INF312H1: Worlds become data

Rohan Alexander

2/27/23

Table of contents

1 Preamble

- Course Instructor: Rohan Alexander
- E-mail: rohan.alexander@utoronto.ca

1.1 Overview

To a certain extent we are wasting our time. We have a perfect model of the world–it is the world! But it is too complicated. Because of this we must simplify the world in order for it to become data. In this course we explore how we do this, and the implications.

1.2 FAQ

- Can I audit this course? Sure, but it is pointless, because the only way to learn this stuff is to do the work.
- What is a tutorial? You write a paper. Then you send it to your tutor. The next day you have a meeting, 'tutorial', where you discuss it with them.
- Why is there so much assessment? The only way to learn this stuff is to actually do the work, and students only do the work when they are assessed. It is unfortunate, but there is no way around it.
- How difficult is the course? The course is not difficult, but the hands-on-projects mean it is a lot of work.

- What is the format of the class? There are rarely old-school lectures because those are not effective. You should read the relevant chapter before class. During class we will focus on tutorials and discussion. We will also have industry guests discuss their experience.

1.3 Syllabus

- 2023 - TBA
- 2022.

1.4 Pre-requisites

- None.

1.5 Textbook

Telling Stories with Data

2 Content

2.1 Week 1

- Drinking from a fire hose
  - The statistical programming language R enables us to tell interesting stories using data. It is a language like any other, and the path to mastery can be slow.
  - The framework that we use to approach projects is: plan, simulate, gather, explore, and share.
  - The way to learn R is to start with a small project and break down what is required to achieve it into tiny steps, look at other people's code, and draw on that to achieve each step. Complete that project and move onto the next project. Each project you will get a little better.
  - The key is to start actively working regularly.

2.2 Week 2

- Reproducible workflows
  - Reproducibility typically begins as something that someone imposes on you. It can be onerous and annoying. This typically lasts until you need to revisit a project after a small break. At that point you typically realize that reproducibility is not just a requirement for data science because it is the only way that we can make genuine progress, but because it actually helps ourselves.
  - Essentially reproducibility implies sharing data, code, and environment. And this is enhanced by using Quarto, R Projects, and Git and GitHub. We use Quarto to build documents that integrate normal text and R code. R Projects enable a file structure that is not dependent on a directory set-up that is user-specific. And Git and GitHub make it easier to share code and data.
  - This is not an unimpeachable workflow, but one that is good enough and provides many of the benefits. We will improve various aspect of it through various tools, but improving code structure and comments goes a long way.
  - As we go through writing R code to come to understand the implications of some dataset, there is typically an awful lot of back-and-forth. This implies that we should restart R often ("Session" -> "Restart R and Clear Output"). There are always errors that occur, and it is important to recognize that debugging is a skill that improves with practice. But one key aspect of being able to get help is to be able to make a reproducible example that reproduces the issue for others.

2.3 Week 3

- R essentials
  - Understanding foundational aspects of R and RStudio enables a gradual improvement of workflows. For instance, being able to use key `dplyr` verbs and make graphs with `ggplot2` makes manipulating and understanding datasets easier.
  - But there is an awful lot of functionality in the `tidyverse` including importing data, dataset manipulation, string manipulation, and factors. You do not need to know it all at once, but it is important to know that you do not yet know it.
  - Beyond the `tidyverse` it is also important to know that foundational aspects, common to many languages, exist and can be added to data science workflows. For instance, class, functions, and data simulation all have an important role to play.

2.4 Week 4

- Writing research

  - Writing is a key skill, perhaps the most important skill, of all the skills required of a data scientist. The only way to get better at writing, is to write, ideally every day.
  - When we write, although the benefits typically accrue to ourselves, we must nonetheless write for the reader. This means having one key message that we want to communicate, and thinking about where they are, rather than where we are.
  - The key is that we get to a first draft as quickly as possible. Even if it is horrible, the difference between a first draft existing and not is enormous. At that point we start to rewrite brutally and removing as many words as possible.
  - We typically begin with some area of interest, and then develop research questions and data in an iterative way. Even as we are writing our research, we are coming to a better understanding of what we are doing.

