Weiai Xu (Wayne)
Assistant Professor
Department of Communication, UMass-Amherst
Email: weiaixu@umass.edu
Office: N334 ILC
Office hour: Tues. & Thur. 11 am-1 pm or by appointment.

COMM 497DB Survey/Digital Behavioral Data

Fall 2019
Tues. & Thur. 2:30PM - 3:45PM; ILC N345
Course website: curiositybits.cc/tutorial/

OVERVIEW

Algorithms and data increasingly power our private and civic life. Companies, nonprofits, and governments have invested heavily in data mining—the bulk collection of user behavior data from web platforms to understand public opinion and to forecast trends. A lot of fashionable terms, such as *artificial intelligence* and *big data*, are being thrown around these days. The public and regulators also become increasingly wary of the dark side of algorithms — the skepticism has culminated after the Cambridge Analytica scandal and the revelation of alleged foreign propaganda in the US through social media.

This course gives a practical understanding of how data mining and algorithms work. You can obtain **(1) marketable computational skills in data analytics and visualization** and **(2) evidence-based critical perspectives on the algorithmic society we live in**.

COURSE OBJECTIVES

Practical computational skills: you will learn to use R, a programming language widely used in data science. Data science is a booming industry. Glassdoor has rated data scientist the best job in the US in 2018 (www.businessinsider.com/best-jobs-in-america-2018-1#1-data-scientist-50). Understanding R will get you a foot in the door in the data science world.

Critical perspectives on algorithms and big data: Higher education is more than vocational training. On top of developing practical skills, I hope the course can help you shape informed perspectives on how algorithms and data may affect civil societies in good and bad ways. Such a perspective should be primarily based on evidence and your experience in working with data, not by ideologies or politically fashionable talking points.

COURSE STRUCTURE

Hands-on workshops: I have developed a series of interactive tutorials on using R for social data analytics. Before each workshop, you will complete a tutorial on your own. During the workshop, we will work to apply R codes in the tutorials to real-world cases.

Peer-to-peer learning: You will find a class partner. You two will discuss any error and difficulty that either of you has encountered. You will learn to debug code collaboratively.

In-class discussions: Discussion sessions occur at the latter part of the semester after you have acquired enough experience with data and R. You are required to contribute to in-class discussions.

LEARNING MATERIALS *- remember, the internet is your oyster*

The interactive tutorials for the class are published on course website curiositybits.cc/tutorial/. On the site, you can access older tutorials and student works produced from previous semesters. Do notice that the current tutorials have incorporated a considerable amount of updates as data science is an ever-evolving field, with new libraries and practices coming out every year.

The internet is your oyster. There are many great online resources about R. You may want to check out data science courses on DataCamp (www.datacamp.com/). You can follow *R-bloggers* (www.r-bloggers.com) to learn about the latest development in the R world. Stack Overflow (stackoverflow.com) is a great place to search and post questions about R (and all things related to data analytics and computer programming).

There is no required textbook. If you need to hold onto something to read, here are two books that might be useful: *Social Media Mining with R* by Richard Heimann, Nathan Danneman and *Humanities Data in R* by Taylor Arnold and Lauren Tilton.

BRING YOUR OWN DEVICES

Sadly, we do not have a computer lab equipped with R. While you can complete the interactive tutorials on virtually any device that has a browser, you must use a laptop/desktop to do the data science work. We will install and use two essential open-source software: *R* (www.r-project.org) and *RStudio* (www.rstudio.com). They run on Windows, Mac OS, and Linux systems (e.g., Ubuntu). They don't work with your iPads and Chromebook.

If your primary device is a Chromebook, try the cloud version of RStudio (rstudio.cloud/). It allows you to run R codes in the cloud within the Chrome browser.

IMPORTANT NOTES

This course does not assume prior knowledge of computer programming. In fact, I built this course particularly for students in humanities & social science majors as a gentle introduction to data science.

