INF2178H: Experimental Design for Data Science

Syllabus

Instructor: Shion Guha
Teaching Assistants: Dashiel Carrerra (Tutorials), Rohith Sothilingam (Projects)
Office: N/A
E-mail:
- **Shion:** shion.guha@utoronto.ca
- **Dash:** dcarrera@cs.toronto.edu
- **Rohith:** rohith.sothilingam@mail.utoronto.ca

Course Web Page: https://q.utoronto.ca/courses/255792/
Github Repository: https://github.com/shionguha/inf2178j-w22-exp-design-datascience
Office hours: Wednesday 4:00pm – 5:00pm (Shion); by appointment (Dash/Rohith)
Location: Synchronous Online in January 2022, followed by in-person February 2022 onwards

Class Times:
Lecture: Monday 1:00pm – 4:00pm
Tutorials:
- 101: Tuesday 10 am – 11 am;
- 102: Tuesday 11 am – 12 pm;
- 103: Wednesday 4 pm – 5 pm;
- 104: Wednesday 5 pm – 6 pm;

Course Description:

At the heart of every Data Science project exists the planning, design and execution of experiments. Such experiments aim at understanding the data, potentially cleaning it and performing the necessary data analysis for knowledge discovery and decision-making. Without knowing the experimental design processes that are used in practice, researchers may not be able to discover what is really hidden in their data. The first aim of this course is to look at existing experimental designs that take into account the questions that need to be answered as well as the nature of the data and the different parameters used by algorithms. Subsequently, the course will introduce different qualitative and quantitative methods to assess the quality of the results.

All concepts will be accompanied by examples and the students will have practical exercises and projects in which they will demonstrate their knowledge.

Prerequisites:

- INF 1344H: Introduction to Statistics for Data Science

- INF 1340: Programming for Data Science.

Course Structure

Three (3) in-class hours per week will be divided into lectures and tutorials, in which we discuss and further probe topics covered in the lectures and readings. Note that for every one (1) hour of class time, students can expect to do 3 hours of reading and preparation work on their own, outside class.

Groups: In order to simulate how data science projects are carried out in industry in reality, students will be assigned to groups in the first couple of weeks of the course. These groups will work on the completion of all assignments related to the semester long group project.

All coursework will be available on Quercus with detailed instructions and submission deadlines (date/time). There will also be an announcement section, which students should be responsible for checking regularly.

Topics:

This course will follow the best principles in human-centered data science as we seek to understand the relationship between research questions, hypothesis testing and experimental design from a quantitative perspective. From a more qualitative perspective, we will discuss, critique and identify challenges and opportunities in modern experimental design, especially those conducted in industry including big tech, healthcare and others. In addition, we will simulate data science project teams by having student groups do a semester long project centered around data science with actual human/social impact. Some suggested topics are:

- Research questions and research design.
- Statistical hypothesis testing
- Linear and logistic regression
- ANOVA
- Introduction to sampling and experiments
- Experimental contrasts
- Block experimental designs
- Factorial experimental design
- Within-subjects experimental design

Learning Objectives:

Upon successful completion of the Experimental Design course, students will be able to:

1. Understand the importance of experimental design and the use of statistics in data science projects

2. Become familiar with the typical experimental designs used in practice

3. Choose an appropriate experimental design based on the given data science project

4. Execute a selected experimental design

5. Analyze the results collected after performing experiments

6. Interpret the experimental results and produce corresponding reports

Relationship to Master of Information (MI) Program-Level Student Learning Outcomes:

Master of Information Program-Level Student Learning Outcomes can be found here.

Conducting experiments is an integral part of every Data Science project. The students in INF2178H will be exposed to the main steps and theoretical foundations of defining the appropriate steps for performing successful experiments (**Outcome 1**). Apart from conducting qualitative experiments, the practical examples of the course will combine theoretical foundations with practical approaches, such that the students can respond to the changing parameters as well as the size and variety of the given data (**Outcome 4**). By employing statistical metrics and evaluation criteria they will be able to provide robust qualitative and quantitative interpretations of experimental results. At the same time they will learn and apply the principles of providing reproducible solutions (**Outcome 5**). Finally, the course will allow students to develop their own goals and continue in life-long intellectual growth beyond graduation (**Outcome 6**).

