Math 343 / 643
Spring 2025 (3 credits)
Course Syllabus

Adam Kapelner, PhD.

Queens College, City University of New York

document last updated Thursday 30th January, 2025 10:46pm

Course Homepage	`https://github.com/kapelner/QC_Math_343_Spring_2025`
Discord Channel	`https://discord.com/channels/1324190933906096180`
Contact	@kapelner on discord in the relevant channel
Office	604 Kiely Hall
Lecture Time / Loc	Tues and Thurs 5:20 – 6:35PM / KY 258
Instructor Office Hours / Loc	see course homepage
TA(s) / Office Hours Time / Loc	see course homepage

Course Overview

MATH 343 / 643. Computational Statistics for Data Science. 3 hr.; 3 cr. Prereq.: MATH 341 or 641. Coreq.: MATH 342W or 642. Mixture models, EM algorithm, Metropolis-within-Gibbs sampling, permutation tests, the bootstrap, the Kaplan-Meier estimator, the Cox model, T and F tests for the linear model, Gauss-Markov theorem, Bayesian linear regression: Ridge and Lasso. Causality and the randomized experiment, randomization tests. Focus on computation. Special topics. Students cannot receive credit for both: MATH 343 and 643. Fall, Spring

The Four Data Science Core Classes

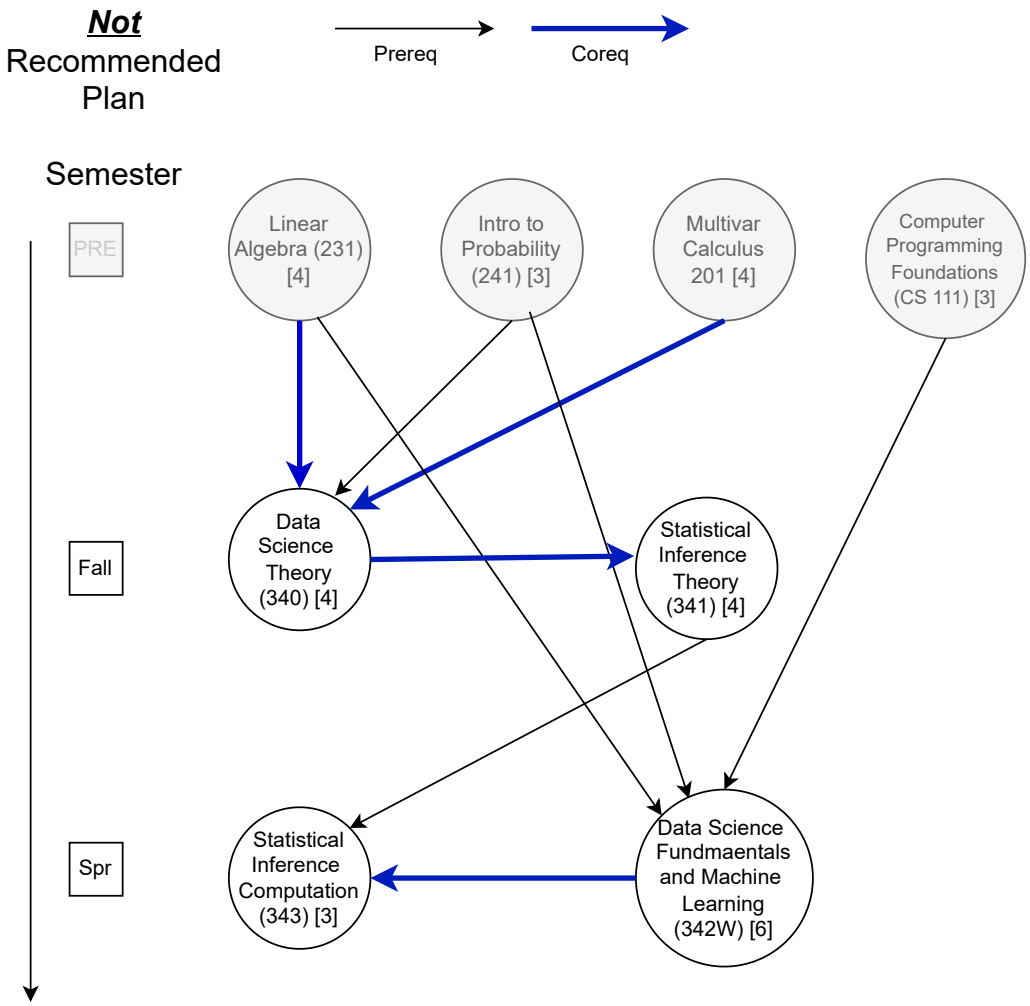
This course is the last and most advanced of the four data science core classes as it builds on the previous three. In Math 340 you learn probability tools and a repertoire of random variable models. In Math 341, you learned the foundations of statistics from both the Frequentist and Bayesian perspectives. Here, we continue with more advanced Frequentist

