CMPT 419

Nicholas Vincent

2025-09-03

Table of contents

Preface

This website contains the CMPT 419 / 980 Fall 2025 Course Materials. It was produced using "Quarto".

To learn more about Quarto books visit https://quarto.org/docs/books.

Outline

You can find the course outline here. The content will also be pasted at the end of this Introduction Chapter for your convenience.

Syllabus and Course Content

If there's anything you're looking for you can't find in this site, check the Canvas homepage – all internal-facing content is there (lecture recordings, notes, etc.).

The syllabus file describe the overall structure of the course and course policies.

The readings page will be updated weekly.

I will post coding assignments here as the semester goes in (note that assignments from previous offerings are online, as I make these GitHub pages public. Some changes will be made to the assignments). The GitHub will list the requirements and grading scheme. Submission will be via Canvas.

Outline Text

Artificial intelligence (AI) technologies have seen a large surge in interest from researchers, investors, businesses, and everyday end-users. These technologies stand on the shoulders of giants – they rely on a large body of research in computing and other fields, as well as modern feats of engineering from organizations that operate them. However, they also rely heavily on data, and thus, people.

Search engines rely on click data from users and content written by volunteers such as blogs and Wikipedia articles. Recommender systems rely on explicit user feedback (e.g., "star ratings") and behavioral data (e.g., browsing history) that reveal user preferences. Supervised learning relies on crowd workers, volunteers, and sometimes unwitting users (e.g., reCAPTCHA participants) to label images and text. And new generative AI systems rely on the wide swathe of content shared on the web. Without this data generated by the public, technologies that use machine learning and statistical models could not exist. The critical role of data suggests an untapped source of power for data creators, i.e., the broad public. Furthermore, it suggests a number of exciting questions about how a data-centric view can advance both AI research and the development of AI products and other systems.

In this course, we will explore AI technologies with a data-centric, and thus human-centric lens. We will discuss topics such as: - Exposure to foundational reading in interdisciplinary AI - The intersection of humanities scholarship and technical computing aspects of AI - Modern research in data valuation - Relevant work in social computing, including the impact of online platform design choices. - The potential for collective action involving data. How might social movements – ranging from protests that withhold data to movements to collect and share data in the public interest – impact the future of AI? - The economics of data. Students will be introduced to recent work on data markets and unique properties of buying/selling data for AI.

We will read papers on these topics together. Students will work together to synthesize and present knowledge from research papers, and present their own opinions on these topics. The course will centre a structured final project that will enable students to conduct interdisciplinary responsible AI research or bring responsible AI concepts to bear in industry contexts.

Students may benefit from having taken a course in AI, ML, or data science (or have equivalent experience from e.g. an internship, a research project, a personal project).

Example SFU courses:

CMPT 310 - Intro Artificial Intelligence CMPT 353 - Computational Data Science CMPT 414 - Computer Vision

Having taken an HCI course or relevant social science course (e.g., sociology, economics) is a plus, but CS students without this experience who want to explore interdisciplinary CS work that is "human-centered" are welcome. Similarly, students in the humanities who have some exposure to data science are also welcome.

We will work through a low-stakes "example assignment" in week 1 so that students can assess their comfort level.

In short, students should ideally be decently comfortable with both (1) working with computational notebooks (Python, R, Julia, etc.), quickly loading and working with quantitative data, training and evaluating machine learning models and (2) reading and critically thinking about new scholarly perspectives and ethical considerations.

This course will include a heavy reading component.

The course will aim to develop the following skills:

- students will become more comfortable reading research papers that take an interdisciplinary approach to study AI
- students will gain experience presenting information from papers
- there will be a project component that incorporates coding
- students will be able to articulate some of the ongoing challenges in "human-centric" AI.

Students will gain exposure to the following concepts:

- interdisciplinary research in AI
- data valuation techniques, and their applications for AI research/practice
- social computing

1 Syllabus

CMPT 419 D200, Nicholas Vincent, Fall 2025

1.1 Lectures and Office Hours

See go.sfu.ca for exact location and time.

Office hours: posted on Canvas.

We can have additional office hours by appointment and/or popular demand.

1.2 General structure of our "lecture" time:

- Each Monday (1 hr sessions), we'll briefly discuss the previous week's readings, I'll introduce any readings and assignments for the week, and I'll start the "lecture content" for the week.
- I'll aim to hold at least 5-10 min every Monday to walk through assignments together and take questions. You're welcome to use this time to start working and see if questions arise.
- On Thursday (2 hr sessions), we'll finish lecture content and have a discussion about the lecture/readings for the first hour, and then typically use the second hour for some kind of activity or "lab time". We may use some of this time to work on assignments and projects and to take quizzes or practice quizzes.
- I'll always take questions at the beginning and end of each lecture session. You're always welcome to email me, but I may take 2-3 business days to respond to emails. Asking questions in class will provide a quicker response and your classmates may benefit from your questions as well! Please include "[CMPT 419]" (or CMPT 980) in your email to help me keep track of requests.