Be prepared to step out of your intellectual comfort zoom. While the course is a gentle introduction to data science, it can feel very technical and challenging to many of you. Five years ago, I took a baby step in learning coding and have enjoyed even the most frustrating moment of the journey. A programming language is just like a foreign language and the culture that comes with it. The more time you spend on it, the better you get at it.

Data science is 10% intelligence and 90% endurance. You will spend a lot of time testing and debugging R codes. The trial and error can be frustrating at times but is the essential part of learning. So, be patient with the learning. You learn the most not from attending lectures but from reverse-engineering and troubleshooting for data products.

Come to class prepared. Before coming to workshops, make sure to complete required tutorials and document errors. Before joining in-class discussions, make sure you have completed required readings. Preparation counts toward your grade.

Responsible use of computing power. At times in the semester, you will have access to my Twitter developer account in order to collect data. I will distribute the private account information to you via email. You have the obligation to keep the account information private in any circumstance. Our interactive tutorials run on a shared cloud server, which may become slow when receiving a lot of web traffic. To keep the server less busy for other students, please remember to close the browser tab after finishing a tutorial.

ABOUT ME

I am a faculty member in the department of Communication. I am specialized in computational communication research. You can track my projects and ideas on my website (curiositybits.cc) and Twitter @WeiaiWayne.

EVALUATION

Preparation and debugging (15%)

- Before each workshop, go to curiositybits.cc/tutorial to complete the tutorial required for the topic. Provide a screen capture showing each completed step;
- Try tutorial codes on your own device and document (screen capture or make a note) any error encountered — this is important because you will share the errors with your class partner;
- Upload screen captures from the previous two steps onto Moodle to receive preparation scores.

Assignments (35%)

The assignments help you practice R codes taught in workshops. In each assignment, you will complete a specific data mining and analytic task (e.g., collect and visualize someone's Twitter timeline). You will receive a specific assignment instruction for each assignment.

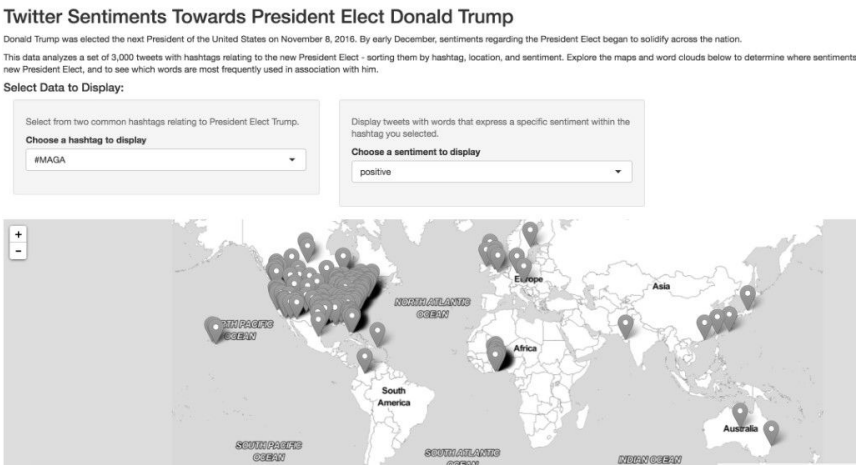
- Please budget enough time for each assignment as it will almost certainly require much more time than expected.
- You are expected to complete assignments independently. That said, you can consult with me or your peers if you encounter a problem. You can also Google to find a solution.
- Assignments should be submitted via Moodle.

