Teaching Statement

Cody Buntain
Assistant Professor
College of Information Studies, University of Maryland

March 14, 2024

Contents

1 Teaching Philosophy and Innovations

1.1 Teaching Philosophy

As students passing through my classes are likely to become data scientists, software developers, and technologists at companies like Meta, Google, or defense contractors, I see it as crucial to empower these students with the tools necessary to reflect on, assess, and understand the societal impacts of their work. I endeavor to teach my students to forefront stakeholder needs and interests rather than focusing on what data or models are easily available. As a consequence, much of my teaching and assessment is focused not on evaluating for perfection but on stimulating exploration, collaboration, and self-guided evaluation. By presenting engaging and unconventional scenarios that cross disciplines and contexts, my goal is to foster individuals who are comfortable with and value working with stakeholders, regardless of their technical capabilities.

Individual aspects of my approach to teaching fall along the following areas:

Let the Student Lead. While I construct my lectures to be humorous and engaging, I find allowing the students to explore a problem via a collaborative exercise, group discussion, or other activity often leads to more interesting and memorable exchanges. Where possible, my classes shift students' focus to each other and introduce topics or solutions via activities or prompt a lecture through student presentations, debates, game-style activities, or other interactive event. In many in-person classes now, for example, I use a deck of cards with students' names as a method to get students to brainstorm for the class on how we might solve a problem encountered in lecture. This aspect is meant to foster student-lead exploration of a topic, allowing them to engage with and justify the "why" before explaining the "how".

Bridge the Disciplines. A great deal of progress has been made in individual fields of study, but many motivating problems in information sciences exist at the intersection of many fields. Social science disciplines come with a wealth of problem areas and research questions that we can appropriate, from mundane daily life or larger societal issues, leading to the burgeoning field of computational social science. These problems are often well-motivated, making them good questions to pose to my students, as they get experience applying computational tools to diverse questions. By encouraging students to engage with stakeholders from outside their disciplines, this aspect is meant to socialize the importance of collaboration among my students.

Explore the Unconventional. Well-motivated and relatable questions from the social sciences are a good starting point, but the most impactful strategy I have employed is to find surprising connections between unexpected contexts. I have had students build analog networks using strings, post-it notes, and dice to mimic the Internet to take the technology out of networking principles. To teach object-oriented design, I had students play boardgames and replicate them via modeling; boardgames already come with well-defined rules, allowing students to focus first on the modeling task. These activities often allow students to engage with the material in a more tactile and visceral way, leading them to important insights. This approach has lead to impactful moments, with one student exclaiming in one of my classes on communication networks, "How does the Internet ever successfully get packets across the globe?!" while discussing error-correction.

Concretize the Course's Value with Useful Artifacts. While collegiate courses have intrinsic value in their educational aspects, it can be difficult to motivate students–especially undergraduates–to engage with the course material, as many courses are seen simply as degree requirements. Hence, a core part of my philosophy is to ensure each course has a clear and concrete set of artifacts that have value to students outside the classroom. In recent courses, I have converted homework assignments into such artifacts by requiring students to write publicly accessible articles on the Medium blogging platform to demonstrate mastery, both in the technical aspects of the work and in communicating decisions and their justifications.

1.2 Experiences in Teaching

Beyond my teaching at UMD, my experience spans multiple institutions, as I have developed and lead four courses as an adjunct professor at American University (AU), and adjunct at UMD following my PhD, and at the Informatics Department at the New Jersey Institute of Technology (NJIT). From AU and NJIT, I have experience in developing multiple courses, from introductory and advanced programming, design, Internet architectures, technology ethics, social media analysis, and crisis informatics. This multi-institution experience has exposed me to a variety of student needs and capabilities, as the student bodies at each institution differ substantially. This variety spans more affluent students in private institutions at AU to many first-generation college students at NJIT to a wide variety of students at Maryland's flagship institution. These variations have provided a grounding for my approach to teaching, as the wide variety of students' trajectories and backgrounds can be a challenge, especially for technically oriented courses. As such, I endeavor to be flexible in teaching and make efforts to pivot, slow, or accelerate courses according to student needs, often including additional exercises to stimulate the more advanced students without leaving less experienced students behind.

1.3 Self-Reflection on Teaching Successes and Failures

Despite my teaching experience, I am well-aware that my approaches to pedagogy and education should not be static, especially as the student population evolves and changes. To this end, I work to maintain an open mind and accept that aspects of my course development may become outdated or outright fail, and these failures present key opportunities to refine my courses. To this end, I have a framework for assessing and improving my coursework:

Internal Course Reflections As I teach courses, I maintain separate archives for each semester I have taught the class, which includes a running notes document about assignments, lecture, and activities that worked or failed. These notes have allowed me to refine courses across semesters and institutions. For example, in revising INST414 between spring 2022 and spring 2023, it became clear that the order of my material was problematic for students, and I reordered the material accordingly, moving material on data hygiene earlier in the semester. Likewise, in preparation for the spring 2024 semester, I further refined the INST414 curriculum to provide more time for in-class coding exercises with live-coding solutions and restructured the schedule to allow students more time for focus on externally facing projects.