This course is designed to have a particularly heavy reading and discussion component. Please be prepared to read quite a bit of material, and to talk about it.

1.3 About course assignments:

Each week has a set of assigned readings:

- There will be a set of mandatory readings.
- There will also be some optional readings. You are encouraged to read the abstracts and/or Introduction sections of the optional readings to see if they align with what you hope to get out of the class. I'll do my best to organize these by theme, and will add more based on the interests you express.
- Each week, you'll submit some relatively brief "reading responses" via Coursys. These will be very lightly graded (there really aren't wrong answers). However, you should be prepared to defend your reading responses live in class (I may cold call students, and you should be able to speak to your reading response in a way that suggests that you did indeed read the required material. You need not agree with the all the arguments presented or understand all the material).
- For reading responses, I strongly recommend against AI assistance. I personally prefer that you submit bullet points rather than bullet points that prompt an LLM to output flowerly text. (I actually read these, and I'm very familiar with all the ChatGPT-isms, and generally don't need to read "This isn't just a great reading suggestion from the professor – it's a groundbreaking article.")

Reading schedule:

- Assigned readings for Week X are considered "finalized" on Monday of the preceding week (Week X-1), and should be completed by Monday of Week X.
  - For example: During class on Monday of Week 2, I'll post and tell you all the required readings for Week 3, which you should finish over the next 7 days.
  - I'll try to provide a solid "look ahead" of course material, but it may be subject to change based on your feedback, course progress, and even current events – so you should check the readings each Monday after class. For instance, in the past, I have extended time to complete readings that students found particularly dense.

1.4 About course organization

The course will be organized roughly in terms of 4 "modules":

- Module 1: Administration and Introduction to Different Frameworks for doing "Human-Centered" or "Data-Centered" Work (Weeks 1-4)
- Module 2: Technical work in data valuation, data scaling, and algorithmic collective action. (Weeks 5-7, 3 in total)
- Module 3: Online platforms, content ecosystems, and data. (Weeks 8-10, 3 in total).

- Module 4: Frontiers in Data Governance: Voting, Markets, and More (Week 11-13, 3 in total).

We will have one assignment per module (coding / data analysis).

1.5 Grading

- 10% reading responses (12 total, drop lowest 2 using Canvas, so each of your top 10 responses effectively is worth 1%)
- 20% coding assignments (4 total; 5/5/5/5, drop lowest 1 using Canvas, so each of your top 3 assignments effectively is worth 6.6.67%)
- 20% quizzes (2 total; 10/10; may adjust scores for difficulty)
- 50% final project (5% project proposal, 45% actual project; must submit a written document and a presentation for both)

1.6 Course FAQs

Q: Is attendance mandatory?

A: While I won't give you direct marks for attendance, you are highly encouraged to attend class whenever you are able to. I do expect all students to participate in class discussion at some point (i.e. I do want everybody to speak up at least once). I will try to facilitate this "softly" via some cold-calling to discuss reading responses but this will not be strictly enforced (e.g., if circumstances arise, we can meet in office hours to discuss your progress in the course). If a very "loose approach" to soliciting participation isn't working at the mid-point to class, we'll discuss (as a class) alternatives.

I am very supportive of students staying home when sick, and understand a variety of personal situations may arise that prevent you from going to class. You do not need to email me to miss class, but are welcome to ask follow up questions (I may just point you to the class notes and encourage you to talk to your classmates). To earn a high mark in this class, I encourage you to plan to attend all lectures you are able to.

Q: Will this class involve coding?

A: Yes, there will be some coding assignments in the class that are designed to give hands-on experience with certain course concepts. You are free to use a variety of programming languages and tools for these assignments, though will be encouraged to use some "standard" solutions based (primarily: Python for ML and data science related components, Javascript

and web-programming for some design components). For coding assignment, LLM assistance will be allowed (with some caveats). I expect available LLM tooling to change quite a bit *during* our semester, so we'll play with tools together as part of the course.

Q: How many assignments will we have?

A: You will complete 4 assignments (involving coding and data analysis) and 1 project.

Q: Can I work in a group?

A: There will be opportunities to do group work, but you must write a contribution statement for everything. You must review all your team's code and writing! Individual assignments that allow group work will have specific details for how this will work.

Q: Are there quizzes, a midterm, and/or a final exam?

A: There will be in-class quizzes, but no "midterm" or "final". They will be announced in advance and some kind of make-up option will be available for sick students. Any "testable" material will be drawn only from in class lecture materials and mandatory readings. The goal of the quizzes is to provide additional incentives to engage with material each week.

Q: What materials do I need?