2.5 Week 5

- Static communication
  - We must show the reader the actual observations in the dataset, or as close as is possible, through graphs and tables. This is because it is only through visualization that we can get a true sense of our data. This means that we need to develop a comfort with a variety of graph options, including: bar charts, scatterplots, line plots, and histograms. We can even consider a map to be a type of graph, especially after geocoding our data.
  - That said, we also must know when to summarize data, for instance using tables. Typical use cases for this include showing part of a dataset, summary statistics, and regression results.

2.6 Week 6

- Farm data
  - Before there can be a dataset, there must be measurement, and this brings a whole host of challenges and concerns. One dataset that is designed to be complete, at least in certain respects, is a census. While not perfect, governments spend a lot of money on censuses and other official statistics, and they are a great foundational data source.
  - However even when we cannot obtain such a dataset, we can use sampling to ensure that we can still make sensible claims. There are two varieties of this—probability and non-probability. Both have an important role. Key terminology and concepts include: target population, sampling frame, sample, simple random sampling, systematic sampling, stratified sampling, and cluster sampling.

2.7 Week 7

- Gather data
  - Sometimes data are available, but they are not necessarily put together for the purposes of being a dataset. We have to go and gather such data.
  - It can be cumbersome and annoying to have to clean and prepare the datasets that come from these unstructured sources, however, the resulting structured, tidy, data are often especially exciting and useful.
  - We can gather data from a variety of sources, including APIs, both directly, including dealing with semi-structured data, and indirectly through R Packages. We can also gather data through web scraping, although it is important to consider reasonable use and ethical concerns. Finally, we may wish to gather data from PDFs, possibly even needing to OCR them.

2.8 Week 8

- Hunt data
  - Establishing treatment and control groups using randomization to estimate average treatment effects and understanding threats to these estimates.
  - Understanding the requirements and implications of internal and external validity.
  - Appreciating why informed consent and establishing the need for an experiment are critical.
  - A/B testing and some of its nuances.
  - Designing and implementing surveys.

2.9 Week 9

- Clean and prepare
  - Cleaning and preparing a dataset is difficult work that involves a great deal of decision-making. Planning an endpoint and simulating the dataset that we would like to end up with are key elements of cleaning and preparing data.
  - It can help to work in an iterative way, beginning with a small sample of the dataset. Write code to fix some aspect, and then iterate and generalize to additional tranches.
  - During that process we should also develop a series of tests and checks that the dataset should pass. This should focus on key features that we would expect of the dataset.
  - We should be especially concerned about the class of variables, having clear names, and that the unique values of each variable are as expected given all this.

2.10 Week 10

- Store and share
  - The FAIR principles provide the foundation from which we consider data sharing and storage. These specify that data should be findable, accessible, interoperable, and reusable.
  - The most important step is the first one, and that is to get the data off our local computer, and to then make it accessible by others. After that, we build documentation, and datasheets, to make it easier for others to understand and use it. Finally, we ideally enable access without our involvement.
  - At the same time as wanting to share our datasets are widely as possible, we must respect those whose information are contained in them. This means, for instance, protecting, to a reasonable extent, and informed by costs and benefits, personally identifying information through selective disclosure, hashing, data simulation, and differential privacy.
  - Finally, as our data get larger, approaches that were viable when they were smaller start to break down. We need to consider efficiency with regard to our data, and explore other approaches, formats, and languages.

2.11 Week 11

- Exploratory data analysis
  - Exploratory data analysis is the process of coming to terms with a new dataset by constructing graphs and tables. We want to explore and understand three critical aspects: 1) each individual variable by itself; 2) each individual in the context of other, relevant, variables; and 3) the data that are not there.
  - During the EDA process we want to come to understand the issues and features of the dataset and how this may affect analysis decisions. We are especially concerned about missing values and outliers.

