CISC/CMPE 251 is the first course in the data analytics certificate, but also makes sense as a standalone introduction to the concepts and practice of data analytics.

Data analytics is a new way of understanding complex systems, and is opening up all sorts of new systems to 'scientific' understanding. There is a chronic shortage of people with the skills to use data analytics, so taking this course (and its follow-ons) opens up new career paths to you, as well as increasing your understanding of the world in the sciences, engineering, business, the social sciences, and the humanities.

It covers these topics:

- Epistemology of data analytics
- Prediction and clustering
- Assessing model quality
- Prediction techniques (Bayes Rule, K-nearest neighbour, decision trees, neural networks, support vector machines, random forests)
- Ensemble techniques
- Attribute selection
- Clustering techniques (k means, expectation-maximisation, hierarchical clustering, density based clustering, online clustering algorithms)
- Visualisation
- Ethical issues in data analytics
- Applications in web search, social networks, natural language, deep learning
- Design principles for data-analytics workflows

For 2020, the course material for this course is made up of a series of narrated powerpoint presentations.

Part of your assessment is based on "class participation" which, in the online asynchronous world means reading and interacting with pdf versions of the slides which I will post, as we go along, on Perusall. Here you can comment, ask questions, or expand on what's talked about in the slides, together in small groups. Your participation marks are based on meaningful interactions on the Perusall site.

The deliverables in the course are designed to help you engage with the material and assess how well you are understanding it.

There are two main skills that you will be learning:

1. The data analytics tools and methods that enable us to build models of complex systems, AND

2.The design skills to know which tools and methods to use when, and how to assemble these into a complete process to analyse and understand a complex system.

Here are the deliverables:

1. Weekly exercises in the first six weeks to get you used to doing pieces of the analytic task and see how the material we are covering fits together: 5 x 2% each = 10%

2. Annotations in Perusall, done in 4 three-week sections: 4 x 5% = 20%

3. A major design project which will occupy the second half of the term: 40%

4. A final exam in two parts; each part will require answering a design question for a particular setting. There will be a ~3 day window to do each part, but once you start on an exam, you will have one hour to complete it: 2 x 15%

There is no official textbook for this course, but you may find:

https://dataminingbook.github.io/book_html/

useful. If you don't find my explanations clear, you can see if you prefer theirs.