Reading materials will be provided digitally by the instructor. There will be no single textbook – rather, we will read an assortment of research papers, book chapters, etc. You will be asked to spend some time installing software tools on your own. You will have some flexibility in which tools you choose – there will always be a free option available.

Q: Can I use ChatGPT (etc.)?

A: You may use generative AI tools to assist with your coursework, but must provide complete logs for any outputs you use directly and any artifacts you submit should indicate the provenance of any generative AI outputs.

e.g.

- "This slide was produced by model XYZ"
- "This summary paragraph or code snippet was produced entirely by ChatGPT"
- "This code was generated with the help of ChatGPT, but heavily edited"

Individual assignments may have specific requirements you should pay attention to.

2 Readings

2.1 Week 1

N/A

2.2 Week 2 (begins Monday September 8, 2025-09-08)

The goal of the week 2 readings is to begin getting some exposure to what different researchers mean when they refer to human and data centered ML/AI. We want to start developing some intuiton for when human-centered practices or data-centred thinking might materially change how we design a system, come up with a research question, or deploy a model.

Reading 2.1: (Chancellor 2023)

First, we'll read "Toward Practices for Human-Centered Machine Learning" by Stevie Chancellor, published in the Communications of the ACM. CACM is a venue in which experts in various fields of computing write broad pieces for the entire computing community.

- How to access: Visit https://cacm.acm.org/magazines/2023/3/270209-toward-practices-for-human-centered-machine-learning/fulltext

Reading 2.2: (Mazumder et al. 2023)

Second, we'll read the Introduction of the DataPerfs paper, published in NeurIPS 2023 Datasets and Benchmarks Track.

- How to access: Visit https://arxiv.org/abs/2207.10062
- Notes: You only need to read the Introduction.

2.2.1 Response Instructions:

1) Please write one to two paragraphs describing why you'd like to work on, or with, ML/AI systems? You can imagine these paragraphs as text you might include in a cover letter. It might be worth expending some serious effort in case you need to use text like this in the future.
2) Please list 1-3 "domains of interest" (e.g., social media, content recommendation, law, health care, mental health, the environment, economics). They can be at any level of granularity (e.g. "AI for health" is OK, as is "AI for oncology"). Similarly to part 1, the purpose of this is to help me identify trends in your interests so I can suggest optional readings that are of interest to you and your classmates!

If you submit any reasonable formatted submission for this reading response, you'll receive full credit. In future response instructions, you might see something along the lines of, "you must quote on of the readings directly to support your point".

For this reading response, you'll submit via Canvas.

2.3 Week 3

The goal of the week 3 reading is to gain further exposure to various framework focusing on humans (Schneiderman reading) and/or data (Sambasivan et al reading and Zha et al reading).

Note this week is a bit longer than than Week 2. We'll check in on how it's going, workflow wise, to complete these readings, and focus on challenges that may come up for those who haven't had many reading heavy computing classes previously.

Reading 1: Shneiderman 2020.

- Citation: Shneiderman, B., 2020. Human-centered artificial intelligence: Reliable, safe & trustworthy. International Journal of Human–Computer Interaction, 36(6), pp.495-504.
- About: This is a paper published in the International Journal of Human–Computer Interaction.
- How to access: visit here on campus or arxiv version off campus

Reading 2: Sambasivan et al 2021.

- Citation: Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P. and Aroyo, L.M., 2021, May. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1-15).

- About: This is a paper published in ACM CHI, the main venue for human-computer interaction research.
- How to access: visit here
- You should try to understand the Introduction, Methods, and Findings (you can skim the Related Work and bookmark for later)

Reading 3: Zha et al 2023.

- Citation: Zha, D., Bhat, Z.P., Lai, K.H., Yang, F. and Hu, X., 2023. Data-centric ai: Perspectives and challenges. In Proceedings of the 2023 SIAM International Conference on Data Mining (SDM) (pp. 945-948). Society for Industrial and Applied Mathematics.
- About: This is a short perspective paper in a data mining conference.
- How to access: visit here
- Notes: you should read the short perspective paper. You may optionally also check out the longer survey paper and repo linked here

Additionally, read Sections 1 and 2 from this paper

2.3.1 Response Instructions

Imagine you are a manager at a large tech company tasked with developing a new AI product. You can pick one of the following three options based on your interests, or suggest your own product:

- A large language model that will read physician notes and make suggestions about how to treat patients
- A recommender system for a video-based content app
- A facial recognition system that will be sold via API credits

Q1: Thought experiment: Please write 1-2 paragraphs describing how adopting any of the suggestions from any of this week's readings (HCAI, Data cascades, data-centric ai, public ai) might change your product's features (first define the product). Please directly reference (e.g. directly quote) one or more of the readings.

Q2: Please list three examples of "harms" that might occur from a failure to do "data work" as defined in the Sambasivan reading. You can use the same AI product you picked for Q1, or discuss one or more different AI products. You don't need to quote the reading directly for this part.

