Instructor: Alex Young, email: alexander_young@fas.harvard.edu
Teaching Fellows:
Lisa Ruan, email: ruan@g.harvard.edu
Evan Arnold, email: evanarnold@college.harvard.edu
Joyce Lu, email: joycelu@college.harvard.edu
Caleb Ren, email: calebren@college.harvard.edu
Philip Nicol, email: phillipnicol@college.harvard.edu
Lectures: Tuesday, Thursday from 3:00 - 4:15 Eastern time.
Course Webpage: canvas.harvard.edu/courses/77317
Office Hours:
Monday: 10:30AM - 12:30PM, 7:00PM - 8:00PM
Wednesday:12:00PM - 2:00PM, 7:00PM - 8:00PM
Friday: 9:00AM - 10:00AM, 12:00PM - 2:00PM, 3:00PM - 4:00PM
Saturday: 3:00PM - 5:00PM
Sunday: 11:00AM - 12:00PM

Sections: This course will not include a section. Rather, the course staff will focusing on holding office hours over a wide range of times to accommodate students from all time zones.

Course Description and Goals: This is an introductory course in dimension reduction. We will cover classical topics such as principal component analysis, nonnegative matrix factorization, and clustering with illustrative applications. The goals of this course is not to provide a complete summary of existing methods or software packages. Rather, using mostly linear algebra, probability, and some coding in R, we will explore selected techniques and their strengths/weaknesses in capturing the curious and often surprising nature of high-dimensional data.

Recommended Texts: This course does not have a required text. All content needed for this course will be presented in class (thus attendance is strongly encouraged). However, the following e-books are great references:
The Elements of Statistical Learning by Hastie, Tibrishani, Friedman
Modern Multivariate Statistical Analysis by Izenman

Attendance: All lectures will be recorded and posted on Canvas so attendance is not required. However, attendance to live lectures is highly encouraged to enable students to ask questions in real time and will include regular breakout rooms allowing students to join their peers in actively working on content during the lesson under supervision of the course staff. To ensure two hours of synchronous class time for each enrolled student, special consideration in the scheduling of office hours will be given to accommodate students in timezones which prevent their attendance in class.

Prerequisites: STAT 110; MATH 21a, 21b or equivalent

Grading: Problem Sets (30%), Group Work (10%), Check-Ins (5%) Midterm (20%), Term Paper (35%)

Problem Sets: A total of seven homework assignments will be assigned as RMarkdown files. Each assignment will contain a written section and a coding section. Students will be asked to complete open portions of the RMarkdown file with answers or code as required and generate a pdf file which they will submit via Canvas. Collaboration with other students is encouraged, but students must write their own solutions in their own words. The lowest homework score will be dropped. *Extensions will be handled on a case by case basis, but will not be accepted if solutions for the assignment have already been posted.*

Group Work: A total of seven group assignments will be assigned throughout the semester. In first five assignments, random groups of 3-4 students will be created and tasked with reviewing key points and concepts from the lectures. For the last two assignments, students will be grouped according to term paper topic. In each case, students will collaborate to write short answers to assess their understanding.

Check-Ins: Each week, there will be two check-ins comprised of one or two multiple choice questions that are expected to take 3-5 minutes to complete. The content of the check-in will come from the previous lecture material. There will be time set aside at the beginning of class to do so. However, students will have a 24-hour window in which to complete the check-in to allow for students in multiple timezones or other issues with the ability to complete quizzes asynchronously as needed.

Midterm Exam: A take home midterm will be posted on TBD and will include theoretical and computational problems. Students will have one week to complete the midterm and post their solutions on Canvas. Collaboration of any type on the midterm is not allowed. *Extensions will not be given unless arrangements are made with the instructor no later than TBD.*