Class Format

The course will consist of lectures, class discussions, and tutorials. Students are expected to attend the classes and to actively participate in the discussions and tutorials. For each class, a series of topics are provided to guide students through the readings and activities, and to frame the lectures, discussions, and studios.

Teaching and learning is a shared responsibility, influenced by individual knowledge and experience, and achieved through expanding our awareness of the different issues and approaches involved in information architecture. Commitment, preparation, and active participation are important ingredients to realize this goal. Your preparation and participation are important to your learning and the learning of your colleagues.

All the course materials will be available on the University of Toronto learning management system (Quercus) together with assignments and announcements.

Readings:

Book: Seltman, H. J. (2018). Experimental design and analysis. *Department of Statistics at Carnegie Mellon (Online Only)*. URL: https://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf

Weekly Readings for Lecture, Reading Responses and Tutorials:

Day	Assigned Reading
Jan 17	• Seltman, Chapter 1: Introduction to Experimental Design • Esther Duflo. "How to Find the Right Questions (Links to an external site.)." 2019. • Paul J. DiMaggio. "Four Mechanisms for Finding (and Being Found by) Research Problems. (Links to an external site.)" *Sociologica*, 2018.
Jan 24	• Seltman, Chapter 3: Review of Probability • Leo Breiman. "Statistical Modeling: The Two Cultures (Links to an external site.)." *Statistical Science*, 2001. • Mario L. Small and Devah Pager. "Sociological Perspectives on Racial Discrimination (Links to an external site.)." *Journal of Economic Perspectives*, 2020.
Jan 31	• Seltman, Chapter 4: Exploratory Data Analysis for Experiments • Abigail Z. Jacobs and Hanna Wallach. "Measurement and Fairness (Links to an external site.)." FAccT 2020. • Graham Scambler. "Covid-19 as a 'Breaching Experiment': Exposing the Fractured Society (Links to an external site.)." *Health Sociology Review*, 2020.
Feb 7	• Seltman, Chapter 7: One Way ANOVA • Michelle N. Meyer et al. "Objecting to Experiments that Compare Two Unobjectionable Policies or Treatments (Links to an external site.)." *Proceedings of the National Academy of Sciences*, • Angus Deaton, "Randomization in the Tropics Revisited: A Theme and Eleven Variations. (Links to an external site.)"
Feb 14	• Seltman, Chapter 7: ANOVA continued • Chelsea Barabas et al. "Studying Up: Reorienting the Study of Algorithmic Fairness Around Issues of Power (Links to an external site.)." FAccT, 2020. • Rediet Abebe et al. "Roles for Computing in Social Change. (Links to an external site.)" FAccT, 2020.
Feb 28	• Seltman, Chapter 10: One Way ANCOVA • Abeba Birhane and Jelle Van Dijk. "A Misdirected Application of AI Ethics (Links to an external site.)." *NOEMA*, 2020. • Nani Jansen Reventlow. "Data Collection is Not the Solution for Europe's Racism Problem (Links to an external site.)." *Al Jazeera*, 2020.
Mar 7	• Seltman, Chapter 11 Two Way ANCOVA • Patrick Ball. "Violence in Blue (Links to an external site.)." *Granta*, 2020. • Mimi Onuoha. "When Proof is Not Enough (Links to an external site.)." FiveThirtyEight, 2020.
Mar 14	• Seltman, Chapter 12: Statistical Power • "Facebook Manipulated User News Feeds To Create Emotional Responses (Links to an external site.)", *Forbes*, 2014. • Why Stanford Researchers Tried to Create a 'Gaydar' Machine (Links to an external site.)," *The New York Times*
Mar 21	• Seltman, Chapter 14: Within Subjects Experiments • Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of economic behavior & organization*, *81*(1), 1-8.