Q3: Describe data augmentation in one sentence. Describe out-of-distribution evaluation data in one sentence.

Q4: Please let me know roughly how the long the readings and responses took so we can calibrate!

(Submit on Canvas)

2.4 Week 4

For this week, there will be just two readings. The goal this week is still to gain exposure to all the different frameworks for thinking that motivate "human-centered AI" and "data-centered AI". Last week, we saw several more frameworks, and in particular learned more about specific "data-centric" task formulations.

In our first reading, Chancellor highlighted that human-centered ML is often tied deeply to specific goals around fairness, justice, and values. This week, we'll dive into this with a reading from a textbook.

This week we'll just read two pieces: one is a longer introduction to a fairness in ML textbook, and the other is the Introduction to another research paper.

Please read the Introduction of FairML: https://fairmlbook.org/introduction.html

```
While our course material will differ in some ways from a Special Topics course that's entire
For our purposes, the concept of the "machine learning loop", and especially measurements and
```

Please read the Introduction of "Value-Sensitive Algorithm Design: Method, Case Study, and Lessons" by Zhu et al, published in CSCW: https://dl.acm.org/doi/10.1145/3274463

The goal of this reading is to see another example of how a research project might concretely seek to incorporate values into design. You don't need to read the full paper, though if you're particularly interested in working on algorithm design you might want to!

Response Instructions:

Q1: Please summarize in your own words the idea of the "machine learning loop". Do your best to capture the key concepts from the FairML intro.

Q2: How does the discussion of feedback loops in FairML Introduction compare to the discussion of feedback in Schneiderman's HCAI? You can just write 2-3 sentences describing major differences or similarities you see. There's not a correct answer here.

Q3: Quick retrieval: What online platform do Zhu et al. use to study value-sensitive design in a real-world setting?

2.5 Week 5

This week, we are going to start reading a long piece that surveys training data influence:

- Citation: Hammoudeh, Z. and Lowd, D., 2024. Training data influence analysis and estimation: A survey. Machine Learning, 113(5), pp.2351-2403.
- How to access: https://arxiv.org/abs/2212.04612

This piece will represent a large jump from reading about high-level frameworks that consider social factors, incentives, etc. to a much more mathematical framework for thinking about data-centricity. Accordingly, we're going to work through this piece (and some excerpts from the key citations) fairly slowly. For this week, you should just read pages 1-10 (on the arxiv version – up to Section 4).

For this week's reading responses, you do not need to answer any questions. Instead, please use the reading response as a chance to record any questions that come up (if you want to just ask them in lecture, that's great too!)

2.6 Week 6

This week we will continue reading the Hammoudeh and Lowd survey.

Please read pages 10-21 (up to Section 5.1.2, "Representer Point Methods"). Grad students should finish the entire reading.

For your response, please answer the following 3 questions:

Q1) Please describe the difference between a leave one out influence value and a Shapley value, in the context of training data influence.

Q2) What is the main issue with calculating retraining-based data values, as described in our reading?

Q3) If you were asked to run a new data market that makes use of influence estimates, which approach from the reading would you use and why? There is no correct answer to question, but you should aim to think through some of the trade-offs.

2.7 Week 7

This week, please read:

- Citation: Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M.M.A., Yang, Y. and Zhou, Y., 2017. Deep learning scaling is predictable, empirically. arXiv preprint arXiv:1712.00409.
- How to access: https://arxiv.org/pdf/1712.00409.pdf

You should also read https://arxiv.org/abs/2012.05345.

For your response, please answer the following 3 questions:

Q1) Please describe the consistent finding across all ML domains in this study.

Q2) What are the three "learning regions" that the authors identify?

Q3) About how long did this reading take?

2.8 Week 8

This week, we're going to start talking about online platforms and their role a key AI training data source. We'll orient much of our discussion around recent advances in Large Language Models, but with the caveat that the core ideas are equally relevant to search, recommendation, and classification systems in many applied domains of interest to our class (e.g. medicine, analytics for sports and games).

First, please read these two short blog posts from 2020 and 2022.

- https://dataleverage.substack.com/p/dont-give-openai-all-the-credit-for
- https://dataleverage.substack.com/p/chatgpt-is-awesome-and-scary-you-deserve-credit

Next, please read Sections 1 and 2 of this pre-print paper:

- https://arxiv.org/abs/2101.00027

For this week, please list three specific online platforms that are useful for AI training.

2.9 Week 9

Please read these two short papers about "Public AI":

- https://arxiv.org/abs/2311.11350
- https://arxiv.org/abs/2507.09296

Starting in Week 9, your reading response will be very short. You should just record any connections you see between the reading and your project (there might be none). You should also record any questions you have after completing the reading. If that's the case, just say so.

