BDA/GEOG 594 Big Data Science and Analytics Platforms

Web site → http://map.sdsu.edu/bigdata (Fall 2019)

Blackboard URL: https://blackboard.sdsu.edu/

Lectures: Tuesday and Thursday: 2:00pm – 3:15pm Storm Hall 324 (SAL lab)
(Available SAL lab hours: Tue/Thurs from 12pm – 2pm)

Instructor: Dr. Ming-Hsiang Tsou Tel: (619) 594-0205
Office: Storm Hall Room 313C
Office Hour: Tuesday/Thursday 3:30pm- 4:30pm
Email mtsou@sdsu.edu

Overview: This course introduces state-of-the-art computational platforms, tools, and skills for big data science and big data analytics with numerous real-world case studies. The big data field provides untapped potential for discovering and analyzing complex problems faced by humankind, including business analytics, disease outbreaks, traffic patterns, urban dynamics, and environmental changes. This class will introduce big data platforms (Amazon EC2) and key concepts (cloud computing, virtualization, information privacy, and crowd sourcing). Students will learn to how to use Amazon EC2, Google Cloud Platform, MongoDB, R, Gephi, ArcGIS Online, and Tableau to conduct big data analytics. The course will provide basic introduction to big database management related to NoSQL databases, Hadoop and MongoDB. This course will have both the hands-on training of analytics tools and computer skills, as well as the fundamental concepts for **big data science with critical thinking and problem solving**. Students will have the opportunity to create their own big data platform on Amazon EC2 virtual servers, manage their own databases in MongoDB, and access and collect big data from sources of their choosing (e.g. Twitter data and business datasets).

Prerequisites: One minimum computer programming or introductory course (from GEOG 104, CS100, CS107, or equivalent computer programming courses) and one fundamental statistics course (from GEOG 385, STAT 250, or SOC 201 or equivalent statistic courses).

Equipment Required: Each student should bring their own laptop computers (preferred Windows OS) to the class for the web-based lab exercises and class assignments. Student may also use the workstations provided in the computer labs if available.

Required Textbooks:
O'Neil, C., & Schutt, R. (2013). *Doing Data Science: Straight Talk from the Frontline*. O'Reilly Media, Inc.

Lectures: Introduction of Big Data platforms, concepts, tools, and technology updates.

Web-based Exercises (Home Works and Lecture Sessions): This course will provide a series of web-based exercises, built upon the Amazon EC2 server and other software. Students can use their own computer or notebook computer to complete the exercises and submit the results to Blackboard. Web-based assignments are due at the beginning of the lecture on the due date (submitted via Blackboard). Late submissions of assignments will be docked 20% per day, beginning on the due date.

Grading: Midterm exam **20%**, Lab exercises **45%**, Class participation (On-line discussion + weekly individual demos) **10%**, **Final group project 25%.**

(Grading of online discussions for graduate students will place emphasis on the demonstration of critical thinking, literature review, domain knowledge, and multidisciplinary perspectives.)

Graduate students will have an additional assignment (literature review in their specialty areas with Big Data applications). Additional **10%** for grading (the rest of grading components will be scaled to 90%). The literature review will ask the students to gather the following information:

1. Find out two web sites which focus on your own special areas and big data analytics and write a 500-word paragraph to introduce each web site.
2. Write a 1000-word essay about the impact of big data science on an area of your interest and identify the potential connections of big data science with your job or your major field.

(Graduate student assignment due day is one week before the final project presentation. Please submit it in **the Blackboard**).

Required Readings: (electronic copies in the Blackboard reading folder and Google Shared Folder).