	• Loftus, G. R., & Masson, M. E. (1994). Using confidence intervals in within-subject designs. *Psychonomic bulletin & review*, *1*(4), 476-490.
Mar 28	• Seltman, Chapter 8: Biases and Threats to Experiments • Gelman, A., & Loken, E. (2014). The statistical crisis in science: data-dependent analysis--a" garden of forking paths"--explains why many statistically significant comparisons don't hold up. *American scientist*, *102*(6), 460-466. • Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, *22*(11), 1359-1366.

Deliverables and Evaluation:

Evaluations	Due Date	Weight
Class Performance and Participation	Mondays 1-4 PM	20%
Reading Responses: Most weeks students will be asked to submit a reading response by Tuesday 8 am EST	Mondays 12 PM	20%
Tutorial attendance and discussions	Tuesdays and Wednesdays at tutorial times	20%
Mid Term Project Submission	TBD (February-March)	20%
Final Project Submission	TBD (April)	20%

The course requirements and weights are final and will not be modified throughout the term. No late submissions will be accepted.

Communication Policy:

If you have a question, there is a high chance that other students in the course have the same question or, at least, will benefit from the answer. Please post all the questions to the INF2178 Quercus Discussion Board so everyone in the course can benefit from your questions and our answers. Students are encouraged to post answers to the questions of other students where appropriate.

Emails to the instructor and TAs must have a subject that starts with "INF2178" and include some more details, e.g., "INF2178: book appointment March 4th", **and must be submitted from your** *mail.utoronto.ca* **student account**.

Readings:

It is important to complete the required readings before the lecture in order to fully benefit from the class activities.

Grading:
Please consult the Faculty of Information's:

- Grade Interpretation Guidelines: http://ischool.utoronto.ca/wp-content/uploads/2020/08/grade_interpretation_revised_August2020.pdf

- The University Assessment and Grading Practices Policy: http://www.governingcouncil.utoronto.ca/Assets/Governing+Council+Digital+Assets/Policies/PDF/grading.pdf (note if nothing will be handed back before the final date to drop without penalty, that should be stated in the syllabus.)

- The Guidelines on the Use of INC, SDF, & WDR: https://www.sgs.utoronto.ca/policies-guidelines/inc-sdf-wdr/

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

Writing Support:
As stated in the iSchool'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 (http://www.sgs.utoronto.ca/currentstudents/Pages/English-Language-and-Writing-Support.aspx). 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-Terms-Courses.aspx) for more information.

Academic Integrity:
Please consult the University's site on Academic Integrity (http://academicintegrity.utoronto.ca). The iSchool has a zero-tolerance policy on plagiarism as defined 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). You should acquaint yourself with the Code. Please review the material in Cite it Right and if you require further clarification, consult the site How Not to Plagiarize (http://advice.writing.utoronto.ca/using-sources/how-not-to-plagiarize/).
Cite it Right covers relevant parts of the UofT Code of Behaviour on Academic Matters (1995). It is expected that all iSchool students take the Cite it Right workshop and the online quiz. Completion of the online Cite it Right quiz should be made prior to the second week of classes. To review and complete the workshop, visit the orientation portion of the iSkills site: https://inforum.library.utoronto.ca/workshops/orientation

The essence of academic life revolves around respect not only for the ideas of others, but also their rights to those ideas and their promulgation. It is therefore essential that all of us engaged in the

life of the mind take the utmost care that the ideas and expressions of ideas of other people always be appropriately handled, and, where necessary, cited. For writing assignments, when ideas or materials of others are used, they must be cited. APA format is suggested, however you may use any formal citation format you are familiar with, as long as it is used consistently in your paper, the source material can be located and the citation verified. What is most important is that the material be cited. In any situation, if you have a question, please post it to QUERCUS. Such attention to ideas and acknowledgment of their sources is central not only to academic life, but life in general.

Accommodations:
Students with diverse learning styles and needs are, of course, welcome in this course. If you have a disability or a health consideration that may require accommodations, please feel free to approach Student Services and/or the Accessibility Services Office (http://www.studentlife.utoronto.ca/as) as soon as possible. The Accessibility Services staff are available by appointment to assess needs, provide referrals and arrange appropriate accommodations. The sooner you let them know your needs, the quicker they can assist you in achieving your learning goals in this course.