2.10 Week 10

Labs – TBA

2.11 Week 11

For this week, please read:

First 10 pages of https://arxiv.org/abs/2402.00159

Section 2 of https://dl.acm.org/doi/abs/10.1145/3531146.3534637

Skim this webpage: https://weborganizer.allen.ai/ and look at linked sample data on HuggingFace.

Starting in Week 9, your reading response will be very short. You should just record any connections you see between the reading and your project (there might be none). You should also record any questions you have after completing the reading. If that's the case, just say so.

2.12 Week 12

Please read the Abstract and Introduction of the following papers. The goal of this set of readings is to get some exposure to different arguments and research directions in the space of data-sharing markets and perspectives on data:

Prainsack, B. and Forgó, N. 2022. Why paying individual people for their health data is a bad idea. Nature medicine. 28, 10 (Oct. 2022), 1989–1991. https://www.nature.com/articles/s41591-022-01955-4

Acemoglu, D. et al. 2022. Too Much Data: Prices and Inefficiencies in Data Markets. American Economic Journal: Microeconomics. 14, 4 (Nov. 2022), 218–256. https://www.aeaweb.org/articles?id=10.1257/mic.20200200

Read: Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., … & Hudson, M. (2020). The CARE principles for indigenous data governance. Data Science Journal, 19, 43-43. Available as HTML at: https://www.adalovelaceinstitute.org/blog/care-principles-operationalising-indigenous-data-governance/

https://data-feminism.mitpress.mit.edu/pub/vi8obxh7/release/4

Starting in Week 9, your reading response will be very short. You should just record any connections you see between the reading and your project (there might be none). You should also record any questions you have after completing the reading. If that's the case, just say so.

3 Assignments

3.1 Assignment 1

Fall 2025 Deadline: Sep 28 11:59pm.

You will submit a short report on tool-related exploration as your first "coding assignment" and do a short coding exercise.

This assignment is very flexible to accomodate for many different backgrounds. You should pick an option that's useful for you to get yourself ready to go in this class.

First everyone should answer the following questions, and submit via Canvas a report which describes your answers. You can submit as a PDF file or in plaintext as a .md file. Please organize your answers in terms of question numbers, as indicated below. It's perfectly OK if some of your answers are very short!

Your choices are not binding – this primarily to get you thinking about these choices early on and to encourage you to explore some of the available options before additional assignments are due.

Q1: Which tools you plan to use for writing code (IDE, AI assistance, version control). e.g. answers might include: VS Code, Sublime text, ChatGPT, Copilot, GitHub

Q2: What open questions or concerns do you have about code writing tools?

Q3: Which ML libraries / frameworks / tools do you have familiarity with already?

Q4: If given the choice, which ML libraries do you prefer to use for any assignments that involve training and evaluating a ML model?

Q5: Which ML libraries / frameworks / tools do you hope to learn more about ("I'm not sure, that's why I'm taking this class" is an OK answer!)

Q6: Which tools you plan to use to read and take notes on papers, if any (pen + paper or PDF reader + notes app is perfectly fine answer!)

Q7: Which tools, if any, you plan to use for project management?

Next, you should complete one "speedrun" modeling project. You can adjust the difficulty for this, but it should involve loading a dataset, creating a test set, experimenting with more

than one modeling approach, and printing more than performance metric with at least 2 presentation modalities (confusion matrix, precision-recall curve, etc.).

If you're new to using Python, you just use the pre-loaded "iris" dataset. If you're a veteran, you might use this as an opportunity to try to "speedrun" a decent baseline for a Kaggle competition.

You should submit one code file (PDF output, .ipynb, .py; if other format, such as using R or Jupypter, include a note.)

Basically, your goal here is to make sure you are able to very quickly produce a "simple baseline" for a classification task from a blank notebook or repo (your code).

To earn the marks, just hit the above requirements: load, split, multiple models, multiple metrics, multiple presentation modalities.

Anything else you want to add is extra!

3.2 Assignment 2

Our Module 2 content is focused on understanding the broad question: *Which groups of observations – or groups of people – are "responsible" for a given model output or "capability"?*

In this assignment, we'll get some hands-on experience with the concept of training data influence.

There are four parts to the assignment. You'll need to write code to train a ML model and produce influence values for some of the training data in the model. Below, the requirements for each part are described.

Each part will have a coding component and a report component. You will turn in one file (or multiple) with code (e.g., a `.py` file or `.ipynb` file) and one report PDF. If using computational notebooks like a Jupyter notebook, you may combine these two into a single file (e.g. a notebook exported to a PDF with code visible.

In your code, you can use comments to designate which parts of your code correspond to each part.

Note 1: you may work on this assignment in groups of 1-3.

