*Mail*

*We recognize and acknowledge that McMaster University meets and learns on the traditional territories of the Mississauga and Haudenosaunee nations, and within the lands protected by the “Dish With One Spoon” wampum, an agreement amongst all allied Nations to peaceably share and care for the resources around the Great Lakes.*

# STATS 4T03 A/B – Senior Research Project

# 2022 -23

**Supervisor:**  Dr. Pratheepa Jeganathan |  **E-mail:** [jeganp1@mcmater.ca](mailto:jeganp1@mcmater.ca) | **Office: HH-212**

**Co-supervisor:**  Jamie McNicol |  **E-mail:** [xxx@mcmater.ca](mailto:xxx@mcmater.ca) | **Office: YYY**

**Course coordinator:**  Associate Chair (Undergraduate) | **Email: mathstatsugadvisor@mcmaster.ca**

## Location and time:

**Weekly meeting:** Tuesday (in person) 2.30 -3.30 PM.

**Bi-weekly meeting:** Tuesday 3.30-4.30 ? (Virtual)

**Course Website**

Github repository. Reading lists, workflows, presentations, will be shared through Github.

## Course Description

6 unit(s)

Build upon an existing Image Analysis and Differential Discovery Pipeline by incorporating Topic Modeling to better characterize the tissue microenvironment (TME). The TME is recognized as an important determinant of patient outcome. Leverage Topic Modeling to better report the elements within the TME that correlate with patient outcomes.

A written report and oral presentation will be required. This course includes a scientific communication component.

Weekly meetings with the supervisor and bi-weekly meetings with the co-supervisors; One occasional tutorial during weekly meetings; two terms

## Prerequisite(s):

Registration in Level IV of any Honours Mathematics and Statistics program; and a GPA of at least 9.0; and permission of the Chair of the Department

Not open to students with credit or registration in MATH 4P06 A/B S, PHYSICS 4P06 A/B or ISCI 4A12 A/B.

## Course and Learning Objectives

In this course, the student will:

1. Experience doing research in the mathematical sciences. This may involve solving one or more challenging mathematical or statistical problems, an in-depth analysis of a large data set, or reading and understanding a research paper that involves absorbing a significant amount of new mathematics or statistics.
2. Improve their mathematical writing by writing a document explaining their work.
3. Enhance their oral presentation skills by giving a public presentation explaining their work.
4. Develop additional skills deemed relevant by their supervisor; e.g. coding, data collection.

## Virtual Course Delivery:

In the unlikely event that the course has to be delivered virtually, please note the following.

**To follow and participate in virtual classes it is expected that you have reliable access to the following:**

* A computer that meets performance requirements [found here](https://cto.mcmaster.ca/technology-resources-for-mcmaster-students/#tab-content-device-recommendations).
* An internet connection that is fast enough to stream video.
* Computer accessories that enable class participation, such as a microphone, speakers and webcam when needed.

If you think that you will not be able to meet these requirements, please contact [uts@mcmaster.ca](mailto:uts@mcmaster.ca) as soon as you can. Please visit the [Technology Resources for Students page](https://cto.mcmaster.ca/technology-resources-for-mcmaster-students/#tab-content-device-recommendations) for detailed requirements. If you use assistive technology or believe that our platforms might be a barrier to participating, please contact [Student Accessibility Services](https://sas.mcmaster.ca/), [sas@mcmaster.ca](mailto:sas@mcmaster.ca), for support.

**Note:**

(When virtual) A synchronous teaching method will be used. Students must attend synchronous virtual lecture via MSTeams. After the synchronous class, recorded lectures and computing sessions will be available on Avenue to Learn.

### Required Materials/ Resources:

Reading lists are selected during weekly meetings.

1. BDA3: Bayesian Data Analysis, Gelman, Carlin, Stern, Dunson, Vehtari, Rubin.

### Software:

This course uses R, RStudio, Python, and Python Notebook, which are free. We recommend to set-up the computing environment earlier. From the first day of the class, you must ensure that you are up and running with the computing environment. There are three ways to do computing.

