Overview

INLS 490-89.

Human-centered data science policies and applications

Fall 2021

Course Description

This course aims to ground students in principles and practices for socially responsible data science. Students will work in teams to define and address a data-intensive problem in a domain of interest.

This course is open to upper-level undergraduate students and any graduate student.

Prerequisites

Permission of instructor, familiarity with basic statistical, programming, and teamwork practices. No specific courses are required but a statement of familiarity with these three skills and primary domain of interest must be presented to the instructor before registration.

Class Times, Location, and Format

Tuesdays, 8:00-10:45

Manning Hall 208

In person sessions, zoom access for viewing only.  https://unc.zoom.us/j/94456913673  Sakai website

Instructor

Dr. Gary Marchionini, Professor and Dean

He, him, his

Office: 100 Manning Hall

Office Hours: 10:30-11:30 Tuesdays and by appointment

Course Objectives

In this course, you will:

- Explain the data life cycle and provide an example for your primary field of interest.
- Identify and document the policies for a public data repository for your primary field of interest.
- Define and participate in a team project that focuses on one or more components of the data life cycle.
- Articulate sets of key applications, advantages, limitations, and biases common in large-scale data collection, analysis, interpretation, and reuse.

Grading
Graduate students will be graded on the H/P/L/F system. Undergraduate students will be graded on the A-F system. Learning activities and their weights are as follows:

- Term Project (40% of grade)
- Participation (30% of grade) This includes both in-class discussion and online posts
- Assignments (30% of grade)

H: The “H” grade is reserved for students whose work consistently goes above and beyond the stated expectations for a course or individual assignment. In this course, that might mean that you engage frequently and deeply in the class discussion forums (beyond the required posts); you ask and/or answer questions in the general discussion forum; your intermediate assignments are especially thorough; and your final proposal is exceptionally comprehensive and polished.
- P: This grade is earned for work which meets all established assignment and course requirements adequately. If you follow the guidelines for each assignment as they are shared with you on the syllabus and in class, you should expect to earn a P.
- L: This grade represents work that is substandard in at least one major way. If you are in danger of earning an L for the course, I will let you know as soon as possible so that you can improve your performance.
- F: Work that falls significantly short of expectations

The following grade scale will be used AS A GUIDELINE (subject to any curve) for undergraduate students: Grade Range Definition* A 90-100% Mastery of course content at the highest level of attainment that can reasonably be expected of students at a given stage of development. The A grade states clearly that the students have shown such outstanding promise in the aspect of the discipline under study that he/she may be strongly encouraged to continue. B 80-89.9% Strong performance demonstrating a high level of attainment for a student at a given stage of development. The B grade states that the student has shown solid promise in the aspect of the discipline under study. C 70-79.9% A totally acceptable performance demonstrating an adequate level of attainment for a student at a given stage of development. The C grade states that, while not yet showing unusual promise, the student may continue to study in the discipline with reasonable hope of intellectual development. D 60-69.9% A marginal performance in the required exercises demonstrating a minimal passing level of attainment. A student has given no evidence of prospective growth in the discipline; an accumulation of D grades should be taken to mean that the student would be well advised not to continue in the academic field. F 0-59.9% For whatever reason, an unacceptable performance. The F grade indicates that the student’s performance in the required exercises has revealed almost no understanding of the course content. A grade of F should warrant an advisor’s questioning whether the student may suitably register for further study in the discipline before remedial work is undertaken. * Definitions are from: [http://registrar.unc.edu/academic-services/grades/explanation-of-grading-system/](http://registrar.unc.edu/academic-services/grades/explanation-of-grading-system/)

The University Honor System

The University of North Carolina at Chapel Hill has had a student-administered honor system and judicial system for over 100 years. Because academic honesty and the development and nurturing of trust and trustworthiness are important to all of us as individuals, and are encouraged and promoted by the honor system, this is a most significant University tradition. More information is available at [http://www.unc.edu/depts/honor/honor.html](http://www.unc.edu/depts/honor/honor.html). The system is the responsibility of students and is regulated and governed by them, but faculty share the responsibility and readily commit to its ideals. If students in this class have questions about their responsibility under the honor code, please bring them to me or consult with the Office of the Dean of Students. The web site identified above contains all policies and procedures pertaining to the student honor system. We encourage your full participation and observance of this important aspect of the University.

Students with Disabilities

The University of North Carolina at Chapel Hill facilitates the implementation of reasonable accommodations, including resources and services, for students with disabilities, chronic medical conditions, a temporary disability or pregnancy complications resulting in difficulties with accessing learning opportunities. All accommodations are coordinated through
the Accessibility Resources and Service Office. See the ARS Website for contact information: [https://ars.unc.edu](https://ars.unc.edu) or email ars@unc.edu.