Participation and Attendance:
Students Discussion and interaction in the classes are an important ways to learn. Sharing your experiences and ideas with your classmates is central to your learning experience in this course. As such, you should attend and participate in every class. There will also be exercises and discussions that you will participate in within your groups in your class. Some of the activities will be very helpful in completing your assignments.

Regrading Policy:
This is primarily a project-based course and as such, usual re-grading policies regarding assignment submission do not apply. Students and/or groups may reach out to the instructor and TA on an ad hoc basis to inquire about their course performance and progress. Instructors and TAs should ensure all communications with the student is in writing (e.g. follow-up e-mail) and keep a copy for later reference.

Academic Dates: https://ischool.utoronto.ca/current-students/academic-resources/academic-calendar/

Statement of Acknowledgement of Traditional Land:

The following is the University approved land acknowledgment statement for official ceremonies (Ceremonial Committee, Governing Council):
See: https://www.provost.utoronto.ca/wp-content/uploads/sites/155/2018/05/Final-Report-TRC.pdf

"I (we) would like to acknowledge this land on which the University of Toronto operates. For thousands of years it has been the traditional land of the Huron-Wendat, the Seneca, and most recently, the Mississaugas of the Credit River. Today this meeting place is still the home to many Indigenous people from across Turtle Island and we are grateful to have the opportunity to work on this land."

See also, the Faculty of Information's Commitment to the Findings and Call for Action of the Truth and Reconciliation Commission (approved at the Feb. 4, 2016 Faculty Council): https://ischool.utoronto.ca/wp-content/uploads/2017/11/iSchools-TRC-Commitment.pdf

Equity, Diversity and Inclusion:
The University of Toronto is committed to equity, human rights and respect for diversity. All members of the learning environment in this course should strive to create an atmosphere of mutual respect where all members of our community can express themselves, engage with each other, and respect one another's differences. U of T does not condone discrimination or harassment against any persons or communities.

Information about Faculty of Information iSkills and co-curricular Workshops:
The following workshop series are exclusively available to the Faculty of Information community. Faculty of Information professors, Inforum librarians, current students, alumni, and a collective of professionals and academics from each program and concentration, work together to create these unique rosters.

Together with the MMSt and MI curricula, these academic, professional, and technical iSkills workshops provide a robust information and heritage graduate educational experience.
iSkills Workshops: https://inforum.library.utoronto.ca/workshops/iSkills

In an effort to ensure your success at the Faculty of Information, key information and skills that all Faculty of Information students must possess, regardless of program or concentration, are covered in these online orientation workshops.
Orientation Workshops: https://inforum.library.utoronto.ca/workshops/orientation

Items Specific to Remote Course Delivery

Absence Declaration Tool
During the COVID-19 pandemic, the University is temporarily suspending the need for a doctor's note or medical certificate for absences from academic participation; students should use the Absence Declaration tool on ACORN to declare an absence if they require consideration for missed academic work; students are responsible for contacting instructors to request the academic consideration they are seeking; students should record each day of their absence as soon as it begins, up until the day before they return to classes or other academic activities.

FIPPA Video Recording Policy:

This course, including your participation, will be recorded on video and will be available to students in the course for viewing remotely and after each session. Course videos and materials belong to your instructor, the University, and/or other source depending on the specific facts of each situation, and are protected by copyright. In this course, you are permitted to download session videos and materials for your own academic use, but you should not copy, share, or use them for any other purpose without the explicit permission of the instructor. For questions about recording and use of videos in which you appear please contact your instructor.

Blackboard Collaborate
Blackboard Collaborate will no longer be supporting these browsers effective July 1, 2020:
- Native Microsoft Edge
- Google Chrome 78 and earlier - **make sure to update Chrome
- Best browsers to use - Chrome and Firefox

Minimum Technical Requirements
The University of Toronto has identified minimum technical requirements needed for students to access remote/online learning: https://www.viceprovoststudents.utoronto.ca/covid-19/tech-requirements-online-learning/

For other syllabus-related items specific to online/remote delivery see also:
https://teaching.utoronto.ca/teaching-support/course-design/developing-a-syllabus/