1. Use your personal computing device (laptop/desktop)
   1. Install R (required): [https://www.r-project.org/.](https://www.r-project.org/)
   2. Install RStudio: highly recommended for syntax highlighting, package management, document generation, and more: <https://www.rstudio.com/>.

* The newest version of RStudio is highly recommended.
  1. Install Latex, which will enable you to create PDFs directly from the RMarkdown in RStudio.
* Install [TinyTex](https://yihui.name/tinytex/) package: install.packages("tinytex", repos = "https://cloud.r-project.org/").
* After installing TinyTex, close RStudio.
* Reopen RStudio.
  1. Install [Python](https://www.python.org/downloads/) (required).
  2. Install [Anaconda](https://www.anaconda.com/).

1. Use interactive computing notebooks hosted by the Digital Research Alliances of Canada (previously known as Compute Canada).
   1. Visit <https://mcmaster.syzygy.ca/>
   2. Sign in with your McMaster credentials.
   3. Choose R or Python.
   4. You can download the PDF of your workflow.
2. Google Collab and RStudio Cloud.
   1. Visit <https://rstudio.cloud/> and create an account.
   2. Visit <https://colab.research.google.com/> and create an account or use your Google account to sign up.

Other recommended tools

1. Install git <https://git-scm.com/> for version control.
2. Create an account at <https://github.com/>. Use your McMaster email to get an education account at <https://education.github.com/pack/join>.
3. Create an account at <https://www.shinyapps.io/> to deploy your Shiny applications.

The following templates will be useful for assignments, projects (report + supplementary materials), and presentations.

1. Download the report template from A2L and recreate it. Written in RMarkdown. You can use either Quarto or (Latex + RMarkdown) or (Latex + Juptyer Notebook. No other tools are allowed. The supplement must be reproducible.
2. Download the presentation template from A2L and recreate it. Written in R Markdown. You can use either Latex or Quarto or any other presentation tools.

## Course Overview and Assessment

### Topics:

1. Introduction to data science.
2. Data transformation, tidy data.
3. Pre-processing (statistical transformation, handling missing value and outliers).
4. Data visualization (grammar of graphics, Shiny, other than tabular data).
5. K-nearest neighbour regression, classification.
6. Cross-validation + Model selection.
7. Generalized linear models.
8. Error analysis.
9. Sensitivity and specificity analysis.
10. CART.
11. Bootstrap method.
12. Bagging, random forests, boosting, Bayesian additive regression tree.
13. Principal component analysis, matrix completion, singular value decomposition, factor analysis, independent component analysis.
14. Clustering.
15. Model-based learning (Gaussian mixture models, discriminant analysis), EM algorithm
16. Feature selection, shrinkage, PCR/PLR, sparse learning methods, information theory.
17. Nonlinear function fitting algorithms (GAM)
18. Support vector machines
19. Introduction to neural networks.
20. Single layer and multiple layer neural networks.
21. Introduction to deep learning (CNN and RNN).
22. Introduction to topic models.

