Instructor: Alexander (Alex) Young, email: alexander_young@fas.harvard.edu
Teaching Fellow: Jameson Quinn, email: jamesonquinn@fas.harvard.edu
Lectures: Tuesday, Thursday 3:00PM - 4:15PM, Science Center 705
Course Webpage: canvas.harvard.edu/courses/62279
Office Hours:

Alex: Tuesday 4:30 - 5:30, Wednesday 1:30PM - 3:00PM or by appointment in Science Center 604
Jameson: Monday 3:00 - 4:00 in Science Center 109A; Friday 12:30 - 1:30PM in Science Center 111

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:

Foundations of Data Science by Blum, Hopcroft, Kannan
The Elements of Statistical Learning by Hastie, Tibrishani, Friedman
Modern Multivariate Statistical Analysis by Izenman

Prerequisites: STAT 110, MATH 21a, 21b
Grading: Homework (40%), Midterm (20%), Term Paper (40%)
Homework Policies: 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 or html file which they will submit via Canvas. Collaboration with other students is allowed, but students must write their own solutions in their own words. The lowest homework score will be dropped. *Extensions will not be given except in special cases at the discretion of the instructor.*
Midterm Exam: An in-class midterm will be given on October 17th. In conjunction, a take-home coding portion will be posted on October 17th after class. Students will have one week to complete the coding portion of the midterm. 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.*
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. Templates will be provided. Suggested topics include: Laplacian Eigenmaps, Hessian Eigenmaps, Archetypal Analysis, Kernel PCA, Principal Curves/Surfaces, Diffusion Maps, MVU, ICA, t-SNE and Spherelets. 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 Sunday November 10th.
Technical and Computational Aspects of the Course: The various dimension reduction techniques discussed in this class ultimately require computational resources to be feasible. 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/.
Important Dates:

Course Registration Deadline: Monday September 9th
In Class Midterm: Thursday October 17th
Midterm Coding Assignment Released: Thursday October 17th
Midterm Coding Due: Thursday October 24th, 3:00 PM (via Canvas)
Term Paper Topic Selection: Sunday November 10th (via email to the instructor)
Term Paper Due: Saturday December 14th, 2:00 PM (via Canvas)

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 important probabilistic foundations for dimension reduction and conclude with a review of clustering techniques.

Week	Date	Topic	Date	Topic
1	9/3	Course Overview (HW#1 out)	9/5	Review: Multivar. Stats.
2	9/10	PCA I (HW#1 due)	9/12	PCA II
3	9/17	PCA III (HW#2 out)	9/19	CCA I
4	9/24	CCA II (HW#2 due)	9/26	NMF I (HW#3 out)
5	10/1	NMF II	10/3	NMF III
6	10/8	MDS I (HW#3 due, HW#4 out)	10/10	MDS II
7	10/15	MDS III (HW#4 due)	10/17	Midterm (MT coding out)
8	10/22	ISOMAP	10/24	LLE (MT coding due)
9	10/29	Probability Review(HW#5 out)	10/31	Johnson-Lindenstrauss
10	11/5	Compressed Sensing (HW#5 due)	11/7	Hierarchical Clustering (HW#6 out)
11	11/12	Center-based clustering	11/14	Spectral Clustering
12	11/19	Spectral Clustering (HW#6 due, HW#7 out)	11/21	Kernel Methods
13	11/26	Google PageRank (HW#7 due)	11/28	Thanksgiving: No class
14	12/3	Variational Autoencoder	12/4	Fall Reading Period: No class
Finals	12/14	Term Papers Due, 2:00 PM		