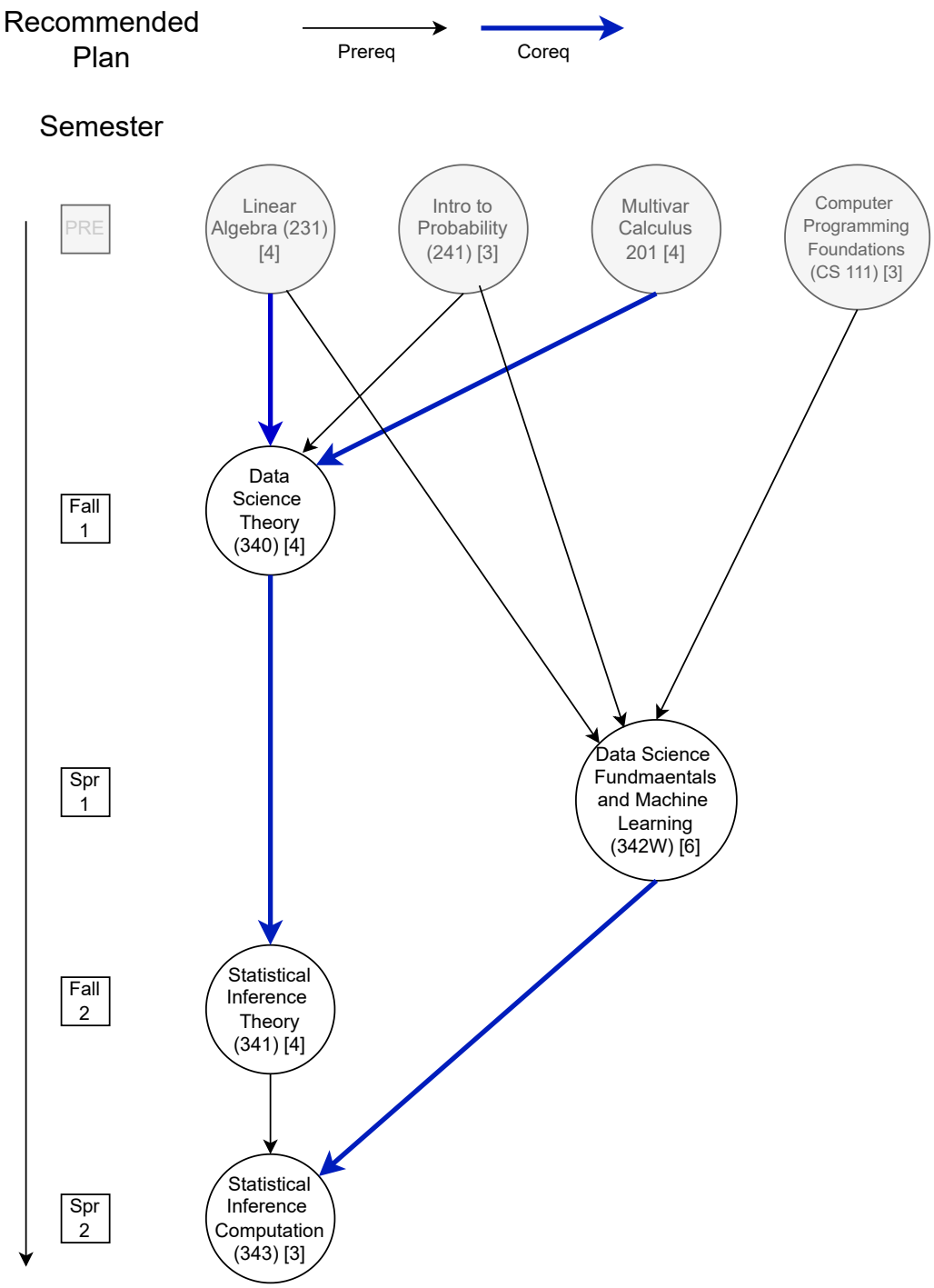
and Bayesian methods. The class will be mostly computational as you now are learning / learned computational skills in Math 342W. Thus, **this is not your typical mathematics course.** This course will do lots of modeling of real-world situations using data via the `R` statistical language.