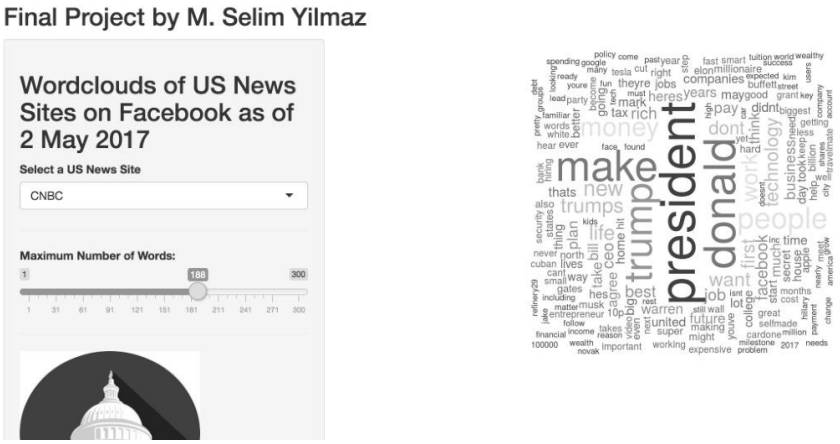
Final project (40%)

A cool feature of R is that you can build interactive visualizations and even interactive apps for the global audience. For the final project, you will build an interactive data visualization using one of the methods taught in class (i.e., trend analysis, sentiment analysis, network analysis, topic modeling, geo-mapping, etc). You can integrate the project into your portfolio for job search.

You are expected to complete the final project independently. You can, of course, consult with me or your peer to debug codes or to find solutions online.

You can make your visualization public by hosting it as an R Markdown file (rpubs.com/marschmi/RMarkdown) or as an R Shiny app (shiny.rstudio.com). An R Shiny app requires considerably more effort but will certainly impress your future employers. We will discuss the option when we get to the stage. Below are examples of some R Shiny apps created by students in previous semesters. You can find the live apps at curiositybits.cc/tutorial





*Some advanced hosting may require a small monthly payment for using a cloud server. I am happy to host your apps using my own server.

Participation (10%)

You will participate in discussions concerning the societal impact of algorithms and datafication. I will note your participation and grade based on your engagement (10%). I will not grade based on your opinion though — please refer to the *Commitment to Mutual Understanding, and Constructive Disagreement* section.

Grade	A	A-	B+	B	B-	C+	C	C-	D+	D	D-	F
Points	94 - 100	90 – 93	87 - 89	83 - 86	80 - 82	77 - 79	73 - 76	70 - 72	67 - 69	63 - 66	56 - 62	<=55

COURSE POLICIES

Attendance
Missing a class could result in significant difficulties in completing assignments and the final project. You are also expected to come to class fully prepared. That means completing required tutorials and readings.

Assignments due to absence can only be made up for legitimate reasons with documentation only. Legitimate reasons include, but are not limited to: absence while under the care of a health professional; absence due to a University-sanctioned event; absence due to your presence at a legal proceeding (documentation required); absence due to religious holiday; and absence due to family crisis, funeral, death, or serious illness. If you miss an assignment due to an excused absence, please contact me as soon as possible.

Deadlines
Meeting deadlines is an essential part of professionalism. So, I expect that all assignments to be submitted in time. I do offer reasonable exception if you encounter significant difficulties in completing the assignments. Contact me at least 2 days before the due date if you want an extended deadline.

Cell phones & non-essential laptop use
R won't run on your cell phones. Nor does your cell phone help you think well. Cell phones must be silenced and put away during class. Laptops are used only for class-related activities.

Course Schedule

This schedule is tentative.

9/3 (Tue.): Course introduction and a gentle introduction of social data analytics
- Complete the *Libraries/packages* tutorial before the class on 9/5

9/5 (Thur.): Set up R and RStudio, install libraries and explore R data frames (workshop)
- Complete the *Data frames* and *Connecting to the Twitter API* tutorials before the class on 9/10

9/10 (Tue.): Introducing Twitter APIs (workshop)
- Complete the *Collect tweets by keywords/hashtags* tutorial before the class on 9/12

9/12 (Thur.): Collect Twitter data part 1 (workshop) - assignment 1 announced
- Complete the *Collect Twitter user timeline* and *Collect Twitter user info* tutorials before the class on 9/17