Mid-Point Anonymous Surveys While teaching evaluations can be useful, by the time such evaluations are available, the course has already ended. As such, it is too late to adjust the course and make it more useful for current students. To address this concern and allow flexibility in courses as I run them, I publish an anonymous survey for students to take at the mid-point of every semester. This survey includes questions about activities and assignments that worked or failed, which aspects the students would like to see more or less

in the course, and what aspects of the course the students found particularly impactful. In INST414 in spring 2022, this survey identified several shortcomings students wanted addressed, primarily around providing additional programming exercises.

Integrating Approaches from External Faculty I recognize that many courses I teach are either long-standing courses at UMD or available at other institutions. Other faculty's teaching of the material I intend to cover is an invaluable resource as we develop new or modify existing courses. My use of the Medium blogging platform is a direct result of this integration, as Prof. Brian Keegan at UC-Boulder has had positive experiences in the same for his class on "information exposition". At the same time, students' needs across institutions and years are ever evolving, and I work to ensure my modifications of others' material meets the needs of our students.

2 Mentoring and Research Supervision

2.1 Mentorship

Beyond teaching, I have mentored numerous students of high school, undergraduate, masters, and doctoral levels at UMD, NJIT, and elsewhere. Since returning to UMD, this mentorship includes multiple doctoral, masters, and undergraduate students as well as a senior high-school student who has since gone on to matriculate at Princeton University. In addition, I have served as a mentor for the annual ARLIS Research for Intelligence and Security Challenges (RISC) internship, starting in the summer of 2022 and continuing in 2023.

I divide my mentoring activities along two lines:

Technical Mentorship Following from my time at NJIT, where many students have a more technical focus,I have a separate approach to mentorship for engineering- and development-oriented work. Students I mentor in this capacity often have shorter availability (one to two years) and a narrower scope of interests, as these students are often in masters programs. The collaborations focus more on technology development than research, experimentation, and understanding social processes, as technology and software development are more central to these students' programs. Consequently, my philosophy on technical mentoring engagements is to prioritize ethical approaches to development, emphasizing the work's motivation and constructing minimum viable products that satisfy these motivations. Students with whom I have engaged in this mentorship generally go to industry or government careers–at least three of whom have become data scientists at Pacific Northwest National Labs (PNNL), Walmart, and Teradata.

Research Mentorship In contrast to more technically and engineering-focused mentorships, my engagement with research-oriented students–often PhD students, but occasionally aspiring masters and undergraduate students–instead follows an apprenticeship model. My approach to this model integrates research-oriented advisees into existing research projects, often offloading small aspects of the project to them, and engaging them in the literature and motivations for these small tasks. At the same time, I push these students to engage in the broader scientific community, first through reading the literature and eventually through inclusion in the publication process around the larger project. My overarching goal is to get students familiar enough with specific research communities so they can identify contradictions or open questions in the literature and, crucially, understand how to frame and make contributions in these fields.

By fostering this familiarity with and capacity to assess contributions to our academic communities, I aim to establish my research-oriented advisees as independent researchers in their own right, armed with the tools to continue in a research-oriented role beyond UMD. To this end, as my students become more senior, I encourage them to pursue a parallel but related research effort on their own, to include collaborations with researchers outside the iSchool and UMD.

2.2 Supervised Research Assistants

Below lists the students for whom I have had a significant level of mentorship. Many of them are engaged on my Young Faculty Award project, Maps, Models, and Metrics for Influence in Conflict Zones (M3I).

2.2.1 Graduate Students

Student	Project	Year	Status	Publications
Do Won Kim	QAnon and Save-the-Children Voter Mobilization in 2020	2023–Present	PhD Student, UMD	–
Matthew Feehan	Vulnerability in Ideology and Moral Foundations	2023–Present	PhD Student, UMD	–
Maharshi Gor	M3I	2023–Present	PhD Student, UMD	–
Marilyn Halbert	M3I	2024–Present	PhD Student, UMD	–
Keng-Chi Chang	Visual Media in Online Influence and Engagement	2022–Present	PhD Candidate, UCSD	[w14]
Yoo Yeon Sung	Identifying Misleading Headlines in Social Media Video Posts	2022–Present	PhD Candidate, UMD	–
Md Main Uddin Rony	Misinforming Headlines	2023–Present	PhD Candidate, UMD	–
Utkarsh Tiyagi	M3I	2023–Present	MS, UMD	–
Purva Chiniya	M3I	2023–Present	MS, UMD	–
Ruizhe “Richard” Fan	Network Embeddings of Web-Domain for Ideology Prediction	2021–2023	PhD Advisee	[o9]
Celia Chen	M3I DARPA Project	2022–2023	PhD Advisee	–
Girish Saraf	Crisis Informatics and Information Retrieval	2022	MS Advisee	–
Shivan Sharma	Crisis Informatics Project	2019–2021	Scientist, PNNL	[c21], [c22]