We will also be able to delve more deeply into the concepts of Math 342W including linear regression from a statistical point of view deriving all the usual linear regression techniques you may have seen in Econometrics. Further, we will also delve into the abyss of modeling and ask questions about correlation vs causation and develop causal models from the ground up. We will discuss the theory of randomization, clinical trials and A/B testing.

If we have time, we will get to advanced machine learning methods such as clustering, reinforcement learning and deep learning.

As we've noted previously, there is an order the four classes need to be taken. Below are two plans, the first is over two semesters and hence it is not *not recommended* as it will be a *very* heavy workload. The second is over four semesters and it is the recommended plan as I believe it will allow students to absorb the material more effectively:




Recommended Plan
Prereq
Coreq
Semester
PRE
Fall 1
Spr 1
Fall 2
Spr 2
Linear Algebra (231) [4]
Intro to Probability (241) [3]
Multivar Calculus 201 [4]
Computer Programming Foundations (CS 111) [3]
Data Science Theory (340) [4]
Data Science Fundmaentals and Machine Learning (342W) [6]
Statistical Inference Theory (341) [4]
Statistical Inference Computation (343) [3]

Examining the above two flowcharts, we note that MATH 341 and 343 form a series of two statistics courses: the first, theoretical with traditional topics and the second, computational with modern topics. Both heavily rely on the theoretical topics taught in MATH 340.

Tentative Day-by-Day Schedule

Lectures and their topics with rough time estimates per topic are below:

D1, Lec 1 [15min] review of Gibbs sampling; [20min] background on Markov chains and invariant distributions; [20min] sketch of proof of why Gibbs sampling converges; [25min] burning and thinning with discussion of autocorrelation

D2, Lec 2 [25min] Poisson count model with hurdle; [25min] the `stan / rstan` package for model-fitting and inference; [25min] Poisson change point model

D3, Lec 3 [50min] Metropolis-Hastings Algorithm with application to Poisson Regression in `stan`; [25min] Permutation testing for different DGP's among two populations

D4, Lec 4 [25min] Choice of test statistics for permutation testing for different DGP's among two populations; [50min] Bootstrap for confidence interval construction and testing with examples

D5, Lec 5 [50min] Estimation for mixture model with Expectation-Maximization algorithm (EM); [25min] Gibbs sampler for mixture model inference

D6, Lec 6 [10min] introduction to survival analysis / churn modeling; [20min] review of the Weibull rv; [10min] invariance of MLE thm; [15min] right censoring at a fixed time point; [20min] Weibull likelihood with right-censoring

D7, Lec 7 [15min] empirical survival function as the sum of indicators; [15min] empirical survival function as the product limit estimator; [15min] inference for survival; [15min] product limit estimator with duplicate survivals; [15min] empirical survival function with right censoring at a fixed time point