Note 2: you may use generative AI on this assignment, and must report your use. FYI – the instructor has tried out several models, and they're definitely useful, but you'll need to be careful about explaining your choices. In fact, I'll even provide you some example outputs of what you get from directly copy-pasting the assignment into several strong models!

Note 3: Finally, as an additional incentive to avoid literally just copy-pasting the assignment into your favorite consumer AI product, I may randomly select some students to explain their solutions in class.

3.2.1 Part 1: Preliminaries

First, you should select a dataset to work with, define a specific classification (must do classification for this assignment) task, and establish a baseline model.

If you're looking for inspiration, you might consider selecting something from https://archive.ics.uci.edu/

You will not be graded based on your dataset choice, task choice, or achieving a certain level of performance.

Rather, you will be graded based on your ability to describe, in a scientifically complete fashion, the choices you've made.

You are recommended to select a dataset from a domain of your interest and then take a small random sample of that dataset (e.g., 10000 rows – though you can lower this if using high-dimensional data, want to use deep learning, etc. – ask us if you're unsure) to ensure that you can complete this assignment quickly, without being burdened by excessive computational costs. What constitutes "excessive" here will depend on your access to computing resources (you may wish to explore using an online tool with some degree of free compute like Google Colab).

If you select a dataset you are interested in, you may be able to reuse some of your code you write for this assignment for your project.

Suggested approach: I recommend first training several models on the "full dataset" (e.g. logistic regression, basic random forest, KNN, XGBoost). See how long this takes. Then, try subsampling 10% or 1% of your data and see if the training time falls low enough that you think you can reasonably retrain a model at least 50 total times. (if a training run takes a day and you have a week left… this is too much training time!)

Specifically, you should write code to do the following:

- Load a dataset into memory. Describe the dataset in your report. (2 marks)
- Process into features and labels. Describe the features and labels in your report. (2 marks)
- Split into train and test sets. Describe your specific approach (e.g. random 80/20 split, time-based split, etc.) (2 marks)

- Train some classifier. It does not need to be the "best" possible performance for your chosen dataset, though you may want to try a few options if feasible to do so. (2 marks)
  - You should list some reason for your choice of classifier.
- Report performance of your baseline classifier: accuracy, confusion matrix. You are encouraged to include a precision-recall curve or TPR vs. FPR curve (i.e. AUROC curve), though if you think it isn't helpful you can just mention why not. You must choose a "primary metric" that you will use for your data value estimates, and you should justify this choice. (2 marks)
  - Why is this measurement appropriate for the data/task you chose?

10 marks total for part 1

3.2.2 Part 2: Brute force LOO influence

Next, you should select (manually or randomly) 10 training data points (i.e., observations) and compute the exact leave-one-out LOO influence of these examples on your chosen primary metric.

You can earn up to 4 marks for clean and correct code.

Report the influence score for each of your observations. You may do this in a table or plot. (2 marks).

Please briefly comment on any trends you observe with your influence scores. Are any points with high influence unusual in any way? It's OK if they're not, but you should demonstrate that you looked. (2 marks)

8 marks total for part 2.

3.2.3 Part 3: Group-level influence

Next, you should select (manually or randomly) 10 different *groups* of data points of different sizes. For instance, you might randomly select 10%, 20%, 30%, etc. of the training data. You should compute the exact leave-entire-group-out influence for each group.

You can earn up to 4 marks for clean and correct code.

Report the influence score for each of your groups. (2 marks)

For part 3, you must also include a plot that shows group size compared with influence. (2 marks)

8 marks total for part 3.

3.2.4 Part 4: Shapley values

Finally, we will roughly estimate Shapley values for our training data.

For each observation and each group, you should compute the Shapley value using Truncated Monte Carlo Shapley Value Estimation (described briefly in our survey reading and in more detail here: http://proceedings.mlr.press/v97/ghorbani19c/ghorbani19c.pdf).

This will involve a coding challenge: implementing this particular Shapley value estimation algorithm.

In the Ghorbani and Zou paper, the authors suggest using a truncation cut-off: if performance for a given point / time step is very close to full performance V(D), we don't need to retrain again.

We will go a step further and use the following rule to ensure our code doesn't take too long to run: we should take our best guess at the Shapley value for each training data point after only 10 total permutations have been examined. In other words, your code should just re-shuffle the training data 10 times, compute the marginal impact of each training point, and then average these across the 10 permutations.

Furthermore, you may further subsample your training data (E.g. if you started with 100k rows and have only been using 10k so far, and need to drop down to 1k… you can) for this part if needed to complete the assignment in time.

You can earn up to 4 marks for clean and correct code.

Here, you need only to plot the distribution of all Shapley values. (2 marks)

If you have extra time, you are encouraged to compute more accurate Shapley value estimates by using more permutations and compare the Shapley values to LOO influence from part 2, but this is optional.

6 marks total for part 4.