2.2.2 Undergraduate Students

Student	Project	Year	Status	Publications
Colin Clifford	M3I – Maps, Models, and Metrics for Influence	2024–Present	Undergraduate, UMD	–
Arvind Jayaraman	M3I	2024–Present	Undergraduate, UMD	–
Zeyu Li	M3I	2023–Present	Undergraduate, UMD	–
Ekenedilichukwu “Glory” Ndu	PPI - Local Health Information-scapes	2023–Present	Undergraduate, UMD	–
Robert Zhang	Intersections of Chinese Influence and Public Diplomacy	2023–Present	Undergraduate, UMD	–
Amogh Joshi	Image-Based Ideology Inference	2021–Present	Undergraduate, Princeton	[w13], [c29]

3 Sample Syllabus and Student Assessments

3.1 UNUH248O: We the (Artificial) People: How AI Has Reshaped Politics

3.1.1 Syllabus

UNUH248O: We the (Artificial) People – How AI has Reshaped Politics

Course Syllabus, v0.1, 15 February 2024

Fall 2024

Instructors

Instructor: Prof. Cody Buntain

Instructor Email: cbuntain@umd.edu

Instructor Office Location: HBK2117B

Course Description

Artificial intelligence (AI) has had profound impacts on the modern political landscape, in the US and abroad. This course encourages the critical evaluation of how AI has impacted political behavior, altered the public's relationships to politicians (for good or ill), and opened new threats like foreign electoral inference, disinformation, and manipulation through deep-fakes and generative language models. Students will debate ethical, fair, transparent, and accountable AI.

Required Background

No prior background is required, as this class will be taught at the introductory level. Familiarity with the US political system or US civics would be valuable but is not necessary.

Student Learning Outcomes

At the end of this course, students will able to do the following:

1. Describe the tensions between an innovation society built on strong protections of privacy and intellectual property alongside technology that relies on large amounts of (potentially sensitive or copyrighted) data to create new and innovative material
2. Explain how AI systems learn from, influence, and impact your information diet, and how these changes can potentially impact political knowledge and behavior
3. Identify and describe strategies that actors may use to leverage, influence, or exploit AI systems for political, economic, or personal gain
4. Identify multiple areas where bias in AI systems or their training data can lead to unequal distributions of harm
5. Describe governance and regulation efforts around artificial intelligence to protect against threats (both real and perceived) to national security, privacy, health, equity, and other areas

6. Develop multiple arguments to prioritize different sensitive aspects of harm reduction in the context of AI regulation, describing potential regulatory interventions to reduce these harms

Things we won't teach– Students should also know that this course will NOT teach:

- How to code AI algorithms
- How to estimate data models
- The mathematical formalisms behind AI or machine learning algorithms
- Nitty-gritty understanding of the kinds of specific algorithmic classes out there (e.g., model architectures in convolutional neural networks, etc.).

Textbooks

While we mostly will be engaging with readings from academic venues and popular press, we will include readings from the following textbooks:

A Political Science Experiment *(APSE)* – We will read several chapters from this book.

Course Modules

Module	Topic
Module 1	Artificial Intelligence and Machine Learning: A Primer
Module 2	AI and Feedback Loops in Your Online Information Diet
Module 3	Media and Political Communication
Module 4	The Effects of Political Information on Citizens
Module 5	Political Participation: Elections and Protests
Module 6	International Relations and Military Uses of AI
Module 7	Regulating Artificial Intelligence Systems

Grade Distribution

Grades for this class are broken down as follows:

Semester Project Students will complete an independent project, exercising the skills and engaging with the topics learned. These projects can be individual or group-based and will culminate with a final report and presentation at the end of the semester. – 25%

In-Class Presentations and Panels Students will give at least two in-class presentations (one pre-midterm, and one post-midterm), outlining one of the assigned readings, highlighting what they feel are the most important aspects of the reading. Students will then sit on a panel with other students presenting that day and answer questions from the instructor and student audience. – 25%

Response Papers Students will write at least two papers recording their responses or reactions to a student's presentation. These reports will include two things: summarize the material and detail the student's reaction to it. – 25%

Class Participation Asking questions, participating in discussion – 25%

Letter Grade Cutoffs

A+	97-100*	B+	87-89.99	C+	77-79.99	D+	67-69.99
A	93-96.99	B	83-86.99	C	73-76.99	D	63-66.99
A-	90-92.99	B-	80-82.99	C-	70-72.99	D-	60-62.99

*Note: To receive an A+ you must have demonstrated significant contributions to the class in addition to achieving this numeric grade. We reserve the right to curve grades upward (but will not curve grades downward).

Policy on Generative AI

The use of generative AI tools such as ChatGPT *is* allowed for all assignments in this class. However, a central goal of the class is to help you become independent and critical thinkers, so we discourage you from the extensive use of generative AI tools as a substitute for your developing your own opinions and ideas. If you do use AI-generated content in your assignments, you must clearly indicate what work is yours and what part is AI-generated through proper attribution. We also ask you provide a short one-paragraph summary at the end of the assignment on how you used AI tools. Please consult this APA post on how to cite AI tools. Failure to do so will be considered plagiarism according to UMD's Academic Integrity policies.