D8, Lec 8 [15min] Right censoring at arbitrary times; [30min] Kaplan-Meier survival function estimator; [10min] inference for one-sample Kaplan-Meier survival estimator; [20min] inference for two-sample Kaplan-Meier survival estimator with log-rank test

D9 Midterm I Review

D10 Midterm I

D11, Lec 9 [10min] review of linear model candidate set and optimal element; [20min] motivation of linear model error multivariate normal core assumption; [10min] proof of normality of OLS estimators; [25min] Cochran's theorem applied to error terms; [10min] computing the correct denominator in the unbiased estimator of σ^2

D12, Lec 10 [10min] Proof of the normality of OLS predictions and OLS residuals; [15min] proof that residuals and slope estimators are independent; [30min] proof that the test statistic for the OLS estimators are Student's t-distributed; [20min] proof that the test statistic for the expectation of a response is Student's t-distributed

D13, Lec 11 [15min] definition and motivation of adjusted R^2; [20min] proof that the test statistic for the the response is Student's t-distributed, prediction intervals; [40min] proving the omnibus F test

D14, Lec 12 [10min] proving that R^2 is beta-distribution under the null of no linear signal in the covariates; [65min] proving the partial F test for arbitrarily nested (full / reduced) models

D15, Lec 13 [15min] proof that the OLS estimator is the maximum likelihood estimator (MLE); [60min] proof that the ridge estimator is the maximum a posteriori estimator under iid normal prior on the slope parameters, ridge regression demos, proof that ridge regression is biased

D16, Lec 14 [75min] proof that the lasso estimator is the maximum a posteriori estimator under iid laplace prior on the slope parameters, lasso vs ridge demos, lasso as a "variable selection" preprocessing algorithm

D17, Lec 15 [10min] introduction to robust linear regression inference without the core error assumption; [10min] employing the bootstrap for linear regression inference; [25min] linear regression inference for mean centered homoskedastic errors with asymptotically normal-distributed estimators and Wald testing; [30min] linear regression inference for mean centered heteroskedastic errors with asymptotically normal-distributed Huber-White estimators and Wald testing

D18, Lec 16 [15min] inference for single effects in probability regression models, a type of generalized linear model (GLM) using MLE's and Wald tests; [10min] likelihood ratio test (LRT) with arbitrarily nested (full / reduced) models; [10min] LRT for probability regression models; [15min] Poisson regression, the log-linear mean link function, and inference for single effects with Wald testing and inference for multiple effects with the LRT; [15min] Negative binomial regression, reparameterization to obtain the log-linear mean link function, and inference for single effects with Wald testing and inference for multiple effects with the LRT, discussion of nuisance parameter; [10min] review of Weibull regression with censoring, the log-linear mean link function, inference for single effects with Wald testing and inference for multiple effects with the LRT

D19, Lec 17 [45min] definition of hazart rates / hazard functions, expressing the PDF with the hazard rate and an integral representation of the survival function; [30min] the Cox proportional hazards model (cox PH), the likelihood of the survival DGP in cox PH modeling

D20 Midterm II Review

D21 Midterm II

D22, Lec 18 [35min] expression that allows inference for the cox PH model, demos of cox PH modeling vs Weibull modeling; [15min] spurious correlations; [15min] causal diagrams as directed acyclic graphs (DAG's); [10min] definition of causality as manipulations within the DAG

D23, Lec 19 [20min] Three common DAGs; [10min] spurious correlation vs correlation vs causation, correlation does not imply causation, but implies causation *somewhere*; [10min] the full picture of y, z's, x's from 342; [45min] proper interpretation of coefficients in G/LM's

D24, Lec 20 [15min] more proper interpretation of coefficients in G/LM's; [10min] the impossibility of assessing causation through a scatterplot or regression; [20min] confounding and regressions with and without confounders for the three common DAGs; [15min] proper interpretation of causal coefficients in G/LM's as *manipulations*; [15min] Simpson's Paradox