2.12 Week 12

TBD

3 Assessment

3.1 Summary

Item	Weight (%)	Due date
Quiz	8	Weekly, end of each week
Personal website	1	Friday, noon, Week 11
SQL quiz	1	Friday, noon, Week 11
Tutorial	10	Weekly, end of each week
Paper 1	25	Friday, noon, Week 3
Paper 2	25	Friday, noon, Week 6
Paper 3	25	Friday, noon, Week 8
Paper 4	25	Friday, noon, Week 10
Final Paper (initial submission)	2	Wednesday, noon, Week 12
Conduct peer review	3	Friday, noon, Week 12
Final Paper	25	Two weeks after that

You must submit Paper 1. You must submit the Final Paper. You must submit and get at least 70 per cent on both the SQL quiz and the Personal website.

Beyond that, you have scope to pick an assessment schedule that works for you. We will take your best three of the twelve tutorials for that 10 per cent, and your best five of twelve quizzes for that 10 per cent. And we will take your two best papers from Papers 1-4 for that 50 per cent (25 per cent for each). The remainder is made up of 2 per cent for submitting a draft of the Final Paper, 3 per cent for conducting peer review of other people's drafts of the Final Paper, and 25 per cent for the Final Paper.

Additional details:

- Quiz questions are drawn from those in the Quiz section that follows each chapter of *Telling Stories with Data*. Some of them are multiple choice, and you should expect to know the mark within a few days of submission.
- Tutorial questions are drawn from those in the Tutorial section that follows each chapter of *Telling Stories with Data*. The general expectation (although this differs from week to week) is about two pages of written content, which the tutor will read, discuss with you, and then provide a mark. You should expect to know the mark within a few days of the tutorial.
- In general papers require a considerable amount of work, and are due after the material has been covered in quizzes and tutorials (i.e. you would draw on knowledge tested in the quizzes, and potentially material could be re-used from the tutorial material). In general, they require original work to some extent. Papers are taken from the Papers appendix of *Telling Stories with Data* and students have access to the grading rubrics before submission.

3.2 Quiz

- Due date: Friday, noon, weekly (with grace period through to Sunday, midnight, to submit without penalty).
- Weight: 8 per cent. Only best five out of twelve count.
- Task: Please complete a weekly quiz.

3.3 SQL quiz

- Due date: Available from Week 1, but due Friday, noon, Week 11 (with grace period through to Sunday, midnight, to submit without penalty).
- Weight: 1 per cent. You cannot pass the course if you do not get at least 70 per cent in this quiz.
- Task: Please complete a quiz about SQL.

3.4 Personal website

- Due date: Available from Week 1, but due Friday, noon, Week 11 (with grace period through to Sunday, midnight, to submit without penalty).
- Weight: 1 per cent. You cannot pass the course if you do not get at least 70 per cent on this assessment.
- Task: Please create a personal website using Quarto and make it live via GitHub Pages. At a minimum, it must include a bio and a CV in PDF form.

3.5 Tutorial

- Due date: Friday, noon, weekly (with grace period through to Sunday, midnight, to submit without penalty).
- Weight: 10 per cent. Only best three out of twelve count.
- Task: Please complete a tutorial question.
- Rubric:
  - 0 - Any typos, major grammatical errors, other table stakes issues for this level. Too short.
  - 0.25 - Grammatical errors, if relevant: tables/graphs not properly labeled, no references, other aspects that affect credibility.
  - 0.6 - Makes some interesting and relevant points, related to course material (including required materials), but lacking in terms of structure and story/argument.
  - 0.80 - Interesting paper that is well-structured, coherent, and credible.
  - 1 - As with 0.80, but exceptional in some way.