Syllabus Change Policy

Once the semester begins, this syllabus will be revised infrequently, and any revised requirements will be posted as announcements and updated course schedule to ELMS. The instructor reserves the right to make changes to the course's schedule, evaluation criteria, policies, etc. through announcements in class and on ELMS, so please check ELMS regularly. Students should email the instructor if there are any discrepancies or questions.

Campus Policies

It is our shared responsibility to know and abide by the University of Maryland's policies that relate to all courses, which include topics like:

- Academic integrity
- Student and instructor conduct
- Accessibility and accommodations
- Attendance and excused absences
- Grades and appeals
- Copyright and intellectual property

Please visit go.umd.edu/ug-policy for the Office of Undergraduate Studies' full list of campus-wide policies and follow up with me if you have questions.

Additional Accommodation Policy

I understand the difficulty and additional constraints you may be facing during this time. I am willing to work with you to discuss possible accommodation and alternative arrangement. Please do not hesitate to contact me when needed.

Accessibility and Learning Support

Students with disabilities should inform me of their needs at the beginning of the semester. Please also contact the Accessibility and Disability Support Office (http://www.counseling.umd.edu/ADS/). ADS will make arrangements with the student and me to determine and implement appropriate academic accommodations. Inclusion is one of the iSchool's core values, and we have attempted to make all

materials and assignments accessible to people with varying abilities. However, if there is something else I can do to make the class more accessible, please schedule a time to come talk to me. These improvements will benefit not only yourself but also future students!

Get Some Help!

Taking personal responsibility for your own learning means acknowledging when your performance does not match your goals and doing something about it. I hope you will come talk to me so that I can help you find the right approach to success in this course, and I encourage you to visit tutoring.umd.edu to learn more about the wide range of campus resources available to you.

In particular, everyone can use some help in sharpening their communication skills (and improving their grade) by visiting ter.ps/writing and scheduling an appointment with the campus Writing Center.

You should also know there are a wide range of resources to support you with whatever you might need (see go.umd.edu/assistance). If you just need someone with whom you can talk, visit counseling.umd.edu or one of the many other resources on campus. Most services are free because you have already paid for them, and everyone needs help... all you have to do is be brave enough to ask for it.

3.2 INST414: Data Science Techniques

3.2.1 Syllabus

INST414: Data Science Techniques

Course Syllabus, v1.0, 25 January 2023

Spring 2023

Instructors

Instructor: Prof. Cody Buntain

Instructor Email: cbuntain@umd.edu

Instructor Office Hours: M/W, 12:30-13:30 PM, and by appointment

Instructional TA: Mofe Barrow

Instructional TA Office Hours: TBD

Grader TA: Madhulika Nambi

TA Email List: inst414-cb-tas@umd.edu

Meeting Time and Spaces

Classroom: In-person in Hornbake Library, HBK 0302H.

Meeting Times: M/W 11:00 AM-12:15 PM

First Day of Class: 25 January 2023

Last Day of Class: 10 May 2023

Course Description

This course explores the application of data science techniques to unstructured, real-world datasets including social media and open data sources. The course will focus on techniques and approaches that allow the extraction of information relevant for experts and non-experts in a wide range of areas including smart cities, transportation or public safety. This course will explore approaches to extract insights from large-scale datasets. The course will cover the complete analytical funnel from data extraction and cleaning to data analysis and insights, interpretation, and visualization. The data analysis component will focus on techniques in both supervised and unsupervised learning to extract information from datasets. Topics will include clustering, classification, and regression techniques. Through homework assignments, a project, exams and in-class activities, students will practice working with these techniques and tools to extract relevant information from structured and unstructured data.

Required Background

Prerequisite Courses: 1 course with a minimum grade of C- from (INST201, INST301); and minimum grade of C- in INST126, INST314, STAT100, MATH115, and PSYC100.

Students are expected to have prior experience with and competency in computer programming. Proficiency in the Python language is preferred but not essential. Course assignments will primarily be written in Python and built on the Jupyter notebook framework, which does not come standard on most platforms and is often installed via the command line, so a familiarity with console applications is also preferred.

Student Learning Outcomes

This course's main goal is to expose students to the collection and analysis of large, web-scale data, collected from online sources, with the goal of extracting actionable insights through applied exercises/hands-on projects. Over this course, students will:

1. Collect and clean large-scale datasets.
2. Articulate the math behind supervised and unsupervised techniques.
3. Execute supervised and unsupervised machine learning techniques.
4. Select and evaluate various types of machine learning techniques.
5. Explain the results coming out of the models.
6. Critically evaluate the accuracy of different algorithms and the appropriateness of a given approach

Textbooks

Textbooks below provide useful background and reference material. They are freely available for UMD students as well.