### Course Schedule (tentative):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Date** | **Week** | **Topics** | **Readings** | **Notes** |
| 09/06/2022 | Lecture 1 | Logistics,  introduction to data science,  data transformation (R) | [Top 10 Ideas in Statistics That Have Powered the AI Revolution](https://news.columbia.edu/news/top-10-ideas-statistics-ai) [How to share data with a statistician](https://github.com/jtleek/datasharing)  [Data science in a time of crisis: lessons from the pandemic (Statistical Science)](https://projecteuclid.org/journals/statistical-science/volume-37/issue-2)  [IEEE Transactions on pattern analysis and machine intelligence (Popular documents)](https://ieeexplore-ieee-org.libaccess.lib.mcmaster.ca/xpl/topAccessedArticles.jsp?punumber=34) |  |
| 09/09/2022 | Lecture 2 | Data transformation, preprocessing,  grammar of graphics (R) | RDS Chapter 3, 5, 7  ISLR Section 12.3,  [IFT6758 Feature engineering lecture notes](https://ift6758.github.io/lectures/feature_engineering.pdf) |  |
| 09/13/2022 | Lecture 3 | More on data visualization + Shiny (R) | [Shiny learning resources](https://shiny.rstudio.com/tutorial/) |  |
| 09/16/2022 | Lecture 4 | Data visualization – text, network, spatial data (R) | <https://krisrs1128.github.io/stat679_notes/> |  |
| 09/20/2022 | Lecture 5 | Function fitting,  bias-variance tradeoff, residual analysis, examples in R and Python | ISLR Sections 2.1, 2.2, 3.2, 3.5 |  |
| 09/23/2022 | Lecture 6 | KNN regression, classification,  curse of dimensionality (R) | ISLR Sections 4.4, 4.5, 4.7 |  |
| 09/27/2022 | Lecture 7 | Cross-validation (Validation set, LOOCV, K-fold CV) (R) | ISLR Sections 5.1, 5.3 | Homework assignment 1 due |
| 09/30/2022 |  |  |  |  |
| 10/04/2022 | Lecture 8 | Linear models, generalized linear models,  sensitivity-specificity analysis (R) | ISLR Sections 4.1, 4.2, 4.3, 4.6 |  |
| 10/07/2022 | Lecture 9 | Model selection measures , Shrinkage methods for linear regression (ridge, lasso, elastic net) (R) | ISLR Chapter 6 |  |
| 10/10 |  |  |  |  |
| 10/16 |  |  |  |  |
| 10/18/2022 | Lecture 10 | Trees (Classification and regression) (Python) | ISLR Section 8.1 | Homework assignment 2 due |
| 10/21/2022 | Lecture 11 | Bootstrapping, bagging, RF (Python) | ISLR Sections 5.2, 8.2 |  |
| 10/25/2022 | Lecture 12 | Boosting, BART (Python) | ISLR Section 8.3 | Mid-term proposal due |
| 10/28/2022 | Lecture 13 | Dimensionality reduction (PCA, SVD, FA, ICA, UMAP) (Python) | ISLR Chapter 12.2,  MSMB Chapter 7  [Ten tips for effective dimensionality reduction](https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006907) |  |
| 11/01/2022 | Lecture 14 | Clustering (Python) | ISLR Sections 12.1, 12.2,  MSMB Chapter 5 |  |
| 11/04/2022 | Lecture 15 | Model-based learning (Gaussian mixture models and discriminant analysis),  EM algorithm  (Python) | ISLR Sections 4.5,  ESL Sections 6.6.3, 6.8 (Gaussian mixture models), Chapter 8 (EM algorithm) |  |
| 11/08/2022 | Lecture 16 | Feature selection (PCR/PLS/sparse learning)  (Python) | ISLR Section 6.3  [An Introduction to Variable and Feature Selection](An%20Introduction%20to%20Variable%20and%20Feature%20Selection),  [IFT6758 Feature selection lecture notes](https://ift6758.github.io/lectures/feature_selection.pdf)  [Laplacian Score](http://www.cad.zju.edu.cn/home/dengcai/Publication/Conference/2005_NIPS_LaplacianScore.pdf)  [for Feature Selection](http://www.cad.zju.edu.cn/home/dengcai/Publication/Conference/2005_NIPS_LaplacianScore.pdf),  [A Framework for](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2930825/)  [Feature Selection in Clustering](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2930825/)  [Spectral Feature Selection for](https://dl.acm.org/doi/abs/10.1145/1273496.1273641)  [Supervised and Unsupervised Learning](https://dl.acm.org/doi/abs/10.1145/1273496.1273641) |  |
| 11/11/2022 | Lecture 17 | Nonlinear function fitting (Gam, SVM)  (Python) | ISLR Chapters 7, 9 | Homework assignment 3 due |
| 11/15/2022 | Lecture 18 | Introduction to neural networks (single layer and multiple layer NN) (Python) | ISLR Sections 10.1-10.2, 10.7 |  |
| 11/18/2022 | Lecture 19 | Convolution neural networks  (Python) | ISLR Sections 10.3-10.4 |  |
| 11/22/2022 | Lecture 20 | Recurrent neural networks  (Python) | ISLR Sections 10.5 |  |
| 11/25/2022 | Lecture 21 | Topic models  (Python)  NMF | [Topic modeling by David M. Blei](http://www.cs.columbia.edu/~blei/topicmodeling.html): general introduction to topic modeling  [Learning the parts of objects by non-negative matrix factorization](https://www-nature-com.libaccess.lib.mcmaster.ca/articles/44565) (1999)  [A comprehensive survey on community detection with deep learning](https://ieeexplore-ieee-org.libaccess.lib.mcmaster.ca/stamp/stamp.jsp?tp=&arnumber=9732192) (2022) |  |
| 11/29/2022 | Students’ Presentation |  |  | Presentation slides due |
| 12/02/2022 | Students’ Presentation |  |  |  |
| 12/06/2022 | Students’ Presentation |  |  |  |
| 12/12/2022 |  |  |  | Final project report due |
|  |  |  |  |  |