3.2.5 Grading

This assignment will be graded based on both code correctness and an accompanying report. You can earn marks for each of these separately (i.e. if you have errors in your influence calculations, you can still earn the marks for reporting and visualizing the potentially erroneous data values).

To recap, there are:

- 10 marks available in part 1

- 8 in part 2
- 8 in part 3
- 6 in part 4
- for a total of 32 marks.

Part 4 will likely be the most difficult, but offers the least marks, so you should consider completing the earlier sections first.

If you submitted with a group, your report must include a 'contribution statement' that describes how each member contributed.

4 Assignment 3

Our Module 3 content will focus on understanding datasets from online platforms and elsewhere. This will be helpful in both understanding LLM pre-training data, and should also be helpful in making progress on your project.

In this assignment, we'll inspect two datasets: a large text dataset and a second dataset from any domain that chose (either because of personal interest or because it helps you make progress on your project).

Some learning goals: - Understand the process by which you might gather a very large web-scale dataset (we will not actually download any full datasets, however!) - Get experience with dataset documentation practices - Get experience with the "just look at your data!" hack

You may want to use: - https://huggingface.co/docs/datasets/en/index - https://github.com/allenai/wimbd

4.1 Part 1: Getting some data (4 marks)

First, you should gain access to a small sample of LLM training data. You may use Dolma (https://allenai.github.io/dolma/), RefinedWeb (https://huggingface.co/datasets/tiiuae/falcon-refinedweb), or any other source you've come across.

The main challenge of part 1 is acquiring a good sample.

Your goal is to acquire 300k tokens (about 0.01% of the 3 trillion token in Dolma).

For part 1, write a short 'methods' section and key code you used to get a random sample of LLM pre-training data onto your machine (3 marks).

Second, write a short 'methods' section that describes the dataset you chose based on your project/interests (1 mark).

4.2 Part 2: Datasheets (8 marks)

Next, you should visit https://arxiv.org/pdf/1803.09010 and answer all the questions in Section 3.2 for both datasets. (4 marks)

4.3 Part 3: Data Assessment (8 marks)

Next, you should prepare a random sample of 10 "observations" from each dataset. We will be manually assessing their quality! You should produce a table that shows each observation and some kind of "assessment column" of your choosing. For instance, you might manually assess the "usefulness" to a certain task. You might consider the toxicity of the content (in the text domain).

To create this "assessment column", you will likely need to make some subjective choices. You could create a quantitative assessment as well (e.g., the number of times of a key word appears in text data).

Please describe and briefly justify your chosen metric. For your project dataset, you're encouraged to select something relevant to your project. The point of this assignment is to provide a forcing function to "look at your data", which is a common adage and suggestion for all kinds of AI projects! (4 marks)

Section 3a of your report will consist of two tables, with 10 rows and at least 2 columns.

Example row:

"this is a sentence in my LLM training data from a blog", Toxicity score: 0 "this is a really angry mean sentence in my LLM training data", Toxicity score: 10

In your report, you should write a paragraph summarize what you found. Perhaps you were surprised by the text, or perhaps everything was just as expected. (4 marks).

4.3.1 Part 2: Data Napkin Math

Next, visit https://nickmvincent.github.io/data_napkin_math/

You will produce a short report (1 page OK, more also OK) that describes some data napkin math estimation about your *project data*.

Please assess, to the best of your ability: - How many hours of human labour were required to create the dataset you are using for your project? - How much money would it cost to "commission" a fresh copy of this dataset (hint: use your hours estimate and make a reasonable guess about hourly costs) - How much money could this dataset generate (hint: make a reasonable guess about this data could be used to make inferences, predictions, detections etc. and what the business value or other value is. The answer might be: not very much!)

You will need to write down a lot of assumptions. You will marked based on your completeness in listing and justifying the assumptions, not the empirical validity of your estimate (i.e., it is better to make wild guesses than to have unexplained details).

If you are in a project group, you may submit this with your group.

4.4 Submission Instructions

You will turn in: a single notebook-style report as a PDF file that fulfills all above criteria. You may use a Jupyter notebook, or a Word/Google doc with key code pasted in.

5 Slides

For now, you can find slides source code and pdfs here.

6 Project Proposal

DEADLINE (Fall 2025): TBA.

We're going to start thinking about our projects relatively early in the term! To scaffold the project ideation, you'll be asked to turn in and present an initial **project proposal** early on in the semester.

You can view some short descriptions of previous projects here.

You can submit a 1-2 page PDF, text, or Markdown file. Exact length is not critical here: as long as it contains the key ideas, you're good to go.

This proposal is not binding, though you will earn some marks for turning it in and presenting it. You can change your project topic, track, or group after the proposal is submitted (though you're encouraged to stick relatively close to your proposal, just for the sake of your own time).

For the project, you can select from three tracks, described below.