Introduction to Machine Learning with Python : A Guide for Data Scientists (IMLP) by Andreas C. Mäller and Sarah Guido, ebook available at UMD library

Python Data Science Handbook: Essential Tools for Working with Data by Jake VanderPlas, available at: https://jakevdp.github.io/PythonDataScienceHandbook/

Software

Jupyter notebooks written in Python 3 will be used for all in-class examples and assignments. The Anaconda distribution of Python 3 is strongly recommended to provide all of these programs and other libraries. If students wish to use an alternative data analysis environment (R, Matlab, Julia, etc.) they are welcome to do so, but instructional support is only guaranteed for Python.

Jupyter also provides a ready-made Docker container for data science-style notebooks, available here: https://jupyter-docker-stacks.readthedocs.io/

Course Modules

Module	Topic
Module 1	Data Science and Motivations
Module 2	Web Data as Graphs
Module 3	Similarity, Dimensionality Reduction, and Cleaning
Module 4	Clustering and Unsupervised Learning
Module 5	Probability and Bayes' Theorem
Module 6	Supervised Machine Learning
Module 7	Evaluating Your Models

Grade Distribution

Grades for this class are broken down as follows:

Module Assignments Students will complete independent projects for each module, exercising the skills learned. These assignments will be submitted in the form of Medium posts, which the professor will aggregate for the class – 35%

In-Class Labs This course includes sections for students to work collectively on lab assignments in class, to practice skills from each module. These lab periods occur weekly, with output submitted via ELMS. – 25%

Final Project Over the semester, students will develop a project that integrates skills over the semester, applies them to a specific problem, and present a final report on this project – 30%

Participation Asking questions, participating in online discussion – 10%

Letter Grade Cutoffs

A+	97-100*	B+	87-89.99	C+	77-79.99	D+	67-69.99
A	93-96.99	B	83-86.99	C	73-76.99	D	63-66.99
A-	90-92.99	B-	80-82.99	C-	70-72.99	D-	60-62.99

*Note: To receive an A+ you must have demonstrated significant contributions to the class in addition to achieving this numeric grade. We reserve the right to curve grades upward (but will not curve grades downward).

Module-Level Assignments

Each module has a corresponding assignment intended to develop students' skill and confidence in using oral and written communication to share the findings of quantitative data analyses to a general audience. The format and evaluation criteria for each Module Assignment will vary around the common themes of "showing" where your data came from, how you analyzed this data, and what the implications are of the choices and limitations in your data. Each Module Assignment should be in a tutorial-like format or "telling" what you found from a data analysis in a reporting-like format.

Each Module Assignments will be published as blog posts via the class's Medium publication, so the class will make extensive use of the Medium blogging platform. Instruction on how to create accounts, read, write, and post to the class publication will be covered in the first week of class. Extensive documentation exists in the Medium Help Center as well as multiple tutorials. Students will write their Module Assignments on Medium and submit links via ELMS with the expectation that their writing will be read by the general public.

If a student is unable or does not want to use the Medium platform, they should email the instructor to work out an alternative arrangement before the third week of class.

Semester Project

While the module assignments are meant to provide students with hands-on learning opportunities in short sprints and on a variety of data, the semester project is meant to give students an opportunity to go deeper on a particular subject and dataset that are of interest to the student. The project consists of five milestones: 1) a proposal, 2) a data collection report, 3) an intermediate report, 4) a final report, and 5) a presentation of methods and results. Additional details are provided in the project assignment.

This project should be viewed as an opportunity to engage in an interesting, self-directed research project that could culminate in a workshop or conference paper.

Participation Policy

While the class is offered online and synchronously, interaction within and among the students is still an essential part of the learning experience. Such points include responses to discussion questions the instructors posts on ELMS, asking questions of fellow students' presentations, and consistent engagement with the instructors. These engagements will count for the class-participation portion of the final grade.

Late Policy

This class *does not* institute a late policy for module-level assignments. Module assignments will be made available throughout the semester with suggested deadlines, and students are encouraged to submit them by that deadline. No late penalty will be assessed against that deadline, however.

The motivation for this "lax" policy comes from pedagogical evidence suggesting late penalties are not helpful in getting students to submit work. Rather, module assignments will shared on Medium, and students are encouraged to assess the rate of submissions from their fellow students.

Syllabus Change Policy

Once the semester begins, this syllabus will be revised infrequently, and any revised requirements will be posted as announcements and updated course schedule to ELMS. The instructor reserves the right to make changes to the course's schedule, evaluation criteria, policies, etc. through announcements in class and on ELMS, so please check ELMS regularly. Students should email the instructor if there are any discrepancies or questions.

Campus Policies

It is our shared responsibility to know and abide by the University of Maryland's policies that relate to all courses, which include topics like:

- Academic integrity
- Student and instructor conduct
- Accessibility and accommodations
- Attendance and excused absences
- Grades and appeals
- Copyright and intellectual property

Please visit go.umd.edu/ug-policy for the Office of Undergraduate Studies' full list of campus-wide policies and follow up with me if you have questions.

