MAT4376E | 5314E | FALL 2023

INSTRUCTOR: P. BOILY

uOttawa

COURSE DESCRIPTION

In October 2012, the *Harvard Business Review* published an article calling data science the **sexiest job of the 21st century**, a long cry from the business-as-usual practice of data geeks playing a supporting role in organizations.

Today's data scientists are not just number-crunchers. As a combination of **data hacker**, **analyst**, **communicator**, and **trusted adviser**, they discover meaningful relationships in ever-growing masses of information and play a leading role in the decision-making processes.

This is a **project-based** course, which is focused on **the delivery of useful analyses rather than on academic technical know-how** (although we will also talk about this).

MULTIPLE I'S FRAMEWORK

Intuition: understanding the data and the analysis context

Initiative: establishing an analysis plan

Innovation: searching for new ways to obtain results, if required

Interpretability: providing explainable results

Insights: providing actionable results

Integrity: staying true to the analysis objectives and results

MULTIPLE I’S FRAMEWORK

Independence: developing self-learning/self-teaching skills

Interactions: building strong analyses through (often multi-disciplinary) teamwork

Interest: finding and reporting on interesting results

Intangibles: putting a bit of yourself in the results/reports; thinking “outside the box”

Inquisitiveness: not just asking the same questions over and over again

etc.

COURSE LOGISTICS

Course Schedule:

- MON 08:30-10:00 | STE J0106
- WED 13:00-14:30 | STE J0106
- OFFICE HOURS: SLACK

Course Website:

- data-action-lab.com/tda
- tda-f23.slack.com

Pre-requisites:

- Programming proficiency
- MAT2122: multivariable calculus
- MAT2141: linear algebra
- MAT2371+MAT2375/MAT2377: prob and stats
- MAT3375: regression analysis

Textbooks:

- Data Understading, Data Analysis & Data Science
- The Practice of Data Visualization

DELIVERABLES

The components all count **equally** towards your final grade.

4376: 1 project for each of the 4 modules.

5314: same + 1 project on a topic of your choice, which I must approve by 22-Sep.

Engagement matters in this course: students who do not participate earnestly and who do not attend at least 17 of the sessions will be docked 10 marks off their final grade.

All communications must take place through Slack; you will be adding me to each team's project channel, as an **insurance policy**. You are of course allowed to DM one another without including me and to use other channels for stuff un-related to the course, or if you want to work something out without involving me (which I encourage whole-heartedly).

Module	Week	Monday	Tuesday	Wednesday	Thursday	Friday	Weekend
Module 0: Preamble	04-Sep	*Storytelling with Data* | in class **Video Lectures:** 0.1 – 0.4 **(3:59:29)** | *Asynchronous*					
	11-Sep						
Module 1: Data Visualization	18-Sep	**Video Lectures:** 1.1 – 1.7 **(4:28:04)** | *Asynchronous*				5314E project topic approval due	
	25-Sep	**Working on Project 1** | in class					
	02-Oct					Project 1	
Module 2: Bayesian Data Analysis	09-Oct	**Video Lectures:** 2.1 – 2.7 **(4:06:59)** | *Asynchronous*					
	16-Oct	**Working on Project 2** | in class *Reading Week:* 23-Oct to 27-Oct					
	23-Oct					Project 2	
Module 3: Queueing Systems	30-Oct	**Video Lectures:** 3.1 – 3.6 **(2:39:29)** | *Asynchronous*					
	06-Nov	**Working on Project 3** | in class					
	13-Nov					Project 3	
Module 4: Anomaly Detection & Outlier Analysis	20-Nov	**Video Lectures:** 4.1 – 4.6 **(5:51:03)** | *Asynchronous*					
	27-Nov	**Working on Project 4** | in class					
	4-Dec					Project 4	
	11-Dec					5314E proj. due	
							No final exam

COURSE CONTENT

1. Data Visualization
2. Bayesian Data Analysis
3. Queueing Systems
4. Anomaly Detection and Outlier Analysis
5. Graduate Project

EXPECTATIONS

I expect you to spend **8-10/12-15 hrs** [4376/5314] per week, on avg, on this course.

Teamwork is crucial to insightful data analysis. You **must** work in teams of 3 or 4 (teams may change from one project to another, but you have to stay within your course code and announce your teams publicly on Slack prior to the start of each project – students who have not formed teams before the deadline will be grouped randomly); **the grade is given to the whole** group (**independently of the quantity/quality of the work performed by each person**).

You may have to use methods or concepts that have not been discussed in the lectures [see Multiple I's]. **First course objective:** start building a **data project portfolio**.

