Worlds become data

Rohan Alexander

1 February 2022

Contents

1 Preamble **2**
1.1 Overview . . . 2
1.2 FAQ . . . 2
1.3 Acknowledgements . . . 2

2 Content **2**
2.1 Week 1 . . . 2
2.2 Week 2 . . . 2
2.3 Week 3 . . . 2
2.4 Week 4 . . . 3
2.5 Week 5 . . . 3
2.6 Week 6 . . . 3
2.7 Week 7 . . . 3
2.8 Week 8 . . . 3
2.9 Week 9 . . . 3
2.10 Week 10 . . . 3
2.11 Week 11 . . . 3
2.12 Week 12 . . . 4

3 Assessment **4**
3.1 Summary . . . 4
3.2 Quiz . . . 4
3.3 Tutorial . . . 5
3.4 Paper #1 . . . 5
3.5 Paper #2 . . . 5
3.6 Paper #3 . . . 5
3.7 Paper #4 . . . 5
3.8 Final Paper . . . 5

4 Other **6**
4.1 Children in the classroom . . . 6
4.2 Description . . . 6
4.3 Learning objectives . . . 6
4.4 Communication . . . 6
4.5 Accommodations with regard to assessment . . . 6
4.6 Minimum submission requirement . . . 7
4.7 Re-grading . . . 7
4.8 Plagiarism and integrity . . . 7
4.9 Late policy . . . 8
4.10 Writing . . . 8

4.11 Accessibility needs . . . 8
4.12 Intellectual Property Statement . . . 8
4.13 Course Learning Outcomes and their Relationship with Assessment . . . 8
4.14 Grading . . . 9
4.15 Writing support . . . 9

1 Preamble

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.

This course is different to many other courses at the University of Toronto. At the end of this course, you will have a portfolio of work that you could show off to a potential employer. You will have developed the skills to work successfully as an applied statistician or data scientist. And you will know how to fill gaps in your knowledge yourself. A lot of scholarships and jobs these days ask for GitHub and blog links etc to show off a portfolio of your work. This is the class that gives you a chance to develop these. It's very important to having something to show that needs to go beyond what is done in a normal class.

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 is to actually do the work, and students only do the work when they are assessed. It is unfortunate, but there's no way around it.

1.3 Acknowledgements

Thank you to the following people for generously providing comments, references, suggestions, and thoughts that directly contributed to this outline: Monica Alexander and Uzair Mirza.

2 Content

Almost all content will closely follow *Telling Stories with Data*.

2.1 Week 1

- Content:
  - Introduction
  - Several end-to-end worked examples

2.2 Week 2

- Content:
  - R Essentials

2.3 Week 3

- Content:

  - Reproducible workflow

2.4 Week 4

- Content:
  - Writing
  - Static communication
- Recording: https://youtu.be/WvcnR-GPg80

2.5 Week 5

- Content:
  - Interactive communication

2.6 Week 6

'Gathering data'.

- Content:
  - Using APIs, scraping.

2.7 Week 7

'Gathering data II'.

- Content:
  - OCR, semi-structured datasets, and text.

2.8 Week 8

'Hunting data'.

- Content:
  - Experiments, sampling,

2.9 Week 9

'Hunting data II'.

- Content:
  - Surveys, and A/B testing.

2.10 Week 10

'Cleaning data'.

- Content:
  - Workflow for cleaning data.
  - Effective naming, checks, and testing.

2.11 Week 11

'Store, retrieve, disseminate and protect' and 'share, but not too much'.

- Content:
  - R packages for data, and documentation including datasheets.

  - Personally identifying information, hashing and salting, GDPR and HIPPA, simulated data, and differential privacy.

2.12 Week 12

'Whoops, I forgot EDA'.

- Content:
  - Coming to terms with a dataset and understanding what is in it.