D25, Lec 21 [10min] Berkson's Paradox; [10min] Fully path-blocked and partially path-blocked DAGs; [10min] Decision tree for all regression scenarios; [45min] introduction to treatment-control experiments (A/B testing)

D26, Lec 22 [15min] Experimental allocation vs experimental design; [25min] unbiasedness of the treatment coefficient for the population average treatment effect in the linear model; [25min] optimal design for the linear model with additive treatment effect; [10min] Finite experimental bias

D27, Lec 23 [10min] Completely randomized design, balanced completely randomized design; [15min] Fisher's randomization test; [10min] Restricted randomization; [10min] Blocking design; [10min] Pairwise-matching design; [10min] Rerandomization design; [10min] multilogit/probit regression introduction

D28 Final Review

Prerequisites

MATH 341 / 641 or a foundations of Frequentist and Bayesian statistics course. Critical is coverage of hypothesis testing, confidence intervals, credible regions, maximum likelihood, priors and posteriors, maximum a posteriori estimation. Implicitly, MATH 340 / 640 is also a prerequisite. At times, we will be drawing on these concepts or extending these concepts (e.g. hazard rates for survival random variables) so keep those notes handy.

Corequisites

MATH 342W / 642 or a foundations of data science course. Critical is coverage of supervised learning, prediction concepts, basic computing using the `R` language, OLS linear regression with matrix algebra and logistic regression.

Course Materials

Textbook: I will be referencing Larry Wasserman's *All of Statistics: A concise course in statistical inference* which can be purchased on Amazon and Casella and Berger's *Statistical Inference* which can be purchased on Amazon. There is no excuse not to have these books. They are *required*. However, I will not ususally be teaching "from the book" — most of the material in the class comes from the lecture notes. The textbooks are a way to get "another take" on the material and they will only cover about only half of the material done in class. For the other half, you will have to make use of other resources. I also recommend Rice's *Mathematical Statistics and Data Analysis*, 3rd edition which can be purchased on Amazon as well but I will not reference it during class.

Computer Software: During lectures, there will be demos using `R` which is a free, open source statistical programming language and console. You can download it from: `http://cran.mirrors.hoobly.com/`.As this course is coreq'd with MATH 342W, this course has a lot of programming in `R` for the homeworks.

Calculator: You can use a TI-84, 85, 89 or any calculator which you wish. I strongly suggest you use Wolfram Alpha and its smartphone app.

Discord Coupled with Github as a Learning Management System

Each assignment and exam will have its own channel. You can feel free to discuss things with your fellow students there. If you are asking me a question, you must do so in the `#discussions` channel for a general questions or the assignment-specific channel (e.g. `#hw03`) so other students can see the question and benefit from the answer. Do not open "issues" on github! If you pm me for help with a class assignment, I will not answer and just ask you to move it to the appropriate public channel. Do not be afraid to ask questions. There are many people who will have your same question!

Discord is to be used professionally so no posting about random stuff!

Announcements

Course announcements will be made via discord in the `#general` channel (not on email). I am known to send a few discord messages per week on important issues.

I can't stress the following enough: if you are not on discord, you will miss all class announcements!!! Discord notifies you when there are messages.

Class Meetings

There are 28 scheduled meetings. Of these, 23 will be lectures, 2 will be midterm exams which are in class and 3 will be review periods during the meeting before the exams (see lecture schedule section above). The exam schedule is given on page 10. The last class of

the semester *may* be rescheduled to be a review period that is conveniently before the final. We will discuss later in the semester.

Homework

There will be 3 theory homework assignments and 3 practice homework assignments (labs). Homeworks will be assigned and placed on the course homepage and will usually be due a week later in class. Homework will be **graded** out of 100 with extra credit getting scores possibly > 100. I will be doing the grading and will grade an *arbitrary subset of the assignment* which is determined after the homework is handed in.

You will submit your finished homework as a PDF by pushing it to your github repo. To set up your github repo, please follow directions on the course homepage (see the link at the top of the README).