1. Aslam AA, Tsou MH, Spitzberg BH, An Li, Gawron JM, Gupta DK, Peddecord KM, Nagel AC, Allen C, Yang JA, Lindsay S. (2014). The Reliability of Tweets as a Supplementary Method of Seasonal Influenza Surveillance, *J Med Internet Res 2014;16(11):e250 URL: http://www.jmir.org/2014/11/e250/, doi:10.2196/jmir.3532*
2. danah boyd & Kate Crawford (2012) Critical Questions for Big Data, Information, Communication & Society, 15:5, 662-679, DOI: 10.1080/1369118X.2012.678878.
3. Lohr, Steve (2014). In Big Data, Shepherding Comes First. The New York Times, 12/15/2014. (URL: http://www.nytimes.com/2014/12/15/technology/in-big-data-shepherding-comes-first-.html) .
4. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. (Executive Summary). URL: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation
5. Marx, V. (2013). Biology: The big challenges of big data. *Nature*, 498(7453), 255-260. http://www.nature.com/nature/journal/v498/n7453/full/498255a.html
6. Mills, S., Lucas, S., Irakliotis, L., Rappa, M., Carlson, T., & Perlowitz, B. (2012). *DEMYSTIFYING BIG DATA: a practical guide to transforming the business of Government*. Technical report. https://www-304.ibm.com/industries/publicsector/fileserve?contentid=239170
7. Nichols, W. (2013). Advertising Analytics 2.0. *Harvard Business Review*, *91*(3), 60-68.
8. Tableau (2016). Visual Analysis Best Practices: Simple Techniques for Making Every Data Visualization Useful and Beautiful. White paper. http://www.tableau.com/learn/whitepapers/tableau-visual-guidebook
9. Tansley, S., & Tolle, K. M. (Eds.). (2009). The fourth paradigm: data-intensive scientific discovery.

10. Tsou, M. H. (2015). Research challenges and opportunities in mapping social media and Big Data. *Cartography and Geographic Information Science*, 42:sup1, 70-74. doi: 10.1080/15230406.2015.1059251.

11. Tsou, M. H., Jung, C. T., Allen, C., Yang, J. A., Gawron, J. M., Spitzberg, B. H., & Han, S. (2015, July). Social media analytics and research test-bed (SMART dashboard). In Proceedings of the 2015 International Conference on Social Media & Society (p. 2). ACM. URL: http://dl.acm.org/citation.cfm?id=2789196

12. Tsou, Ming-Hsiang, Jiue-An Yang , Daniel Lusher , Su Han , Brian Spitzberg , Jean Mark Gawron , Dipak Gupta & Li An (2013) Mapping social activities and concepts with social media (Twitter) and web search engines (Yahoo and Bing): a case study in 2012 US Presidential Election, Cartography and Geographic Information Science, 40:4, 337-348, DOI: 10.1080/15230406.2013.799738

NOTE: If you are a student with a disability and believe you will need accommodations for this class, it is your responsibility to contact Student Disability Services at (619) 594-6473. To avoid any delay in the receipt of your accommodations, you should contact Student Disability Services as soon as possible. Please note that accommodations are not retroactive, and that I cannot provide accommodations based upon disability until I have received an accommodation letter from Student Disability Services. Your cooperation is appreciated.

WEEK		**LECTURE and WEB Exercises**		**Additional Resources (Reading)**
1	27 Aug 29	Introduction – What is Big Data? (from a Human Centered Perspective).	Ch. 1 Manyika	**No Lab this week.**
2	**03 Sep** 05	Big Data Collection and Process -1	Ch. 2 Lohr Tansley	Online Data Science Learning Resources and GitHub.
3	10 Sep 12	Big Data Collection and Process -2 (Sampling and Re-sampling, Data Biases and Noises, Data Filtering, Social Media APIs, and GeoJSON).		Introduction of R and R Studio
4	17 Sep 19	Common Technologies for Big Data Science and Analytics (HPC, Cloud Computing, EC2 and Google Compute Engine).	Mills	Introduction of Tableau 10 and some business data exercises
5	24 Sep 26	Software Packages (R, Tableau, and Pythons) and Database Management for Big Data Analytics (NoSQL, MongoDB, Hadoop and Spark).	Ch. 14 Tableau	Tableau 10 Exercises (continue)
6	01 Oct 03	Methods and Concepts in Data Science and Data Analytics – PART-1 (Statistical Inference and Machine Learning). **Introduce Group Project (Oct 03)**	Ch. 3 Nicoles	Installation of Amazon EC2 and MongoDB
7	08 Oct 10	Methods and Concepts in Data Science and Data Analytics – PART-2 (Social Network Analysis, Linguistic Analysis, and Noise Filtering). **Group Project Proposal Presentation (Oct 10).**	Ch. 4	Installation of Amazon EC2 and MongoDB