3 Assessment

3.1 Summary

Item	Weight (%)	Due date
Quiz	20	Weekly before the lecture
Tutorial	20	Weekly the day before the tutorial
Paper 1	25	End of Week 4
Paper 2	25	End of Week 6
Paper 3	25	End of Week 8
Paper 4	25	End of Week 10
Final Paper (initial submission)	1	Middle of Week 12
Final Paper (peer review)	4	End of Week 12
Final Paper	25	Two weeks after that

You must submit Paper 1. And you must submit the Final Paper.

'End of' means Sunday 11:59pm.

Beyond that, you have scope to pick an assessment schedule that works for you. We will take your best 3 of the 11 tutorials, or your best 8 of 11 quizzes for that 20 per cent—whichever results in a better grade for you (i.e. you can choose to do either quizzes or tutorials). 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 1 per cent for submitting a draft of the Final Paper, 4 per cent for peer reviewing 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*. Almost all of them are multiple choice, and you should expect to know the mark within two 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 three 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

- You should choose to do either tutorials or quizzes.

- Due date: Weekly before the lecture.
- Weight: 20 per cent. Only best eight out of eleven count and only if that is better for you than counting tutorials.
- Task: Please complete a weekly quiz in Quercus.

3.3 Tutorial

- You should choose to do either tutorials or quizzes.
- Due date: Weekly the day before the tutorial.
- Weight: 20 per cent. Only best three out of eleven count and only if that is better for you than counting quizzes.
- Task: Please complete a tutorial question and submit it via Quercus.
- 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.

3.4 Paper #1

- You must submit this paper.
- Task: 'Mandatory Minimums' (details will be added to Quercus).
- Due date: End of Week 4.
- Weight: 25 per cent (for Papers #1-#4 the best two of four count).

3.5 Paper #2

- Due date: End of Week 6.
- Task: 'The Short List' (details will be added to Quercus).
- Weight: 25 per cent (for Papers #1-#4 the best two of four counts).

3.6 Paper #3

- Due date: End of Week 8.
- Task: TBA (details will be added to Quercus).
- Weight: 25 per cent (for Papers #1-#4 the best two of four counts).

3.7 Paper #4

- Due date: End of Week 10.
- Task: TBA (details will be added to Quercus).
- Weight: 25 per cent (for Papers #1-#4 the best two of four counts).

3.8 Final Paper

- Task: TBA (details will be added to Quercus).
- You must submit this paper.
- Due dates:
  - Initial submission: Middle of Week 12.
  - Peer review: End of Week 12.
  - Final Paper: Two weeks after that.

- Weight: 30 per cent
  - Initial submission: 1 per cent
  - Peer review: 4 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: https://familycare.utoronto.ca/childcare/breastfeeding-at-u-of-t/. A list of baby change stations is also available: https://familycare.utoronto.ca/childcare/baby-change-stations-at-u-of-t/. 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 Description

This course covers issues in the practices of translating phenomena to data and algorithmic description. What happens, what is gained, what is lost, when things that happen in the world are recorded and made into information or recorded as a document? The course explores representation, modeling, correctness, reliability, and bias in different types of data and algorithms. We will learn about diverse topics such as cultural and algorithmic bias, challenges of big data, what happens when the world is transformed into images, what are the implications of having your social status determined by data and scores on your social media profile, and what we gain or miss when we deal with geographical information systems.

4.3 Learning objectives

- Design a survey or sample that appropriately gathers information of interest.
- Carry out a variety of statistical analyses in R to make inference on the data collected from a survey/sample.
- Identify and implement different sampling techniques and different study designs and the trade-offs involved in each.
- Identify sources of bias within a study and comment on a study's design, including weaknesses, strengths, and appropriate analyses.
- Clearly communicate results of statistical analyses to technical and non-technical audiences.

4.4 Communication

If you have a question, there is a decent chance that others have the same question or, at least, will benefit from the answer. Please post all questions to Piazza so that everyone in the course can benefit from your questions and our answers. You are encouraged to post answers to the questions of other students, where appropriate. Of course, if you have a concern of a personal nature then please email the TAs or me and you should begin your subject line with the course code 'INF312', and then an appropriate subject.

Emails and the message board are not checked or responded to by either the TA or me after hours or on the weekend.

Please be polite. We continue to be in a pandemic.