**SILS Diversity Statement**

In support of the University’s diversity goals and the mission of the UNC School of Information and Library Science, SILS embraces diversity as an ethical and societal value. We broadly define diversity to include ability, age, ethnicity, gender, gender identity, gender expression, immigration status, national origin, race, religion, sexual orientation, and socioeconomic status. As an academic community committed to preparing our graduates to be leaders in an increasingly multicultural and global society we strive to:

- Ensure inclusive leadership, policies, and practices
- Integrate diversity into the curriculum and research
- Foster a mutually respectful intellectual environment in which diverse perspectives and experiences are valued
- Recruit and retain students, faculty, and staff from traditionally underrepresented groups
- Participate in outreach to underserved groups in North Carolina and beyond.

The statement is our commitment to the ongoing cultivation of an academic environment that is open, representative, and reflective of the concepts of equity and fairness. *The Faculty and Staff of the UNC School of Information and Library Science*

**Counseling and Psychological Services (CAPS)**

CAPS is strongly committed to addressing the mental health needs of a diverse student body through timely access to consultation and connection to clinically appropriate services, whether for short or long-term needs. Go to their website: [https://caps.unc.edu/](https://caps.unc.edu/) or visit their facilities on the third floor of the Campus Health Services building for a walk-in evaluation to learn more.

**Title IX Resources**

Any student who is impacted by discrimination, harassment, interpersonal (relationship) violence, sexual violence, sexual exploitation, or stalking is encouraged to seek resources on campus or in the community. Please contact the Director of Title IX Compliance (Adrienne Allison – Adrienne.allison@unc.edu), Report and Response Coordinators in the Equal Opportunity and Compliance Office (reportandresponse@unc.edu), Counseling and Psychological Services (confidential), or the Gender Violence Services Coordinators (gvscl@unc.edu; confidential) to discuss your specific needs. Additional resources are available at [http://safe.unc.edu](http://safe.unc.edu)

**Course Schedule**

**Course Outline**

**Topic 1.** Introduction: Data Science as an emergent field

August 24

**Key concepts**

- The emergence of data science: Science and commerce drivers; the Vs (Volume, Variety, Velocity, Variability, Veracity, Visualization, and Value.
- The data-information-knowledge pathway
Data science is team science
Technology and Knowledge neutrality arguments from philosophy, law, and politics perspectives
Data jobs, roles, and opportunities
Professional codes of conduct

Activities

Post these two examples with pointers (e.g., URL) to the class Sakai site.

a. Find one example of how large-scale data have driven human progress.
b. Find one example of how large-scale data have harmed individuals, groups, or society.

Read/View

Watch video on https://www.data.org/

Scan proposals for technology for public good https://www.dayoneproject.org/policy-proposals


https://hbr.org/2007/11/eight-ways-to-build-collaborative-teams More focused on corporate, large teams, but some useful guidance for understanding corporate and large organization actions that may affect your success

Topic 2. Teamwork, collaboration, and the changing nature of work

August 31

Key Concepts

Collaboration, cooperation, and teamwork pros and cons

Jobs, roles, and titles for data science

Activities

Pick 6 of the following job titles and go to http://monster.com and http://indeed.com and find how many positions are available in the US and in NC. Post to Sakai.

Business Analyst, Database Engineer, Data Analyst, Data Steward, Data Engineer, Data Scientist, Research Scientist, Software Engineer, Statistician, Product Manager, Project Manager, Analytic developer, Analytic production steward, Chief data officer, Collection steward, Data architect, Data custodian, Data engineer, Data modeler

Read/View

https://www.go-fair.org/fair-principles/

Data curation and the emerging field of data science. 21 minutes https://web.microsoftstream.com/video/ea0aa868-6fa2-4e2a-a9ee-dc7af1828da2

Scan: https://www.dataone.org/


**Topic 3.** The Data Life Cycle and FAIR Principles

September 7

**Key Concepts**

Data Life Cycle Models

Importance of the full life cycle for effective data management and use

Requirements for FAIR: Findable, accessible, interoperable, and reusable

**Activities**

For class discussion:

Give examples of how you have used DLC in making a decision (e.g., where to go to college, which phone to buy)

Do a web search for ‘data life cycle.’ What kinds of variations do you find? What are the tradeoffs for a circular vs linear layout? What about a word cloud?

**Read/View**

Read https://research.aimultiple.com/data-cleaning/ Practical posting in Cem Dilmegani’s AlMultiple blog.


Explore: http://Onthebooks.lib.unc.edu

Optional:

L. Arbuckle and F. Ritchie, "The Five Safes of Risk-Based Anonymization" in IEEE Security & Privacy, vol. 17, no. 05, pp. 84-89, 2019. doi: 10.1109/MSEC.2019.2929282 keywords: {data privacy;law;computer security;risk management;data analysis;ethics} url: https://doi.ieeecomputersociety.org/10.1109/MSEC.2019.2929282

Participate in SILS 90th symposium on AI and Knowledge Work (Sept 10)

**Topic 4.** Problem driven data collection and cleaning

September 14

**Key Concepts**

Why collect data? What data to collect? Problem ontology

Data cleaning

Documenting work flows

Bias in data (statistical, cultural, cognitive)
Activities

Write the rationale for an IRB that specifies a data collection and data management plan

Read/View

Explore:

https://research.unc.edu/human-research-ethics/consent-forms/

https://research.unc.edu/human-research-ethics/

Read: MIT Technology Review article: https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/


Optional:


[a recent meta-analysis of data quality factors in biomedical repositories]

Topic 5. Data quality, Informed consent, and responsible outcomes

September 21

Key Concepts

Attributes of data quality

Strategies and metrics

Informed consent for data on humans

Activities

Outline a set of steps you and your team will take to document data quality for each step of the data life cycle.