8	15 Oct 17 Oct	Human Dynamics and GIS Applications – Part-One (GIS and Geocoding Methods)	Ch.10 Tsou-10	**Creating a Group Project Website.**
9	**22 Oct** **24**	Distribute Midterm Exam Questions: **Oct 22** Human Dynamics and GIS Applications – Part-Two (Social Network Analysis)	Tsou-11 Tsou-12	MongoDB and NoSQL databases.
10	**29Oct** 31 Oct	**Mid-term Exam: Oct. 30 (20 points)** Human Dynamics and GIS Applications -Part-Three (Spatial Analysis Methods and Spatiotemporal Analysis)		MongoDB and NoSQL databases.
11	05 Nov 07	Big Data Case Studies in GPS data, Social Media, Epidemiology, and Cancer Data.	Ch. 12 Aslam Marx	Social Network Analysis and Gephi
12	12 Nov 14	Methods and Concepts in Visualization and Cartography – Part-One (Tools, Software, Colors and Symbols, Graphic Design).	Ch. 9	Introduction of ArcGIS Online and Story Maps
14	19 Nov 21 Nov	Methods and Concepts in Visualization and Cartography – Part-Two (User Centered-Design, 3D, Animation, and Virtual Reality).		Introduction of other Web Mapping Tools (CartoDB and MapBox).
13	**26 Nov** **28**	**Thanksgiving Week** **(NO Lectures, NO Labs)**		**NO LABS (SAL open on Tuesday, but no TA).**
15	03 Dec 05	Misconception and Miseuse in Big Data– Part 1 **(Graduate student assignment DUE: Dec 05)**	Ch. 11 danah	Group Project Time (How to use Camtasia for making videos)
16	**10 Dec**	Ethic issues and privacy concerns in Big Data	Ch. 13, 16	Group Project Time
	12 Dec	**Final Group Project Presentation Dec 12 (Thursday) (25 points) (SH 324) – as the Final Exam (12:30pm – 3:15pm). Please check your other class final exam schedules to make sure that you can attend this presentation (sign-in required).**		
	17 Dec	Email the Final Group Project Report to the Instructor (TSOU) by 5PM (**mtsou@sdsu.edu**)		

BDA/GEOG 594 Big Data Science and Analytics Platforms 2019 Fall
(25% of total course grade)

Big Data Analytic Research Group Project:

Two or three students will form an "Big Data Analytic Research Project Team". Each group will submit one page proposal on **October 10, 2019** and choose a possible project topic. Each team will select a team coordinator, who will coordinate the work progress of your project. The proposal will list the following items in a single page:

- The title of your project,
- Members' names,
- Coordinator's name,
- One paragraph to explain your project (200-300 words), and
- Weekly schedules and individual assignments.

Each team will spend five minutes to introduce their project proposal to the class on October 10, 2019.

Each team will give a brief group project progress report (two minutes) at the beginning of lecture on Thursday each week (after October 10).

At the end of semester, each team has to submit a "Big Data Analytic Research Project" in paper format and publish the result to group project web pages. **Each team will create a short 3 minutes video to introduce your group project.** The whole team members will present your project and video in front of the class as the final exam. **The final report presentation will be held in December 12 from 12:30pm - 3:15pm in SAL lab.** Each team has 3 minutes for video and 12 minutes for presentation and 5 minutes for questions. (If you need to use Powerpoint slides, save the slide in a USB drive or **upload to the Blackboard** before your presentation.) The contents of your presentation should follow your group report. (Everyone is required to attend the presentation classes and sign-up your name). **The final report (paper format) is due on December 17 (5pm) by email to the instructor's email address (mtsou@sdsu.edu).**

The Final report should include:

Group report (10-15 pages, double space, submit by each group) should include the following items:

- Team members
- Problem statement (why are you doing this project? why Internet mapping?)
- Literature review (other similar projects or fundamental theories – scientific journals or on-line resources)
- Database management and Data Process procedure (where do your data sets come from? Where do you put them on the Web and which analytic tools do you use?)
- Results (introduce your web design and published data analytics)
- Discussion

Link to the Group Project Video (3 minutes) using Camtasia Software.

Overview of your group project
Demo of the project website and analytic tools.
Promote your data analytic project and get more views!

Group Project Grading:

Final presentation 15%, Website Design 40%, Group project report 30%, Video 15%.