Additional Accommodation Policy

I understand the difficulty and additional constraints you may be facing during this time. I am willing to work with you to discuss possible accommodation and alternative arrangement. Please do not hesitate to contact me when needed.

Accessibility and Learning Support

Students with disabilities should inform me of their needs at the beginning of the semester. Please also contact the Accessibility and Disability Support Office (http://www.counseling.umd.edu/ADS/). ADS

will make arrangements with the student and me to determine and implement appropriate academic accommodations. Inclusion is one of the iSchool's core values, and we have attempted to make all materials and assignments accessible to people with varying abilities. However, if there is something else I can do to make the class more accessible, please schedule a time to come talk to me. These improvements will benefit not only yourself but also future students!

Get Some Help!

Taking personal responsibility for your own learning means acknowledging when your performance does not match your goals and doing something about it. I hope you will come talk to me so that I can help you find the right approach to success in this course, and I encourage you to visit tutoring.umd.edu to learn more about the wide range of campus resources available to you.

In particular, everyone can use some help in sharpening their communication skills (and improving their grade) by visiting ter.ps/writing and scheduling an appointment with the campus Writing Center.

You should also know there are a wide range of resources to support you with whatever you might need (see go.umd.edu/assistance). If you just need someone with whom you can talk, visit counseling.umd.edu or one of the many other resources on campus. Most services are free because you have already paid for them, and everyone needs help... all you have to do is be brave enough to ask for it.

3.2.2 Exercise: In-Class Exercise on Introductory Network Analysis

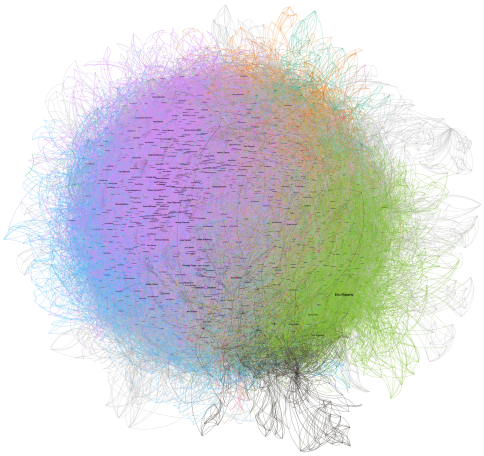
In-class Exercise, Week 5

Published | Edit

- Please submit your code as a Python or Jupyter notebook file on ELMS, and any visualizations as PDFs.
- Exercises are expected to be uploaded to ELMS at the end of class, but the assignment will be available on ELMS until 11:59 PM.
- Discussion is encouraged, but everyone should come up with solutions independently. If you received help from anyone, you should list the name(s) in your submission.

Exercise 1. Visualizing Movie Subgraphs

1. Download **imdb_movies_2000to2022.prolific.json** (https://github.com/cbuntain/umd.inst414/blob/main/data/imdb_movies_2000to2022.prolific.json?raw=true) from GitHub (or on Canvas files)
2. Using **Week03-Exercise** (https://umd.instructure.com/courses/1340623/modules/items/11911650) scaffolding, create a network of actors (nodes represent actors)
3. Add edges based on common movies
4. Use the `**nx.ego_graph()** (https://networkx.org/documentation/stable/reference/generated/networkx.generators.ego.ego_graph.h` function to extract a subgraph centered around Chris Evans
   - Chris Evans IMDb actor is is "nm0262635"
   - Set `radius=3` to include actors within 3 hops from Chris Evans
5. Use `**nx.write_graphml()** (https://networkx.org/documentation/stable/reference/readwrite/generated/networkx.readwrite.graphml` to export your subgraph of actors to a GraphML file
6. **Download Gephi** (https://gephi.org/), a tool for graph visualization, and open your GraphML file with Gephi
7. Using Gephi, create a visualization of the graph, and upload PDF/PNG or Gephi file
   - Example visualization is below, where color represents community ID from modularity, and size is eigenvector centrality



Exercise 2(a). Finding Similar Actors based on Genre

1. Download **imdb_movies_2000to2022.prolific.json** (https://github.com/cbuntain/umd.inst414/blob/main/data/imdb_movies_2000to2022.prolific.json?raw=true) from GitHub (also available on Canvas in Exercises/data)
2. Create a data frame, where each row corresponds to an actor, each column represents a genre, and each cell captures how many times that row's actor has appeared in that column's genre
3. Using this data frame as your "feature matrix", select an actor (called your "query") for whom you want to find the top 10 most similar actors based on the genres in which they've starred
   - As an example, select the row from your data frame associated with Chris Hemsworth, actor ID "nm1165110", as your "query" actor
4. Use **sklearn.metrics.DistanceMetric** (https://scikit-learn.org/stable/modules/generated/sklearn.metrics.DistanceMetric.html) to calculate the cosine distances between your query actor and all other actors based on their genre appearances
5. Print a list of the top ten actors most similar to your query actor using cosine distance
   - The top 10 for Chris Hemsworth are
     - nm1165110 Chris Hemsworth
     - nm0000129 Tom Cruise
     - nm0147147 Henry Cavill
     - nm0829032 Ray Stevenson
     - nm5899377 Tiger Shroff
     - nm1679372 Sudeep
     - nm0003244 Jordi Mollà
     - nm0636280 Richard Norton