3.6 Paper #1

- You must submit this paper.
- Task: Donaldson Paper
- Due date: Friday, noon, Week 3 (with grace period through to Sunday, midnight, to submit without penalty).
- Weight: 25 per cent (for Papers #1-#4 the best two of four count).

3.7 Paper #2

- Due date: Friday, noon, Week 6 (with grace period through to Sunday, midnight, to submit without penalty).
- Task: Mawson Paper
- Weight: 25 per cent (for Papers #1-#4 the best two of four counts).

3.8 Paper #3

- Due date: Friday, noon, Week 8 (with grace period through to Sunday, midnight, to submit without penalty).
- Task: Howrah Paper
- Weight: 25 per cent (for Papers #1-#4 the best two of four counts).

3.9 Paper #4

- Due date: Friday, noon, Week 10 (with grace period through to Sunday, midnight, to submit without penalty).
- Task: Dysart Paper or Spofforth Paper
- Weight: 25 per cent (for Papers #1-#4 the best two of four counts).

3.10 Final Paper

- Task: Final Paper
- You must submit this paper.
- Due dates:
  - Initial submission: Wednesday, noon, Week 12 (no grace period and no late submissions accepted).
  - Conduct peer review: Friday, noon, Week 12 (no grace period and no late submissions accepted).

    - Final Paper: Two weeks after that (with grace period through to Sunday, midnight, to submit without penalty).

- Weight: 30 per cent
    - Initial submission: 2 per cent
    - Conduct peer review: 3 per cent
    - Final Paper: 25 per cent

4 Other

4.1 Children in the classroom

Babies (bottle-feeding, nursing, etc) are welcome in class as often as necessary. You are welcome to take breaks to feed your infant or express milk as needed, either in the classroom or elsewhere including here. A list of baby change stations is also available here. Please communicate with me so that I can make sure that we have regular breaks to accommodate this.

For older children, I understand that unexpected disruptions in childcare can happen. You are welcome to bring your child to class in order to cover unforeseeable gaps in childcare.

4.2 Accommodations with regard to assessment

Please do **not** reveal your personal or medical information to me. I understand that illness or personal emergencies can happen from time to time. The following accommodations to assessment requirements exist to provide for those situations.

Straight-forward (will automatically apply to all students - there's no need to ask for these):

- Quiz: Only best five quizzes count.
- Tutorial: Only best three tutorials count.
- Papers #1-#4: Worst two are dropped.

So for those (with the exception of Paper #1), if you have a situation, then just don't submit.

Slightly more involved:

- Paper #1: You must submit something for Paper #1, even if it gets zero. If you have a medical reason that makes it impossible for you to submit Paper #1, then you are welcome to continue with the class, but then one of the remaining term papers (Papers #2 - #4), must be done individually to ensure fairness with the rest of the class.

- Peer review: No accommodation or late submission is possible for this because it would hold up the rest of the class. If you cannot submit then email me before the deadline and the weight will be shifted to the final paper.
- Final paper: The final paper is a critical piece of assessment. It is also up against deadlines for submission of grades. Extensions for valid reasons may be granted for a maximum of three days, however this isn't possible for all students (i.e. there are restrictions around graduating students). This means the exact extension needs to be at my discretion. To be considered, an extension request must be sent to rohan.alexander@utoronto.ca by the business day before the due date so there is time to get advice from a faculty/department/college advisor about your particular circumstance.

4.3 Re-grading

Requests to have your work re-graded will not be accepted within 24 hours of the release of grades. This is to give you a chance to reflect. Similarly, requests to have your work re-graded more than seven days after the release of the grades will not be accepted. This is to ensure the course runs smoothly.

Inside that 1-7 day period if you would like to request a re-grade, please email rohan.alexander@utoronto.ca. Please specify where the marking error was made in relation to the marking guide. The entire assessment will be re-marked and it is possible that your grade could reduce.

Plenty of students get 0 on the first paper, but go on to get an A+ overall in the course. The nature of the work in this course requires students to adjust from what is expected in other courses, and the forgiving assessment weighting is designed to allow this.