4.5 Accommodations with regard to assessment

You do **not** need to 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: Worst three quizzes are dropped.
- Tutorial: Worst eight tutorials are dropped.
- 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: I'm open to a day without penalty to account for situations. Beyond that it begins to slow down the class. You must submit something for Paper #1.
- 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's 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 may be restrictions around graduating students). Hence, 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 the Department and your college about your particular circumstance.

4.6 Minimum submission requirement

If you are going to not be able to submit at least two problem sets, 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 registrar get in touch with me as early as possible if this may be the case for you.

4.7 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 with a subject line that starts with 'INF312'. You must specify where the marking error was made in relation to the marking guide. Your entire assessment will be re-marked and it is possible that your grade could reduce.

4.8 Plagiarism and integrity

Please do not plagiarize. In particular, be careful to acknowledge the source of code - if it's extensive then through proper citation and if it's 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 include (but are not limited to) plagiarism, cheating, copying R code, communication/extra resources during closed book assessments, purchasing labour 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 http://academicintegrity.utoronto.ca/. Please also see the definition of plagiarism in section B.I.1.(d) of the University's Code of Behaviour on Academic Matters http://www.governingcouncil.utoronto.ca/Assets/Governing+Council+Digital+Assets/Policies/

PDF/ppjun011995.pdf. Please read the Code. Please review Cite it Right and if you require further clarification, consult the site How Not to Plagiarize http://advice.writing.utoronto.ca/wp-content/uploads/sites/2/how-not-to-plagiarize.pdf.

4.9 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.

4.10 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.

4.11 Accessibility needs

Students with diverse learning styles and needs are welcome in this course. In particular, if you have a disability/health consideration that may require accommodations, please feel free to approach me and/or Accessibility Services at 416 978 8060 or visit studentlife.utoronto.ca/as.

4.12 Intellectual Property Statement

Course material that has been created by your instructor is the intellectual property of your instructor and is made available to you for your personal use in this course. Sharing, posting, selling, or using this material outside of your personal use in this course is not permitted under any circumstances and is considered an infringement of intellectual property rights.

4.13 Course Learning Outcomes and their Relationship with Assessment

Students who have successfully completed this course will:

1. Know how to how to clean and prepare a new dataset and quickly generate summary statistics, tables, and graphs. *Demonstrated through completion of problem sets.*
2. Be able to identify ethical considerations and adjust approaches accordingly. *Demonstrated through all assessment items.*
3. Create graphs and tables that help to illustrate and support claims. *Demonstrated through completion of all aspects of assessment.*
4. Conduct statistical analysis in a reproducible way and interpret results. *Demonstrated through completion of all aspects of assessment, especially the Final Paper*
5. Communicate results in a way that is clear about their strengths, weaknesses, and the assumptions that underpin them. *Demonstrated through completion of all aspects of assessment, especially the Final Paper*
6. Discuss the importance of openness in science, of making their analysis and datasets public. *Demonstrated through completion of all aspects of assessment.*

4.14 Grading

Please consult: - the Faculty of Information's Grade Interpretation Guidelines: https://ischool.utoronto.ca/wp-content/uploads/2016/11/grade-interpretation.pdf
- The University Assessment and Grading Practices Policy: http://www.governingcouncil.utoronto.ca/Assets/Governing+Council+Digital+Assets/Policies/PDF/grading.pdf - The Guidelines on the Use of INC, SDF, & WDR: http://www.sgs.utoronto.ca/facultyandstaff/Pages/INC-SDF-WDR.aspx

These documents will form the basis for grading in the course.

4.15 Writing support

As stated in the Faculty of Information's Grade Interpretation Guidelines, 'work that is not well written and grammatically correct will not generally be considered eligible for a grade in the A range, regardless of its quality in other respects.' With this in mind, please make use of the writing support provided to graduate students by the SGS Graduate Centre for Academic Communication. The services are designed to target the needs of both native and non-native speakers and all programs are free. Please consult the current workshop schedule http://www.sgs.utoronto.ca/currentstudents/Pages/Current-Years-Courses.aspx for more information.