The amount of available data is quickly outpacing our ability to understand and use it in meaningful ways. The focus of this course is to teach students some of the ways in which they can make sense of large amounts of unstructured data, with an emphasis towards exploratory data analysis.

The course will be structured seminar-style with a focus on semester-long projects that work on specific data sets and the needs of associated “data clients”.

**Prerequisites**

CS206, CS/MATH 231, Linear Algebra

**Textbooks**


**Class Times and Office Hours**

- Class meets T/Th 12:55pm-2:15pm in Park 337
- Office hours are Tuesdays 3:45pm-5:45pm, and by appointment

**Speaker Series**

There will be a number of visitors giving important and related talks/presentations in this course. You are required to attend each of the talks, as listed below. Please make sure to put them in your calendar now. If you can not attend a talk, please discuss with me within the first week of class. Attendance of these talks count for at least 5% of your total grade.

- Thursday 9/4, 12:55pm-2:15pm (in class) - Richard Freedman (Music, Haverford) and Sorelle Friedler (Computer Science, Haverford)
- Tuesday 9/9, 12:55pm-1:15pm (in class) - Sydne Record (Biology, Bryn Mawr)
- Friday 9/19, 11am-12:30pm - Bruce Maxwell (Distinguished Visitor) and the ImPACT data set
- Thursday 9/25, 12:55pm-2:15pm (in class) - Sydne Record and Joshua Shapiro (Biology, Bryn Mawr)
- Tuesday 11/25, 1pm-2:15pm (in class) - Matthew Wright (Distinguished Visitor), talk on topological data analysis

Planned topics

- Statistical Methods (3 weeks)
  - probability distributions and central limit theorem
  - random numbers and Monte Carlo
  - parameter estimates, maximum likelyhood, least $\chi^2$ fit, confidence intervals and coverage
  - regression analysis
  - Bayesian methods
- Basics of Machine Learning (2-3 weeks)
  - multi-variate regression
  - logistical regression
  - other methods
- Dimensionality Reduction (3 weeks)
  - principal component analysis
  - singular value decomposition
  - Kernel PCA
  - Other non-linear methods, i.e. locally linear embedding, laplacian eigenmaps, metric multidimensional scaling, isomap, semidefinite embedding, etc.
- Topological Data Analysis (3 weeks)
  - Euler characteristics and surface fitting
  - Manifold learning
  - Introduction to TDA
- Graph Theory or Network Analysis (2-3 weeks)
  - Details TBA