robert-pienta/50/165/542)	Wed, 4-5pm, common area next to Klaus 1324
TA	Long Tran (http://www.cc.gatech.edu/~tqlong/)	Mon, 4-5pm, Klaus 1305
Grader	Alan Zhang	

Schedule (tentative)

Date		Topic	Tue	Thu	Events
Jan	7, 9	* Course introduction, project overview * Big data analytics process & building blocks	Slides (lectures/CSE6242-20140107-Intro.pdf)	Slides (lectures/CSE6242-20140109-ProcessSqlite.pdf)	
	14, 16	* Data Collection, Simple Storage (SQLite) & Cleaning * Data Integration			HW1 out (Tue)
	21, 23	• Visualization fundamentals • Data visualization for the web (D3) • How to present your analysis (to your boss, or for research)			by Chad Stolper (http://chadstolper.wordpress.com)
	28, 30	Classification (techniques, visualization & interaction)			HW1 due (Mon)
Feb	4, 6	Clustering			
	11, 13	Dimensionality Reduction: techniques, visualization, practitioner's guide			

	18, 20	• Graph analytics
		  ◦ basics; power laws; centrality
		  ◦ how to build and store graphs
		  ◦ graph statistics and how to compute them
		  ◦ scalable single-machine graph algorithms			

	25, 27	• Graph analytics (cont'd)
		  ◦ interactive tools
		  ◦ applications
		• Scaling up (Hadoop)			

Mar	4, 6	Scaling up (Pig, HBase, Hive, Pegasus)			
	11, 13	Project proposal presentations			
	18, 20	Spring Break	X	X	

	25, 27	Human Computation			
Apr	1, 3	Time series: algorithms, visualization, & applications			
Apr	8, 10	Text analytics: concepts, algorithms (LSI=SVD), visualization			
	15, 17	Course review			
	22, 24	Project presentations			

Grading

- 40% Homework
- 50% Project
- 10% Class and Piazza participation

Late Submissions Policy

- Homework: each student has *4 slip days* total. No questions asked.
- Project: each team has *3 slip days* total. No questions asked.
- After all slip days are used up, *5% deduction* for every 24 hours of delay. (e.g., 5 points for a 100-point homework)
- No penalties for medical reasons or emergencies. You *must* submit a doctor's note or an official letter explaining the emergency.

Homework (tentative)

HW1 (./hw1/6242_hw1.pdf)	5%
HW2	10%
HW3	15%
HW4	10%

Project

Team project: 3-4 people. Description and grading policy (project.html) (proposal + presentation, progress report, final report + presentation).

Auditors

Auditors must first obtain instructor's permission of the instructor, then enroll in the course. The auditor must attend all lectures, and optionally complete the assignments.

Textbooks and reading materials

None required.
Highly recommended good reads:

- Data Science for Business (http://amzn.com/1449361323) by Foster Provost and Tom Fawcett
- FREE probability book (http://theanalysisofdata.com/probability/0_2.html), by Prof. Guy Lebanon. (From Amazon (http://www.amazon.com/Probability-Analysis-Data-Guy-Lebanon/dp/1479344761/).)
- A nice D3 tutorial (http://alignedleft.com/tutorials/)
- Human Computation book (http://www.morganclaypool.com/doi/abs/10.2200/S00371ED1V01Y201107AIM013) by Edith Law and Luis von Ahn

Prerequisites

For both CSE 6242 (grad) and CX 4242 (undergrad)

Students are expected to complete significant programming assignments (homework, project) that may involve higher-level languages or scripting (e.g., Java, R Matlab, etc.). Basic algebra, probability knowledge is expected.

Additional formal prerequisites for CSE 6242

None.

Additional formal prerequisites for CX 4242

(Undergraduate Semester level MATH 2605 Minimum Grade of D or
Undergraduate Semester level MATH 2401 Minimum Grade of D or
Undergraduate Semester level MATH 24X1 Minimum Grade of D) or
and
(Undergraduate Semester level MATH 3215 Minimum Grade of D or
Undergraduate Semester level MATH 3225 Minimum Grade of D or
Undergraduate Semester level ECE 3077 Minimum Grade of D or

Undergraduate Semester level ISYE 2027 Minimum Grade of D)
and
(Undergraduate Semester level CS 1371 Minimum Grade of C or
Undergraduate Semester level CS 1372 Minimum Grade of C or
Undergraduate Semester level CX 4010 Minimum Grade of C or
Undergraduate Semester level CX 4240 Minimum Grade of C)

Large datasets

- Yahoo WebScope (http://webscope.sandbox.yahoo.com/)

Previous offerings

Spring 2013 (http://poloclub.gatech.edu/cse6242/2013spring) - CSE 6242 / CS 4803-DVA – Polo Chau
Spring 2011 (http://smlv.cc.gatech.edu/dava) - CSE 8803-DVA / CS 4803-DVA - Guy Lebanon
Spring 2010 (http://www.cc.gatech.edu/~lebanon/teaching/DAVA/) - CSE 8803-DVA - Guy Lebanon

Acknowledgements & Related Classes

We thank Amazon's AWS in Education (http://aws.amazon.com/grants/) grant program for providing support for Amazon Web Services (http://aws.amazon.com)

Many thanks to my colleagues for sharing their course materials:

> Prof. John Stasko - Information Visualization - Fall 2012 (http://www.cc.gatech.edu/~stasko/7450)
> Prof. Jeff Heer - Research Topics in Interactive Data Analysis - Spring 2011 (http://hci.stanford.edu/courses/cs448g/)
> Prof. Christos Faloutsos - Multimedia Databases and Data Mining - Fall 2012 (http://www.cs.cmu.edu/~christos/courses/826.F12/)