EXPECTATIONS

Second objective: learn to navigate tight deadlines and plan your analysis/reporting accordingly (**10-page limit, PDFs, uploaded to Slack by the deadline**).

In this flipped approach, you get class time to work on the project, but **it will not be sufficient**. Do not wait until the last minute before starting work on your projects. Also: don't forget to **back your work up** as you go!

There may be times when you are unable to deliver the projects by the deadline due to reasons outside your control. You are requested to inform me (and to submit the work you have already completed) **as soon as you become aware of such a situation arising** (within reason) so that we can discuss alternatives.

Q&A

Q: How will you grade the projects?

A: If

- you don’t entirely answer the question, or
- there are too many errors/glaring mistakes/omissions, or
- your deliverable is difficult to read because it’s too long, not arranged in a logical manner, and/or inconsistently written (I am not talking about the quality of written English)

you cannot get higher than a **B+** for a project.

[I will be expecting more from 5314E students, btw].

Q&A

To obtain an **A-**, **A**, or **A+** on a project, you also need to clearly address the “Multiple I’s” in a way that demonstrates that you understand their importance to the project at hand and that you’ve thought about the project in a critical way.

That is not easy to do – I’ll recognize an **A-**, **A**, or **A+** when I see one, but **I can’t tell you ahead of time** what is required for you to get such a grade.

You can ask me for my thoughts on things that you think could add to your write-up/ analysis (I will be more lenient with the marking if I have some evidence that you have thought about this), and I will try to give you practical answers, but **until I see how you’ve implemented your analysis plan or how you’ve written things up**, I can’t guarantee that it will yield an A or an A+.

Q&A

I don't intend to grant anybody a project grade below a **C**: if your work is not good enough to get a **C**, I will simply hand out an **F** for the project

- 2+ **F** $\Longrightarrow$ final grade = **F**; 1 **F** $\Longrightarrow$ final grade $\leq$ **B**

But I will definitely err on the side of generosity**: you would need to flat-out not hand in a project or only do half of it (or less) to get an F**. I reward honest attempts: you get a chance to "fail", without it affecting your grade much (I will discuss this further).

TL;DR Summary: getting at least an **A-** in this course is within everybody's reach, with a reasonable amount of work. Historically, undergrad grades are distributed as follows:

- 4376: **A+** (5%), **A** (25%), **A-** (30%), **B+** (10%), **B** (10%), **C+** (10%), **C** (10%)
- 5314: **A+** (5%), **A** (45%), **A-** (50%)

Q&A

Q: But 10 pages is not enough!!! We need more time!!! And there are too many projects!!!

A: That's not a question… but, yes. I'm afraid it's true. 10 pages is **not a lot…** and 3 weeks is a **really, really short** period of time to work on any project. That's … the whole point.

But I'm **NOT** asking you to find the optimal solution or to be technically perfect. I want you to focus on the Multiple I's framework:

- think about the problem;
- come up with an analysis plan, and decide what to keep/omit;
- implement it; and
- report on your analysis results in a way that is going to be **useful**.

Q&A

Q: I'm working with dataset *XYZ*, and I'm wondering what variable *W* stands for?

A: A-ha! Excellent question! How could you find out?

That is often going to be the answer… "how could you find out?"

Don't just stand there and hope for inspiration: ask questions! Profs/classmates/search engines are your friends!

Q&A

Q: In this project, you ask to produce a “definitive” so-and-so. What does that mean?

A: I mean, what should definitely be found in a report, a dashboard, a blog article, etc.

The "definitive" so-and-so’s are the ones that are essential to the story.

Q&A

Q: Do we need to provide code?

A: Not necessarily, however if you can find a way to include it in your report in a natural manner (either in an appendix or in the text itself), that's probably a good idea…

Sometimes, though, you'll have no choice. If not having the code **STOPS** the story from being conveyed, somehow, then you need to add the code. Use your judgment.

It's ok to be wrong about that, too.

Q&A

If you feel that your project stands strong without code, you don't include the code.

If you feel that it would be stronger with code, you add code.

The only **restrictions** (if you stick to a report) are:

- the 10-page limit;
- meeting the stated objectives (are you actually solving the problem?), and
- are you dealing with the Multiple I’s.

Q&A

Q: Why do you care about reports? We're not English majors!

A: The purpose of writing the reports is to give you a bit of a chance to practice the communication aspect of data analysis/data science, which is **substantially more important** out in the real world than you might hitherto have been led to believe.