### Evaluation:

* **Assignments:** 45%. There will be three assignments, each worth 15% of the final grade. Instructions for each assignment will be available at least two week before the due date.
  + Assignment 1 due on **September 23rd, 2022, 11:00 (EST).**
  + Assignment 2 due on **October 18th, 2022, 11:00 (EST).**
  + Assignment 3 due on **November 11th, 2022, 11:00 (EST).**
* **Class project:** 55%. There will be one project, which includes project proposal, an oral presentation, and write up. The class project has three parts.
  + **The project proposal report + supplementary material**: 10%. 2-3 pages limit (single-spaced, 12 pt, 1-inch margins, including graphs and tables) + supplementary material (less than 5 pages of code and results), an overview of two methods you plan to compare and the real dataset on which you will do one of the comparisons. Dataset must have mixed datatypes of more than 15 variables. Due on **October 25th, 2022, 22:00 (EST**).
  + **Presentation slides**: 5% 6-8 slides limit (outline, motivation, data, methods, results, discussion), the dimension of each slide must be no larger than 1024 X 768, and font size should be no smaller than 16pt. 6-8 mins, **due on November 29th, 2022, 22:00 (EST)**.
  + **Oral presentation**: 10%. 6-8 minute presentation and 2-4 minutes Q&A. Due on **either 11/29 or 12/2 or 12/6, 2022 during class time**.
  + **Final project report + supplementary material**: 30%. Less than 15 pages (single-spaced, 12 pt, 1-inch margins, include inclusive of tables, figures, and bibliography) + (less than 10 pages of code and results), due on **December 12th, 2022, 22:00 (EST).**

## Course Policies and Procedures

Class participation

* In-person participation and Avenue Discussion Forum are encouraged.
  + Avenue discussion forum is highly catered to getting help fast and efficiently from classmates, the TA, and myself. Rather than emailing questions to the teaching staff, we encourage you to post your questions on Avenue Discussion Forum.
  + When homework involves simulations and data analysis, you will use R statistical computing software. Please post your R or RStudio or RMarkdown or Python or Jupyter Notebook questions to Avenue Discussion Forum.
  + Instructor or TA or other students in your class can answer your questions.
  + When asking questions about code, be specific (copy and paste the exact error, relevant code, and describe what you are attempting to do). We will not answer questions that are too similar to the problem sets or better-answered in-office hours. You should conduct yourself courteously and respectfully on Avenue Discussion Forum.

Homework assignments

* Homework assignment will be assigned in [Avenue to Learn](https://avenue.mcmaster.ca/).
* Homework assignment due date will be announced on the assignments posted date (at least ten days, including a weekend).
* Prepare your completed homework assignment (report and supplementary material) using RMarkdown (or other tools listed in software section).
  + Render your RMarkdown to PDF. All the codes (rendered to PDF) must be in the technical appendices.
  + Name your PDF file as "LastName\_FirstName\_HWK", where K=1,2,… , 4.
  + Submit a copy to Avenue to Learn.
* **Late penalty for assignments:** 15% will be deducted from assignments each day after the due date (rounding up). Assignments won’t be accepted after 48 hours after the due date.
* After attempting homework problems on an individual basis, students may discuss a homework assignment with their classmates. However, students must write up their solutions individually and explicitly indicate who (if anyone) or resources students received help at the top of their homework solutions.