To generate the PDF file of your homework, **you must do one of two things:**

- **Print out the homework and handwrite your answers in the alotted space for each question. Then scan your homework as a PDF. There are a ton of good photo $\Rightarrow$ PDF apps for both iPhone and droid.**
- **Open the PDF on your device and use a PDF editing program to electronically handwrite your answers and save the PDF.**

I will NOT accept homework that is not atop the original rendered homework PDF file. Remember to write your name.

You are highly recommended to work with each other and help each other. You must, however, submit your own solutions, *with your own write-up* and in *your own words*. There can be no collaboration on the actual *writing*. Failure to comply will result in severe penalties. The university honor code is something I take seriously and I send people to the Dean every semester for violations.

Philosophy of Homework

Homework is the *most* important part of this course.[1] Success in Statistics and Mathematics courses comes from experience in working with and thinking about the concepts. It's kind of like weightlifting; you have to lift weights to build muscles. My job as an instructor is to provide assistance through your zone of proximal development. With me, you can grow more than you can alone. To this effect, homework problems are color coded green for easy, yellow for harder, red for challenging and purple for extra credit. You need to know how to do all the greens by yourself. If you've been to class and took notes, they are a joke. Yellows and reds: feel free to work with others. Only do extra credits if you have already finished the assignment. The "[Optional]" problems are for extra practice — highly recommended for exam study.

[1] In one student's observation, I give a "mind-blowing homework" every week.

Labs

Labs are the "practice" part of the homework assignment. They will complement the theory homework and are due on the same day also by pushing it to your github repository.

Time Spent on Homework

This is a three credit course. Thus, the amount of work outside of the 2.5hr in-class time per week is 6-9 hours. I will aim for 7.5hr of homework per week on average. You can think of doing the homework well as "sufficient" as *my* job is to ensure that by *you* doing the homework you will study and understand the concepts in the lectures and you won't have all that much to do when the exams roll around.

Late Assignment Policy

Late homework will be penalized 10 points per business day (Monday–Friday save holidays) for a maximum of five days. *Do not ask for extensions*; just hand in the homework late. After five days, **you can hand it in whenever you want** until *the last scheduled class meeting according to the official academic calendar*. As far as I know, this is one of the most lenient and flexible homework policies in college. I realize things come up. Do not abuse this policy; you will fall far, far behind.

LaTeX Homework Bonus Points

Part of good mathematics is its beautiful presentation. Thus, there will be a 1–5 point bonus added to your theory homework grade for typing up your homework using the LaTeX typesetting system based on the elegance of your presentation. The bonus is arbitrarily determined by me.

I recommend using overleaf to write up your homeworks (make sure you upload both the hw#.tex and the preamble.tex file). This has the advantage of (a) not having to install anything on your computer and thus not having to maintain your LaTeX installation (b) allowing easy collaboration with others (c) alway having a backup of your work since it's always on the cloud. If you insist to have LaTeX running on your computer, you can download it for Windows here and for MAC here. For editing and producing PDF's, I recommend TeXworks which can be downloaded here. Please use the LaTeX code provided on the course homepage for each homework assignment.

If you are handing in homework this way, read the comments in the code; there are two lines to comment out and you should replace my name with yours and write your section. The easiest way to use overleaf is to copy the raw text from hwxx.tex and preamble.tex into two new overleaf tex files with the same name. If you are asked to make drawings, you can take a picture of your handwritten drawing and insert them as figures or leave space using the "\vspace" command and draw them in after printing or attach them stapled.

Since this is extra credit, do not ask me for help in setting up your computer with LaTeX in class or in office hours. Also, **never share your LaTeX code with other students** — it is cheating and if you are found I will take it seriously.

Homework Extra Credit

There will be many extra credit questions sprinkled throughout the homeworks. They will be worth a variable number of points arbitrarily assigned based on my perceived difficulty of the exercise. Very high homework scores are not unheard of. They are a good boost to your grade; I once had a student go from a B to and A- based on these bonuses.