As I suspect that most of us are not native English speakers (myself included), and since stylistics are mostly a matter of taste, **you will not be marked on grammar and style** (unless there are issues of consistency).

Q&A

Q: Can we use ChatGPT?

A: Even though it goes against everything I believe in, **I'll allow it**.

The tool could prove useful to clean up your writing and/or to translate text/code. But you should remain aware that AI-generated technical writing only looks impressive to people **who don't understand technical matters in the first place**.

If you use it to run the analyses, you will end up with egg on your face. Not a good look.

Q&A

Q: What exactly goes in a report/deliverable?

A: When in doubt, go back to the **Multiple I'**s:

- come up with a narrative, a story, and sell it;
- or talk about the process, the challenges;
- or share some insight into how people could misread your visualizations/results, etc.

If you're not sure what the objective of the exercise is, **make up your own objective**, then go out there and meet it. The instructions to the **B+** are fairly clear and easy to meet if you justify your work. The instructions to the **A-/A/A+** are left vague by design.

Q&A

There is no guarantee that any of this will net you an **A-/A/A+**; execution still matters, and you need to justify your choices. But that is the kind of stuff I am looking for.

So add code, or music, or video, or … well, whatever you want.

I am aware that students are not typically fond of vagueness and open-endedness, but … **that's the entire point:** try things, tell a story, learn something along the way.

Q&A

Q: I am not good with people and I hate teamwork: . Can I work alone?

A: Nope. This is the **third course objective**: data science is a team sport. I am not asking you to become extroverts or to change your attitude towards people: I am asking you to make teamwork… work. If this isn't your cup of tea, then this course is not for you. It's as simple as that. No hard feelings.

BUT… I am also expecting respect for yourself and your teammates (both academically and socially). The university rules governing interactions apply (in class, on Slack, etc.).

Q&A

Q: Ok, ok. If I partner up with some students for a project, do I have to partner up with them for the other projects?

A: No. If somebody does not pull their weight, you don't have to work with them again. If working with somebody stresses you out, you don't have to work with them again.

At times, your teammates may have to cover for you, and vice-versa. But **communication is key**: if you wait until the day before the deadline to tell them you can't do your part, your team's grade is affected, and your teammates will understandably not want to work with you again. This is not the place for secrets.

I'm going to be treating all of you as adults. Behave accordingly.

Q&A

Q: You “taught” us `R`, but `R` is not a real language and most jobs require Python. Why?

A: I’m agnostic when it comes to the choice of software/coding language. Frankly, it’s the **wrong question to ask:** they’re all more or less equivalent and once you know how one works, it’s easy to figure out the others. We could be doing all of this in Visual Basic.

In this course, your grade depends on the quality of the work and the write-up, not on what software/language you use (use whatever allows you to do the work).

And of course, on how you deal with the **Multiple I'**s.

Q&A

Q: Would it be a good idea to include plots/results/etc. that don't end up being useful, then explaining why that was so and how we changed them to make them better?

A: That really depends on what your own personal objectives/interests are.

Showing the process could be a positive outcome (learn from my mistakes!), including a useless plot/result/etc. could be a negative outcome (adds unnecessary length and distracts the audience from the point).

Either way, given the time constraint, you need to figure out what you are really trying to convey and organize yourself accordingly **BEFORE** you start producing the deliverable.

WHAT SOME FORMER STUDENTS ARE SAYING

“This was the first course I took where there were no midterms or final exams. The workload seemed heavy at first but being able to complete a small, concrete assignment in each section of the course made learning very motivating! As the term progressed, I became more confident in my abilities and more motivated!”

“I have never learned so much in a course, and I have never acquired so many tools that I will be able to use outside of university!”

“We were rewarded for our efforts: in the correction of the projects, the prof emphasized that we should not be afraid to make mistakes, as it is part of the learning process. The most important thing was to try to go over and beyond the bare minimum while thinking of innovative ideas, a very important asset to develop for the job market.”

WHAT SOME FORMER STUDENTS ARE SAYING

“Unless you’re writing a thesis, you’re asked a question and your answer is either right or wrong - there is little nuance to be found. In this course, there was rarely a right or wrong answer. This approach not only motivates students to think outside the box but also promotes mathematical curiosity.”

“The intentionally ambiguous instructions helped me prepare for the workforce. They allowed me to step out of my comfort zone and push myself in terms of creativity, which I would have been too scared to attempt in a work setting.”

A WORD ABOUT DATA ETHICS

Fourth objective: One of my goals is to make you think about what effects our analyses/algorithms might have on individuals/societies/planet **in the short and in the long run**. There are plenty of ugly examples out there which could easily have been avoided.