Late Withdrawal

McMaster University provides a Late Withdrawal option to assist students who have become irretrievably behind in a course. Students who have fallen behind with assignments and/or are not prepared to write final examinations (or equivalent) in one or more courses are encouraged to make use of this option and must contact their Academic Advisor in the Faculty/Program Office. Students will work with their Academic Advisor to discuss the situation and what steps they can take to prevent a recurrence. The maximum number of units for which students may request a Late Withdrawal is 18 units throughout their undergraduate degree.

Students may request a Late Withdrawal, without petition, no later than the last day of classes in the relevant Term. However, it is important to note that:

* Requests for Late Withdrawal cannot be made in courses for which the final exam (or equivalent) has been attempted or completed. This also includes courses where a final grade has been assigned (e.g. clinical courses).
* Such requests will be cancelled or revoked if it is determined that the student attempted or completed the final exam (or equivalent).
* Students cannot use the Late Withdrawal option for courses in which they are under investigation or for which they have been found guilty of academic dishonesty.

Course(s) approved for Late Withdrawal will be:

* Assigned a non-numeric grade of LWD, in lieu of an alpha/numerical grade
* Excluded from the calculation of the GPA
* Ineligible for tuition refund

Approval of a late withdrawal is final, and requests to be re-enrolled in the withdrawn course(s) will not be considered. A withdrawal will not preclude students from enrolling in the course(s) in a subsequent term.

**For this course, the final examination equivalent is the final project. Requests for Late Withdrawal cannot be made after submitting the final presentation slides**.

## Requests for Relief for Missed Academic Term Work

[McMaster Student Absence Form (MSAF):](https://secretariat.mcmaster.ca/university-policies-procedures-guidelines/msaf-mcmaster-student-absence-form/) In the event of an absence for medical or other reasons, students should review and follow the Academic Regulation in the Undergraduate Calendar “Requests for Relief for Missed Academic Term Work”.

**MSAF Course Specific Information**

Absences for a longer duration or for other reasons must be reported to your Faculty/Program office, with documentation, and relief from term work may not necessarily be granted. Please note that the MSAF may not be used for any part of the project work.

**The weight (15%) for the missed assignment due to MSAF will be used for the next assignment to calculate the final grade. We will not accept MSAF for any project component.**

## Academic Accommodation of Students with Disabilities

Students with disabilities who require academic accommodation must contact [Student Accessibility Services (SAS](https://sas.mcmaster.ca/)) at 905-525-9140 ext. 28652 or [sas@mcmaster.ca](mailto:sas@mcmaster.ca) to make arrangements with a Program Coordinator. For further information, consult McMaster University’s [*Academic Accommodation of Students with Disabilities*](https://secretariat.mcmaster.ca/app/uploads/Academic-Accommodations-Policy.pdf) policy.

## Academic Accommodation for Religious, Indigenous Or Spiritual Observances (Riso)

Students requiring academic accommodation based on religious, indigenous or spiritual observances should follow the procedures set out in the [RISO](https://secretariat.mcmaster.ca/app/uploads/2019/02/Academic-Accommodation-for-Religious-Indigenous-and-Spiritual-Observances-Policy-on.pdf) policy. Students should submit their request to their Faculty Office ***normally within 10 working days*** of the beginning of term in which they anticipate a need for accommodation or to the Registrar's Office prior to their examinations. Students should contact their instructors as soon as possible to make alternative arrangements for classes, assignments, and tests.

## Academic Integrity

You are expected to exhibit honesty and use ethical behaviour in all aspects of the learning process. Academic credentials you earn are rooted in principles of honesty and academic integrity.