Read/View

Explore: https://odum.unc.edu/survey-research/

https://www.pewresearch.org/methods/u-s-survey-research/questionnaire-design/

Topic 6. Surveys, transaction logging and screen scraping
September 28

Key Concepts
Problem ontology and mapping to stimuli
Kinds of surveys
Format of stimuli
Skip patterns
Tools: Qualtrics, Mechanical Turk, Survey Monkey

Activities
If your term project will use a survey to collect data, define your question set and begin pilot testing with other groups in the class

Read/View
Read: https://www.nature.com/sdata/policies/repositories

Topic 7. Public repositories
October 5

Key Concepts
Public repository mission and governance
Open repositories (see https://or2021.openrepositories.org/)

Policies of access, contribution, editing, and attribution

Activities
What are the key public repositories in your area of interest?
What are the policies for acquisition, editing, accessing, acknowledging?

Read/View
The digital representation blur 14 minutes https://web.microsoftstream.com/video/65898160-210e-4358-a85b-efdceab119b8
Indexing 37 minutes https://web.microsoftstream.com/video/99a72fbc-b397-4964-acdb-fa0bdf175679

October 12: University Day. No class meeting

Topic 8. Metadata, indexes and indexing theory and practice
October 19

Key Concepts
Metadata for machines and surrogates for humans

Indexing principles and techniques

**Activities**

Identify metadata you will use for your team project and how it will be created or harvested

**Read/View**

Links TBD (guest speaker)

**Topic 9.** Tool and Toolkits

October 26

**Key Concepts**

Tool kit suites and stacks for different phases of the data life cycle

Guest Professor Arcot Rajasekar

**Activities**

Post a structured description for a tool or tool kit of interest

**Read/View**


**Topic 10.** Curation, Governance, and Preservation

November 2

**Key Concepts**

Data professional titles, roles and responsibilities

Curation as added value process

Data governance

Preservation principles and techniques

Guest Dr. Michael Barker

**Activities**

Go to [http://monster.com](http://monster.com) and [http://indeed.com](http://indeed.com) and enter terms: data officer, data manager, data steward. How many jobs for each? How many in North Carolina? What kinds of salaries?

Report (verbally) in class project update

**Read/View**

TBD (guest)
**Topic 11.** Communicating Results

November 9

**Key Concepts**

Visualizations

User interfaces

Technical reports and video messaging

**Activities**

Post a pointer to an effective data visualization

**Read/View**

TBD (guest speaker)

**Topic 12.** Data systems: end to end

November 16

**Key Concepts**

Data storage

Data workflows

Data sharing

Data Security

Data privacy

**Activities**

Prepare for project presentations

Project Presentations Part 1

November 23

Project Presentations Part 2

November 30

Final Written Projects Due Dec 7

**Resources**

INLS 490-89 Resources: Fall 2021

**Publications/Reports**


Nature. https://www.nature.com/sdata/policies/repositories


Websites

https://www.data.org/

https://www.dayoneproject.org/policy-Proposals

Data science and humanities: https://www.turing.ac.uk/research/interest-groups/humanities-and-data-science

https://datacarpentry.org/ Online, free lessons for fundamental data skills in different research areas.

https://www.go-fair.org/fair-principles/

https://bdtechtalks.com/2020/07/15/machine-learning-adversarial-examples/ (easy read, overview of high profile cases of image recognition errors)
https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/

https://cacm.acm.org/magazines/2018/7/229030-making-machine-learning-robust-against-adversarial-inputs/fulltext?mobile=false (more detailed explanation of some of the cases and potential solutions)

https://nnlm.gov/data/thesaurus/data-lifecycle


https://www.thebalancecareers.com/tips-for-better-teamwork-1919225  Tips on better teamwork. Easy read, practical

http://Onthebooks.lib.unc.edu

https://www.pbs.org/independentlens/documentaries/coded-bias/


https://research.unc.edu/human-research-ethics/consent-forms/

https://research.unc.edu/human-research-ethics/

https://odum.unc.edu/survey-research/

https://www.pewresearch.org/methods/u-s-survey-research/questionnaire-design/

https://or2021.openrepositories.org/


TED Talks

Zeynep Tufekci

**Machine intelligence makes human morals more important**

Posted Oct 2016

Sharon Weinberger

**Inside the massive (and unregulated) world of surveillance tech**

Posted Dec 2020

Audrey Tang

**How digital innovation can fight pandemics and strengthen democracy**

Posted Jun 2020

Natsai Audrey Chieza

**Possible futures from the intersection of nature, tech and society**

Posted Mar 2021