As data analysts, we can't just talk the talk, we also need to be able to walk the walk.

It's one thing to realize after the fact that others have done shady things, but there aren't going to be big glowing letters floating in the sky warning you that you are about to embark on a project that might not meet your ethical standards 2 years down the road.

A WORD ABOUT DATA ETHICS

All the nice principles and good sentiments in the world will be worth absolutely nothing if they are not put in practice **EVEN WHEN IT IS NOT CONVENIENT TO DO SO**!!

Ultimately, data people need to know where they stand.

It's easier to back out of a project before it starts, but it might be unrealistic to hope that you will never be led astray.

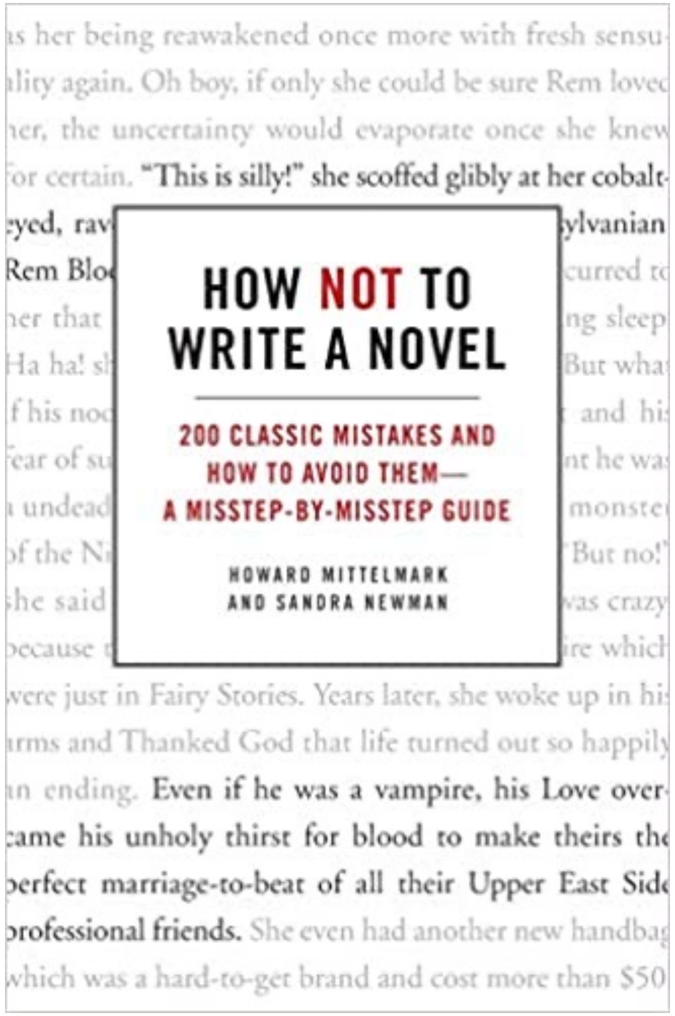
Doubt is your ally: ask yourself frequently whether your projects are going where you think they should be going, and what consequences they lead (or might lead) to. Your answers might change over time, and it is never too late to pull the plug.

FINAL REMARK

Not all objectives are created equal, and you might need to do a fair bit of work to justify your choice of objective, and **even then I might not buy it**. But I do want you to have an objective **outside the framework of what I ask you to do**.

Above all, **I want you to try**. An honest attempt, not a perfunctory one. Embrace the vagueness, don't be afraid to make mistakes, etc.

Play along with me for one term: in the Winter, you can go back to being a by-the-book analyst if you want.



HOW NOT TO WRITE A NOVEL

200 CLASSIC MISTAKES AND HOW TO AVOID THEM— A MISSTEP-BY-MISSTEP GUIDE

HOWARD MITTELMARK AND SANDRA NEWMAN

as her being reawakened once more with fresh sensu-
ality again. Oh boy, if only she could be sure Rem loved
her, the uncertainty would evaporate once she knew
for certain. "This is silly!" she scoffed glibly at her cobalt-
eyed, rav ylvanian
Rem Blo curred to
her that ng sleep
Ha ha! sh But wha
f his noc and his
ear of su nt he was
undead monster
of the N 'But no!'
she said was crazy
because t re which
were just in Fairy Stories. Years later, she woke up in his
arms and Thanked God that life turned out so happily
an ending. Even if he was a vampire, his Love over-
came his unholy thirst for blood to make theirs the
perfect marriage-to-beat of all their Upper East Side
professional friends. She even had another new handbag
which was a hard-to-get brand and cost more than $50