Term Paper: In lieu of a final exam, each students will be asked to complete a term paper written in LaTeX or R Markdown. In the paper, the student will review a method of dimension reduction not covered in the class. The term paper must discuss the mathematical foundation of the method, any necessary assumptions, and its strengths and weaknesses. Examples will be provided. Suggested topics include: Laplacian Eigenmaps, Hessian Eigenmaps, Kernel PCA, Diffusion Maps, MVU, ICA, t-SNE, Spherelets, Functional PCA, and wavelets. However, students are welcome to propose methods pertaining to their own (research) interests. Students must indicate their proposed topic via email to the instructor no later than October 30th. A bibliography and draft of the introduction must be submitted by November 20th. The final paper must then be submitted on Canvas as a single pdf file by December 9th.
Technical and Computational Aspects of the Course: The various dimension reduction techniques discussed in this class ultimately require computation. To balance the technical discussions and ideas presented in class and homework sets, students will be expected to follow guided assignments using R code and to interpret the results. Additionally, the final paper must be completed in an acceptable, legible format. As such, familiarity with R and LaTeX will be beneficial. However, accommodations will be made to assist students develop proficiency with these tools. Tutorials for R and R Markdown may be found at https://www.rstudio.com/online-learning/ and for LaTeX at https://www.latex-tutorial.com/.

Graduate Student Option: Some departments request that courses specify additional course requirements for graduate students wishing to receive credit. In this case, graduate students will be required to complete additional problems on select problem sets, two additional problems on the midterm, and their term paper must contain one complete proof guaranteeing theoretical performance of their selected algorithm. Students using this option should inform the instructor of their intent to use the graduate student option and to discuss an acceptable topic and result for the term paper.

Challenges and Community in the age COVID-19: Remote instruction challenges our ability to develop and support one another in many ways such the brief moments of conversation before and after class and the natural development of study groups. It is the goal of the course staff to support and facilitate these types of interactions through alternative formats and assignments so that we can recreate the innumerable often unseen advantages of in-person classes. In particular, we will make use of the following tools, assignments, and formatting changes this semester:

> **Slack:** A Slack channel for the course, which can be accessed through Canvas, has been created with specific channels allowing students and course staff to interact on questions related to the lectures and assignments.
>
> **Group work:** the primary goal of group is to help students meet one another and establish habits of scheduling times for study groups.
>
> **Check-ins:** The check-ins are intended to provide low stakes opportunities for course staff to regularly assess students and make sure everyone is staying on pace with course material. We have chosen the short multiple choice format so that these assignments are not onerous.
>
> **Lecture format:** To recreate the feel and pacing of a traditional lecture, content will be presented via with handwritten notes rather than prepared slides. Additionally, breakout rooms will be used regularly for small in-class problem sessions and peer-to-peer conversations.

Communication is critical so do not hesitate to reach out to anyone on the course staff should you have questions or concerns about course assignments, content, due dates, or unexpected circumstances which are having a deleterious effect on your ability to get the most out of this class.

Important Dates:

Problem Set Due Dates: September 7, 21; October 5, 19; November 2, 16; December 3
Group Work Due Dates: September 18 ; October 2, 16, 30; November 6, 20; December 2
In Class Midterm: Posted October 19; Due October 26
Term Paper Topic Selection: Friday October 30th, 2020 (via email to the instructor)
Term Paper Introduction Draft and Bibliography: Friday November 20th, 2020

Term Paper Due: December 9, 2020

Tentative Schedule: Following a brief review of multivariate probability and statistics, this course will begin with a review of classic linear dimensionality reduction techniques followed by a review of select nonlinear techniques. The second half of the course will introduce some probabilistic foundations for dimension reduction and conclude with a review of clustering techniques. If time permits, kernel methods will be introduced with connections to earlier topics. A summary calendar is available on the next page.

Week	Date	Topic	Date	Topic
1			9/3	Review: Multivar. Stats. & Linear Algebra
2	9/8	PCA I	9/10	PCA II
3	9/15	PCA III	9/17	SVD I
4	9/22	SVD II	9/24	CCA I
5	9/29	CCA II	10/1	NMF I
6	10/6	NMF II	10/8	NMF III
7	10/13	MDS I	10/15	MDS II
8	10/20	MDS III	10/22	ISOMAP (Midterm Due)
9	10/27	LLE	10/29	Johnson-Lindenstrauss
10	11/3	Johnson-Lindenstrauss	11/5	Hierarchical Clustering
11	11/10	Hierarchical Clustering	11/12	Center-based Clustering
12	11/17	Center-based Clustering	11/19	Spectral Clustering
13	11/24	Model-based Clustering	11/26	Thanksgiving: No class
14	12/1	Kernel Methods	12/3	Kernel Methods II
Finals	12/9	Term Papers Due		