4.4 Plagiarism and integrity

Please do not plagiarize. In particular, be careful to acknowledge the source of code - if it is extensive then through proper citation and if it is just a couple of lines from Stack Overflow then in a comment immediately next to the code.

You are responsible for knowing the content of the University of Toronto's Code of Behaviour on Academic Matters.

Academic offenses includes (but is not limited to) plagiarism, cheating, copying R code, communication/extra resources during closed book assessments, purchasing labor for assessments (of any kind). Academic offenses will be taken seriously and dealt with accordingly. If you have any questions about what is or is not permitted in this course, please contact me.

Please consult the University's site on Academic Integrity. Please also see the definition of plagiarism in section B.I.1.(d) of the University's Code of Behaviour on Academic Matters available here. Please read the Code. Please review Cite it Right and if you require further clarification, consult the site How Not to Plagiarize.

4.5 Late policy

You are expected to manage your time effectively. If no extension has been granted and no accommodation applies, then the late submission of an assessment item carries a penalty of 10 percentage points per day to a maximum of one week after which it will no longer be accepted, e.g. a problem set submitted a day late that would have otherwise received 8/10 will receive 7/10, if that same problem set was submitted two days late then it would receive 6/10.

4.6 Writing

Papers and reports should be well-written, well-organized, and easy to follow. They should flow easily from one point to the next. They should have proper sentence structure, spelling, vocabulary, and grammar. Each point should be articulated clearly and completely without being overly verbose. Papers should demonstrate your understanding of the topics you are studying in the course and your confidence in using the terms, techniques and issues you have learned. As always, references must be properly included and cited. If you have concerns about your ability to do any of this then please make use of the writing support provided to the faculty, colleges and the SGS Graduate Centre for Academic Communication.

4.7 Minimum submission requirement

If you are going to not be able to submit at least two term papers, and/or be unable to submit the final paper then it would be unfair on the other students to allow you to pass the course. Please ensure you and your college registrar or faculty/department advisor get in touch with me as early as possible if this may be the case for you.

4.8 Course Learning Outcomes

1. Critique the major methodological and computational challenges related to the creation and representation of structured and unstructured data (All papers and quizes)
2. Analyze, and synthesize the major ethical debates and recurring patterns pertaining to data stewardship and circulation (All papers)
3. Develop innovative research design to overcome the methodological, sociopolitical and ethical challenges associated with current methods of analysis of complex information practices involving structured and unstructured data. (All papers)

4. Communicate and present research proposal to overcome the methodological, sociopolitical and ethical challenges associated with current methods of analysis of complex information practices. (Papers and tutorials)

These will be measured through assignments that support understanding of the methodological and computational debates related to data creation and representation (CLO #1) and the ethical debates pertaining to data stewardship and circulation (CLO#2), reflective papers that help develop and relate insights regarding the methodological, sociopolitical and ethical challenges associated with current methods of analysis of complex information practices involving structured and unstructured data (CLO#1, CLO#2, CLO#3), and produce work that help relate their reflections in textual and non-textual forms. (CLO #4.)

4.9 Relation to Program Learning Outcomes

Huge amounts of data are produced everyday including different types of data such as structured quantifiable data, unstructured text, and multimedia data. Studying the social world through shaping it into one of the above-mentioned data types has its own advantages and limitations. This course will help students to understand and assess the social, political, economic, and ethical entailments of information creation, ownership, stewardship, and circulation, especially in light of enduring and emerging ethical and political questions (PLO1r). Given the case studies applied in class and practical assignments in this course, the students will be able to critique the conceptual and philosophical foundations of representation and computation, and to recognize recurring patterns of unresolved intellectual and social tension (PLO3i and PLO11r). With the knowledge acquired in this course, the students will be able to develop, defend, and use methods of analysis of complex information practices and the political, economic, technical, and cultural contexts in which they occur (PLO10i).