**It is your responsibility to understand what constitutes academic dishonesty.**

Academic dishonesty is to knowingly act or fail to act in a way that results or could result in unearned academic credit or advantage. This behaviour can result in serious consequences, e.g. the grade of zero on an assignment, loss of credit with a notation on the transcript (notation reads: “Grade of F assigned for academic dishonesty”), and/or suspension or expulsion from the university. For information on the various types of academic dishonesty please refer to the [*Academic Integrity Policy.*](https://secretariat.mcmaster.ca/app/uploads/Academic-Integrity-Policy-1-1.pdf)

**The following illustrates only three forms of academic dishonesty:**

* plagiarism, e.g. the submission of work that is not one’s own or for which other credit has been obtained.
* improper collaboration in group work.
* copying or using unauthorized aids in tests and examinations.

## Authenticity / Plagiarism Detection

***Some courses may*** use a web-based service (Turnitin.com) to reveal authenticity and ownership of student submitted work. For courses using such software, students will be expected to submit their work electronically either directly to Turnitin.com or via an online learning platform (e.g. A2L, etc.) using plagiarism detection (a service supported by Turnitin.com) so it can be checked for academic dishonesty.

Students who do not wish their work to be submitted through the plagiarism detection software must inform the Instructor before the assignment is due. No penalty will be assigned to a student who does not submit work to the plagiarism detection software. **All submitted work is subject to normal verification that standards of academic integrity have been upheld** (e.g., on-line search, other software, etc.). For more details about McMaster’s use of Turnitin.com please go to the [McMaster Office of Academic Integrity’s](https://www.mcmaster.ca/academicintegrity/) webpage.

## Conduct Expectations

As a McMaster student, you have the right to experience, and the responsibility to demonstrate, respectful and dignified interactions within all our living, learning and working communities. These expectations are described in the [*Code of Student Rights & Responsibilities* (the “Code”).](https://secretariat.mcmaster.ca/app/uploads/Code-of-Student-Rights-and-Responsibilities.pdf) All students share the responsibility of maintaining a positive environment for the academic and personal growth of all McMaster community members, **whether in person or online**.

It is essential that students be mindful of their interactions online, as the Code remains in effect in virtual learning environments. The Code applies to any interactions that adversely affect, disrupt, or interfere with reasonable participation in University activities. Student disruptions or behaviours that interfere with university functions on online platforms (e.g. use of Avenue 2 Learn, WebEx or Zoom for delivery), will be taken very seriously and will be investigated. Outcomes may include restriction or removal of the involved students’ access to these platforms.

## Copyright and Recording

Students are advised that lectures, demonstrations, performances, and any other course material provided by an instructor include copyright protected works. The Copyright Act and copyright law protect every original literary, dramatic, musical and artistic work, **including lectures** by University instructors.

The recording of lectures, tutorials, or other methods of instruction may occur during a course. Recording may be done by either the instructor for the purpose of authorized distribution, or by a student for the purpose of personal study. Students should be aware that their voice and/or image may be recorded by others during the class. Please speak with the instructor if this is a concern for you.

## Extreme Circumstances

The University reserves the right to change the dates and deadlines for any or all courses in extreme circumstances (e.g., severe weather, labour disruptions, etc.). Changes will be communicated through regular McMaster communication channels, such as McMaster Daily News, A2L and/or McMaster email.

## Important Dates

|  |  |
| --- | --- |
| **Description** | **Dates** |
| Classes Begin | Sep 6, 2022 |
| Last day for enrollment and course changes (drop/add) | Sep 14, 2022 |
| National Day for Truth and reconciliation; No classes | Sep 30, 2022 |
| Mid-term recess | Oct 10-16, 2022 |
| Last day for canceling courses without failure by default | Nov 11, 2022 |
| Test and Examination Restriction | Dec 2-8, 2022 |
| Classes End | Dec 8, 2022 |
| Final Exams | Dec 9-22, 2022 |
|  |  |

*Disclaimer: This outline provides a general plan; deviations may be necessary. Any change in the course outline will be posted on the course website, and the details will be announced in class.*