Well before you turn your project in, you will be provided with a much more detailed rubric describing how your project will be graded. For the initial proposal, however, you should just focus on selecting a project that:

- fits your personal interests in the course (including your career goals)
- will give you an opportunity to explore and demonstrate understanding of the key concepts from our readings and lectures.

The two heuristic questions I recommend you ask while brainstorming project ideas:

- Does this project meet the unique individual incentives of all group members (e.g., a chance to work with a particular ML library, a chance to work on a task of interest, a chance to produce a high quality report or prototype to include in my portfolio).
- Does this project offer an opportunity to demonstrate understanding of key concepts from the course? For instance, does it fit into any of the frameworks for human-centered ML and AI that we've seen, or does it relate to any of the calls for data-centric we've seen?

6.0.1 Track 1: Tools and interfaces for human/data-centered AI

Track 1 will be a good fit for front-end focused projects. For this track, you can propose and develop some kind of tool or interface for data-centric AI. This interface might be a web application, mobile application, or even a user-focused CLI prototype.

To fit the project criteria, this tool should help users accomplish some kind of data-related action or some kind of data exploration task. In other words, it should either be targeted at users who want to control the flow of their data, or at data scientists who want to explore data in some way.

Please note that if you're very uncomfortable doing prototyping and frontend development, you may not want to select this track. While I'm happy to support you if you want to learn these topics on the fly, we probably won't have much time to cover core design, frontend, or software engineering concepts in this course, so this project is best suited to students who already have some of those skills and specifically want to use their project work time to advance in this area.

Examples:

- A new interface for interacting with large language models that allows user to save or export conversation data (you might consider forking and contributing to something likehttps://github.com/ollama-webui/ollama-webui)
- A browser extension that helpers user collect and use data generated by their own browsing (e.g. export my YouTube watch history and train a local personalization / recommender system)
- A browser extension that blocks data collection and informs the user how data that's collected might impact AI systems
- A web interface for visually exploring aspects of a dataset, aimed at ML developers

6.0.2 Track 2: ML Project with Data Exploration Component

Track 2 will be the closest to what you might do in a typical project-focused ML course. For this project, you should select a machine learning task of interest and produce a thorough report describing how you might tackle the relevant ML challenges. What will set your project apart from a pure ML focused course is that you will also be asked to conduct a data-centric exploration of the task. This might involve using data valuation techniques we learned in the course, exploring different dataset selection choices, etc.

The DataPerf reading will be particularly useful to projects on this track.

Examples:

- You might select a medical imaging dataset from a research lab or research challenge and show how selecting or deselecting certain training observations impact performance on a carefully chosen held out test set
- You might fine-tune an open language model with a variety of different fine-tuning sets and explore the impact on benchmark performance or quality as perceived by humans

6.0.3 Track 3: Dataset Documentation and AI Auditing

Later in the course, we will discuss some research on dataset documentation and AI auditing. To summarize, this work involves carefully scrutinizing existing datasets and/or the outputs of AI systems to check for potential biases, performance gaps, unusual behavior, etc.

As your project, you might pick a famous dataset or AI system and conduct a systematic documentation effort or "audit".

Examples:

- You might select a popular dataset that's been used to train LLMs like ChatGPT and use a mix of manual inspection and ML-powered investigation to try and understand the demographics of dataset contributors, or biases in the underlying the content.
- A fun example of this might involve a question like, "How much do various fandom communities discussing their favorite movie, book, anime, etc." contribute to the success of ChatGPT?

If you wish to pursue this option, please consult with the instructor first to discuss properly scoping this kind of project (obviously, investigating every single piece of training data underlying ChatGPT will not be possible with the time we have).

References:

- BookCorpus datasheet:https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/54229abfcfa5
  Paper-round1.pd
- Mozilla's Common Crawl data investigation:https://foundation.mozilla.org/en/blog/
  Mozilla-Report-How-Common-Crawl-Data-Infrastructure-Shaped-the-Battle-Royale-
  over-Generative-AI/

6.0.4 Mixing the tracks

If you have an idea for a project that involves mixing multiple tracks, that is totally great! Please let us know via the initial proposal draft.

In particular, mixing tracks might make sense if you have a larger group of students who want to work on multiple parts of a particular problem. For instance, if you want to build

a prototype system that hooks up with a ML model and reports the results of a dataset documentation effort, you can definitely do so.

7 Project Rubric

TBA.

References

Chancellor, Stevie. 2023. "Toward Practices for Human-Centered Machine Learning." *Communications of the ACM* 66 (3): 78–85.

Mazumder, Mark, Colby Banbury, Xiaozhe Yao, Bojan Karlaš, William Gaviria Rojas, Sudnya Diamos, Greg Diamos, et al. 2023. "Dataperf: Benchmarks for Data-Centric Ai Development." *Advances in Neural Information Processing Systems* 36: 5320–47.