Instructor: Alex Young, email: alexander_young@fas.harvard.edu, office: Science Center 604
Teaching Fellows: TBD
Lectures: Tuesday, Thursday from 10:30 - 11:45
Sections: TBD
Office Hours: The office hour schedule will be updated following a survey of enrollees during the first week of class.

Day	Time	Location
TBD	TBD	TBD

Ed: An Ed 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.

Prerequisites: STAT 110; MATH 21a, 21b or equivalent
Course Description and Goals: This is an introductory course in unsupervised learning. We will cover classical topics such as principal component analysis, nonnegative matrix factorization, and clustering with illustrative applications. This course is not intended 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 central to unsupervised learning. The primary goal of this class is to provide students with a solid foundation to critically assess a novel unsupervised learning method not covered in this class.

Recommended Texts: This course does not have a required text. All content needed for this course will be presented in class. However, the following e-books are helpful references:

An Introduction to Unsupervised Learning by ALY (a working e-book based on the lecture notes)
The Elements of Statistical Learning by Hastie, Tibrishani, Friedman
Modern Multivariate Statistical Analysis by Izenman

Attendance: Attendance to live lectures is expected so that students are able to ask questions in real time. If recording is available, classes will also be recorded and posted to Canvas. Unfortunately, this is highly dependent on our assigned room and is not guaranteed. All lecture notes will be posted on the course website after class for your reference. Additionally, a working textbook based on these notes will be provided.

Important Dates:

Problem Set Due Dates: September 13, 27; October 11; November 11, 15, 26
In Class Midterms: October 17, December 3
Final Report Due: December 10
Oral Exams: December 12 - 19

Grading: Problem Sets (30%), Check-Ins (5%) Midterm (30%), Term Paper/Final Exam (35%)

Problem Sets: A total of six, approximately biweekly problem sets will be assigned as RMarkdown files. Each assignment will contain a combination of theoretical and data/coding problems. Students will be asked to complete open portions of the RMarkdown file with answers or code as required and generate a single pdf file which they will submit via Gradescope. Handwritten solutions of the theoretical portions of the assignments will be accepted, but all solutions must be presented in the order they appear on the assignment. Staff will not compile any files on your behalf when grading assignments.

Collaboration with other students is encouraged and an Ed page will be open for students to post and share thoughts on problem sets and other aspects of the course. However, students must write their own solutions in their own words. The lowest homework score will be dropped. *Extensions will not be offered. Late submissions will be accepted with penalty of 1% per hour the assignment is late.*

Check-Ins: Each week, there will be two check-ins comprised of one or two short questions that are expected to take 3-5 minutes to complete and will be administered at the beginning of each class. The content of the check-in will come from the previous lecture material. The lowest 10 check-in scores will be dropped.

Midterm Exams: (15% each) In-class midterms will be given on October 17th and December 3rd. 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 October 10th.*

Final Report (20%) and Oral Exam (15%): A primary goal of this course is to equip students with the ability to critically assess a new (to them) unsupervised learning method. As such, each student will complete a final report wherein they review an unsupervised learning method not covered in the class. The report must discuss the mathematical and/or algorithmic foundation of the method, any necessary assumptions, and its strengths and weaknesses. Two examples with supporting figures must be included. Suggested topics include: Hessian Eigenmaps, Diffusion Maps,UMAP, ICA, t-SNE, Spherelets, and wavelets. However, students are welcome to propose methods pertaining to their own (research) interests. Students must indicate their proposed topic when they submit problem set 5 on November 11.

Each report is expected to be approximately five to six pages (excluding figures) of single spaced, single column text in 12 point font and one-inch margins. Templates in LaTeXand RMarkdown will be provided through Canvas. The course staff will review the submission, and based on the results, will select three follow up questions to ask during a twenty minute oral exam during the finals period. Scheduling of the oral exams will done through a lottery. The oral exams will be conducted over Zoom and will be recorded for use in grading.

For reference, rubrics for the final report and oral exam will be made available through Canvas.

Technical and Computational Aspects of the Course: The various 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 and are strongly encouraged. 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/.

Policy on the Use of Generative AI: The use of generative AI to draft any problem set solutions or any portion of the final report is not allowed; however, you may use these tools to help proofread your writing and to debug code. Using generative AI to assist with preliminary research in the report including the identification of references is acceptable but should be done with extreme caution. Be aware that in many cases existing software has misstated facts, provided inaccurate references, and in some cases, completely fabricated research results. In other cases, Large Language Models have simply restated the writing from research papers and submitting these passages will be considered plagiarism. You are responsible for the content of your report so carefully review any references provided by AI or revisions to your work. Should you have questions about your use of generative AI, please contact Alex.

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 final third of the course will review a number of clustering techniques. A tentative summary calendar is available on the next page.

Week	Date	Topic	Date	Topic
1	9/3	Multivar. Stats. Review	9/5	Linear Alg. and PCA
2	9/10	PCA	9/12	PCA
3	9/17	SVD	9/19	SVD
4	9/24	NMF	9/26	NMF
5	10/1	MDS	10/3	MDS
6	10/8	Kernels and Kernel PCA	10/10	Kernel PCA
7	10/15	Manifolds	10/17	Midterm (in class)
8	10/22	ISOMAP	10/24	LLE
9	10/29	Laplacian Eigenmaps	10/31	LEs and Autoencoders
10	11/5	Autoencoders	11/7	Hierarchical Clustering
11	11/12	Center-based Clustering	11/14	Spectral clustering
12	11/19	Model-based Clustering	11/21	Model-based Clustering
13	11/26	Assessing clustering	11/28	Thanksgiving Break: No class
14	12/3	Midterm II (in class)	12/5	Reading period begins