9/17 (Tue.): Collect Twitter data part 2 (workshop)

9/19 (Thur.): Review & Catch up

- Complete the *Make Wordclouds* and *Visualizing virality* tutorials before the class on 9/24

9/24 (Tue.): Visualizing trends in Twitter data (workshop)

- Complete the *Sentiment analysis* tutorial before the class on 9/26

9/26 (Thur.): Sentiment analysis of tweets (workshop) - assignment 2 announced

10/1 (Tue.): Review & Catch up

- Complete the first three *Text Mining* tutorials before the class on 10/3

10/3 (Thur.): Text as data part 1, produce wordclouds and topic models (workshop)

- Complete the last two *Text Mining* tutorials before the class on 10/8

10/8 (Tue.): Text as data part 2, produce wordclouds and topic models (workshop)

10/10 (Thur.): Review & Catch up - assignment 3 announced

10/15 (Tue.): NO CLASS; MONEY SCHEDULE

10/17 (Thur.): Surveillance and privacy in the age of algorithms and datafication (discussion)

10/22 (Tue.): Advanced topics: detect impolite tweets using AI (workshop)

10/24 (Thur.): Advanced topics: predict Twitter users' ideology (workshop)

- Review the *Insights from networks* tutorial before the class on 10/29

10/29 (Tue.): Network as data: social network analysis (workshop)

10/31 (Thur.): Review & Catch up

11/5 (Tue.): Network as data: semantic network analysis (workshop) - assignment 4 announced

11/7 (Thur.): Review & Catch up

- Review the *Geo-mapping* tutorial before the class on 11/12

11/12 (Tue.): Geo-mapping Twitter users part 1 (workshop)

11/14 (Thur.): Geo-mapping Twitter users part 2 (workshop) - assignment 5 announced

11/19 (Tue.): Disinformation and propaganda in the age of algorithms and datafication (discussion)

11/21 (Thur.): R Shiny part 1- final project requirement announced

12/3 (Tue.): R Shiny part 2

12/5 (Thur.): Review & Catch up

12/10 (Tue.): Work on your final project

Academic Honesty Policy Statement

Since the integrity of the academic enterprise of any institution of higher education requires honesty in scholarship and research, academic honesty is required of all students at the University of Massachusetts Amherst. For more information about what constitutes academic dishonesty, please see the Dean of Students' website: http://umass.edu/dean_students/codeofconduct/acadhonesty/

Please note that in data science, it is customary to borrow and adapt computer codes produced by others. For instance, one may look for open-source on Github (github.com) and RPubs (rpubs.com) for inspiration. So feel free to use open-source codes for your work.

Disability Statement

The University of Massachusetts Amherst is committed to making reasonable, effective and appropriate accommodations to meet the needs of students with disabilities and help create a barrier-free campus. If you are in need of accommodation for a documented disability, register with Disability Services to have an accommodation letter sent to your faculty. Please initiate these services and to communicate with faculty ahead of time to manage accommodations in a timely manner. For more information, consult the Disability Services website at http://www.umass.edu/disability/.

Commitment to Mutual Understanding, and Constructive Disagreement

There are divergent views on the societal impact of algorithms and data, depending on the ideological lens. In order to create a classroom environment that supports respectful, critical inquiry through the free exchange of ideas, the following principles will guide our work (adapted from https://heterodoxacademy.org/teaching-heterodoxy-syllabus-language/):

- Treat every member of the class with respect, even if you disagree with their opinion;
- Bring light, not heat;
- Reasonable minds can differ on any number of perspectives, opinions, and conclusions;
- Because constructive disagreement sharpens thinking, deepens understanding, and reveals novel insights, it is not just encouraged, it is expected;
- No ideas are immune from scrutiny and debate;
- You will not be graded on your opinions.

Lastly, enjoy the ride!