- nm0607884 Mark Mortimer
- nm2018237 Taylor Kitsch

6. Describe how this list changes based on Euclidean distance

(Extra Practice) Exercise 2(b). Finding Similar Actors based on Co-stars

1. Download **imdb_movies_2000to2022.prolific.json** (https://github.com/cbuntain/umd.inst414/blob/main/data/imdb_movies_2000to2022.prolific.json?raw=true) from GitHub (also available on Canvas in Exercises/data)
2. Create a data frame, where each row and column correspond to actors, and each cell captures how many times that row's actor has appeared in a movie with the column's actor
   - You can also use the **networkx.adjacency_matrix()** (https://networkx.org/documentation/stable/reference/generated/networkx.linalg.graphmatrix.adjac function to create this matrix from the graph you've created previously
3. As in the genres exercise, select a query actor and find the top 10 most similar actors based on costar similarity
   - Select Scarlett Johansson, "nm0424060", as your target
4. Use **sklearn.metrics.DistanceMetric** (https://scikit-learn.org/stable/modules/generated/sklearn.metrics.DistanceMetric.html) to calculate the cosine distances between your query actor and all other actors based on their genre appearances
5. Print a list of the top ten actors most similar to your query actor
   - Example for Scarlett Johansson is:
     - nm0424060 Scarlett Johansson
     - nm0749263 Mark Ruffalo
     - nm0812307 Peter Sohn
     - nm1165110 Chris Hemsworth
     - nm0538683 Mako
     - nm9125822 Max Ivutin
     - nm10309420 Dayna Hilton
     - nm10558382 Ricochet
     - nm1812991 Tori Kelly
     - nm0262635 Chris Evans

Points	100
Submitting	a file upload

3.2.3 Assessment: Medium Post Describing Some Network Analysis Task

Module 2 Assignment

Published | Edit

Write a 1,000-word Medium post on analyzing the structure of a Web-based network. Relevant networks or graphs include the Web graph (e.g., edges between domains or edges between articles in a single web site) or online social networks. In your analysis, **define what it means for a node to be "important"** and describe how you might find such a node in the graph.

Your post should include the following:

- Describe a question you think can be answered using network data, what specific stakeholder is asking this question, and what decision(s) the answer to this question will inform.
- Describe the data that could answer this question, what fields it contains, and why it is relevant to your question.
- Explain how you collected some subset of this data (e.g., libraries like requests, BeautifulSoup, tweepy, praw, etc. or from data archive).
- Define what entity is represented by a node/vertex in the graph, and what relationship between these nodes does an edge represent.
- Define "importance" in your graph and identify a set of **at least three** important nodes.
- Provide an answer to your question, explaining your analysis of the data you collected, and how it answers that question.
- Include figures or tables summarizing your findings.
- Describe how you cleaned up this data, common bugs you think others might encounter, and how you fixed them, etc.
- Discuss the limitations of your analysis. What's missing? How might it be biased?
- Include a link to one of your GitHub repositories that contains the code you have developed for this assignment.

When you have written your post, publish it via Medium, **add your post it to the class publication via Medium**, and submit the URL to it via Canvas.

Tag your story as "inst414spr24a02" - Articles with this tag will automatically be added to the appropriate assignment tab after I accept on the class publication.

You can review the grading rubric for this assignment here: **A2-Rubric.doc (https://umd.instructure.com/courses/1361527/files/77671128?wrap=1)** **(https://umd.instructure.com/courses/1361527/files/77671128/download?download_frd=1)**

Points	100
Submitting	a website url

INST414 – Medium Post Grading Rubric, Assignment 2

Name _____ Topic_____

#	Item Name	Max	Earned
1	Describe a question you think can be answered from data, what specific stakeholder is asking this question, and what decision(s) the answer to this question will inform.	10	
2	Describe the data that could answer this question, what fields it contains, and why it is relevant to your question.	10	
3	Explain how you collected some subset of this data (e.g., libraries like requests, BeautifulSoup, tweepy, praw, etc. or from data archive).	10	
4	Define what entity is represented by a node/vertex in the graph, and what relationship between these nodes does an edge represent.	10	
5	Define "importance" in your graph and identify a set of at least three important nodes.	10	
6	Provide an answer to your question, explaining your analysis of the data you collected, and how it answers that question.	10	
7	Describe how you cleaned up this data, common bugs you think others might encounter, and how you fixed them, etc.	5	
8	Include figures or tables summarizing your findings.	5	
9	Discuss the limitations of your analysis. What's missing? How might it be biased?	10	
10	Include a link to one of your GitHub repositories that contains the code you have developed for this assignment.	10	
11	Excellence	5	
12	Length	5	
	TOTAL	100	

Comments:

1

3.2.4 Assessment: Semester Project

Semester Project Overview

Overview

A key element of this course is a semester project, wherein you will gain hands-on knowledge with collecting data from web sources, filtering this web-scale data, and extracting useful insights from it. As with any analysis, to get people to pay attention and understand your insights, you will need to be able to present your work efficiently in written and oral form.