Examinations

Examinations are solely based on homeworks (which are rooted in the lectures)! If you can do all the green and yellow problems on the homeworks, the exams should not present any challenge. I will *never* give you exam problems on concepts which you have not seen at home on one of the weekly homework assignments. There will be three exams and the schedule is below.

Exam Schedule

- Midterm examination I will be on [see course homepage] with the first review session on the class meeting prior
- Midterm examination II will be on [see course homepage] with a review on the class meeting prior.
- The final examination will be on [see course homepage] with a review on the final class meeting.

Exam Policies and Materials

I allow you to bring any calculator you wish but it cannot be your phone. The only other items allowed are pencil and eraser. I do not recommend using pen but it is allowed. Food is not allowed during exams but beverages are allowed.

I also allow "cheat sheets" on examinations. For midterm I, you are allowed to bring *one* 8.5" $\times$ 11" sheet of paper (front and back). Two sheets single sided are not allowed. Midterm II, you are allowed to bring *one* 8.5" $\times$ 11" cheat sheet. For the final, you are allowed to bring *three* 8.5" $\times$ 11" cheat sheets. Six sheets single sided are not allowed. On these sheet(s) of paper you can write anything you would like which you believe will help you on the exam. I will be handing back the cheat sheets so you can reuse your midterm cheat sheets for the final if you wish.

Cheating on Exams

If I catch you cheating, you can either take a zero on the exam, or you can roll the dice before the University Honor Council who may choose to suspend you.

Missing Exams

There are no make-up exams. If you miss the exam, you get a zero. If you are sick, I need documentation of your visit to a hospital or doctor. Expect me to call the doctor or hospital to verify the legitimacy of your note.

Accommodations for Students with Disabilities

Candidates with disabilities needing academic accommodation should: 1) register with and provide documentation to the Special Services Office, Frese Hall, Room 111; 2) bring a letter indicating the need for accommodation and what type. This should be done during the first week of class. For more information about services available to Queens College candidates, contact: Special Service Office; Frese Hall, Room 111; 718-997-5870 (Monday – Thursday 8:00 a.m. to 5:00 p.m. & Friday 8:00 a.m. to 4 p.m.).

Class Participation

This portion of your grade is assessed based on your level of interaction during the course lectures e.g. asking / answering questions. Participation on discord counts towards this total.

Grading and Grading Policy

Your course grade will be calculated based on the percentages as follows:

Theory Homework	10%
Labs	10%
Midterm Examination I	20%
Midterm Examination II	20%
Final Examination	35%
Class participation	5%

The semester is split into three periods :

(a) From the beginning until midterm I. Midterm I covers material during this time.

(b) From midterm I to midterm II. Midterm II covers material in this period only.

(c) From midterm II until the final. The final is cumulative and covers all course material.

Each of the periods is assessed evenly. Thus, each period must count the same towards your grade. Since there is 75% of the grade allotted to exams, there is 25% allotted to each period. Thus, the final is upweighted towards the material covered in the third period. In summary, the final will have 5/35 points $\approx$ 14% for the first period's material, 5/35 points $\approx$ 14% for the second period's material and 25/35 points $\approx$ 71% for the last period's material. A good strategy for the final is to just study the material after Midterm II and minimal studying for the previous material.

The Grade Distribution

As this is a small and advanced class, the class curve will be quite generous. I'm expecting approximately 40% A's and 40% B's. If you do your homework and demonstrate understanding on the exams, you should expect to be rewarded with an A or a B. C's are for those who "dropped out" somewhere mid-semester or who cano only demonstrate rudimentary understanding. F's are rare and are for those who miss exams or are not demonstrating any understanding.

Checking your grade and class standing

You can always check your grades in real-time using `https://qc.gradesly.com`. You will enter in your QC ID number (or CUNYfirst email address). I will provide you with your password by email the first week of class.

Auditing

Auditors are welcome. They are encouraged to do all course assignments. I will even grade them. Note that the university does not allow auditors to take examinations.