You may work in teams of 1, 2, or 3, but the amount of work you perform will need to scale accordingly. Larger teams might be allowed under unusual circumstances but will require special permission from the instructor. Although the project work will scale with students, the administrative parts will remain constant, so having a large group will make it easier for you. All students will need to have clearly defined roles as demonstrated in the final report and presentation.

Scale of Project

The specifics of the project will be very flexible. I expect each group to collect data from some web service or platform and apply data mining techniques to this data to extract insights. The goal is to gain more in depth experience in some aspect of the class and to do so in a setting where the instructor can provide guidance.

The lectures will cover the material, homework will reinforce the material, and this project should provide deep understanding of some aspect of the material. Students who demonstrate a deep understanding of some aspect of analysis covered in the class through their project will receive high grades.

Project options may include:

- Applying touchstone techniques from the class to real-world web-scale data to extract unique insights.
- Comparing several techniques from the class on a real-world data and synthesizing an analysis of the advantages and disadvantages of each.
- Proposing and testing new extensions to techniques from class on a real-world web data.

Project Elements

Your project consists of the following elements:

- Project Proposal
- Data Collection Report

- Midterm Report
- Final Presentation
- Final Report

Semester Project - 1 - Proposal

Published | Edit

Prepare an at least **200-word** document outlining your plan, which should contain at least the following:

1. Who is in your group?
2. What problem are you solving?
3. What data do you plan to use, and how will you collect it?
4. What technology do you intend to use or build in this project?
5. What insights or predictions do you want to extract from this data (be specific)?
6. Why this problem is interesting, or what is its business use case?
7. What metric(s) will you use to measure success of your project?
8. What do you expect to submit at the end of the project?

These questions are meant to establish that this project is more than just another homework assignment.

Also, you must collect your own data. Do not use pre-filtered datasets from sources like Kaggle. As a central element of this course is about collecting and cleaning your own data, this project is meant to provide you hands-on experience with APIs, web-based collection, or similar.

It is common for the instructor to provide feedback to alter or modify your proposed plans. This feedback can either happen by students engaging with the instructor during virtual office hours before the proposal is due, or will come in feedback on the specific proposal. This step is most important when the topic is related to material that is covered later in the class.

If there are multiple students on a project team, the project should scale accordingly. This scaling may for instance involve trying three extensions to a technique and comparing to see which works the best (for 3 students) or processing a data set in two phases using two consecutive techniques (for 2 students). The work does not need to be split as such, but it should be clearly more than for a single student.

If this project is part of a larger project (e.g., part of your thesis work, extending your lab's work), please indicate its scope with respect to previous work. As a way to separate this class from other courses on machine learning, we will not allow projects with central focus on building classifiers. A classifier may be used to evaluate a result, but the primary contribution of the project should be focused on extracting insight from the data.

Topics

The best projects occur on topics with which students have experience and/or passion. First, look within your own experience, whether that is your research, your work, or your hobbies. You are welcome to use data that's available in any web-based system, which includes social media data, API-provided data, review data, etc.

You can also find examples and data sources here:

- Examples of short projects: **https://medium.com/information-expositions-f2019** (https://medium.com/information-expositions-f2019)
- Large datasets from the US govt: **https://www.data.gov/** (https://www.data.gov/)
- Public datasets in Google's BigQuery database: **https://cloud.google.com/bigquery/public-data** (https://cloud.google.com/bigquery/public-data)
- Twitter APIs: **https://developer.twitter.com/en** (https://developer.twitter.com/en)
- Reddit APIs: **https://www.reddit.com/dev/api/** (https://www.reddit.com/dev/api/)
- Yelp Reviews: **https://www.yelp.com/dataset** (https://www.yelp.com/dataset)

If you have an advisor, they may also be a good source of problems and data.

Evaluation

Your proposal will be evaluated by how well it answers the above questions, according to this rubric: **Rubric.doc** (https://umd.instructure.com/courses/1340623/files/70460553?wrap=1) (https://umd.instructure.com/courses/1340623/files/70460553/download?download_frd=1)

Points	100
Submitting	a file upload
File Types	pdf

Due	For	Available from	Until
Feb 15	Everyone	-	-

+ Rubric

INST414 – Semester Project Proposal

Name _____ Total Points: _____

Item	Item Name	Max	Earned
Sections			
1.	Who is in your group?	5	
2.	What problem are you solving?	15	
3.	What data do you plan to use, and how will you collect it?	15	
4.	What technology do you intend to use or build in this project?	10	
5.	What insights or predictions do you want to extract from this data?	15	
6.	Why this problem is interesting, or what is its business use case?	15	
7.	What metric(s) will you use to measure success of your project?	10	
8.	What do you expect to submit at the end of the project?	15	
Comments			
	TOTAL	100	