Course Syllabus for Data Science: Data to insights
https://mitprofessionalx.mit.edu

Important Course Dates:

- **February 21, 2017** - Course Officially Begins - **04:00 UTC**
- **March 14, 2017** - Mid-Course Survey Available
- **March 22, 2017**- Mid-Course Survey Due
- **April 3, 2017** - Course Assessments Due - **23:30 UTC**
- **April 4, 2017** - Course Officially Ends - **23:30 UTC**
- **April 5, 2017** - Final Course Survey Available
- **April 12, 2017** - Certificates Posted
- **April 12, 2017** - Exclusive LinkedIn Group Invite
- **April 19, 2017** - Final Course Survey Due
- **April 27, 2017** - CEU Awarded
- **July 3, 2017** - Close Archive Course Access

COURSE DESCRIPTION:

Read the full course description here.

Time Requirement/Commitment | Who Should Participate? | Learning Objectives | Course Staff | Course Requirements | Course Schedule: Week 1, 2, 3, 4, 5, 6

TIME REQUIREMENT/COMMITMENT

Taking into consideration various time zones, this course is self-paced with online accessibility 24/7. Lectures are pre-taped and you can follow along when you find it convenient as long as you finish by the course end date. You may complete all assignments before the course end date, however, you may find it more beneficial to adhere to the suggested weekly schedule so you can stay up-to-date with the discussion forums. There are approximately 2 hours of video every week. Most participants will spend about 3 to 4 hours a week on course-related activities. However, when you do the optional case study activities, the time required varies depending on your experience and programming background. We suggest to plan somewhere between 1 to 3 hours.

Please note that for assessment due dates, the edX platform uses Coordinated Universal Time (UTC). To convert times to your local time zone, please use the following tool: http://www.timeanddate.com/worldclock/converter.html

« Back to Top

WHO SHOULD PARTICIPATE?

Prerequisite(s): This course is designed for data scientists and data analysts, as well as professionals who wish to turn large volumes of data into actionable insights. Because of the broad nature of the information, the course is well suited for both early career professionals and senior managers. Participants represented include:

- Technical managers
- Business intelligence analysts
- Management consultants
- IT practitioners
- Business managers
- Data science managers
- Data science enthusiasts

« Back to Top

LEARNING OBJECTIVES

After taking this course, participants will:

- Accelerate learning from research to industry dissemination and expose participants to latest techniques and how to use them;
- Understand common pitfalls in big data analytics and how to avoid them;
- Develop a better understanding of machine learning and how it works in practice;
- Better understand how to interpret model results and what questions you should be asking before you use the results to make business decisions;
- Better understand the challenges and constraints associated with scaling big data algorithms.

Methodology: Online recorded lectures, optional discussion boards, case studies, assessments, and a community Wiki.

Learning Activities Planned for the Program:

- Optional participation in threaded discussions on designated forums
- End of topic assessments
- Video learning sequences

- Wiki for sharing of resources and external links

« Back to Top

COURSE STAFF

Faculty Co-Directors:

Devavrat Shah
Professor, Laboratory for Information and Decision Systems (LIDS), Computer Science and Artificial Intelligence Laboratory (CSAIL) and Operations Research Center (ORC) at MIT

Devavrat Shah is a Professor with the department of Electrical Engineering and Computer Science at MIT. He is the director of Statistics and Data Science Center, Institute for Data, Systems and Society. He is a member of LIDS, CSAIL and ORC at MIT.
His current research interest is in developing large-scale machine learning algorithms for unstructured data with particular interest in social data. He has made contributions to development of “gossip” protocols and “message-passing” algorithms for statistical inference which have been pillar of modern distributed data processing systems.
Devavrat’s work has received broad recognition, including prize paper awards in Machine Learning, Operations Research and Computer Science, and career prizes including 2010 Erlang prize from the INFORMS Applied Probability Society, awarded bi-annually to a young researcher who has made outstanding contributions to applied probability. He is a distinguished young alumni of his alma mater IIT Bombay.
He co-founded Celect, Inc. which helps retailers decide what to put where by accurately predicting customer choice using omni-channel data. His work has been covered in popular press including NY Times, Forbes, Wired and Reditt.

Philippe Rigollet
Associate Professor, Mathematics department and Center for Statistics at MIT

Dr. Rigollet works at the intersection of statistics, machine learning, and optimization, focusing primarily on the design and analysis of statistical methods for high-dimensional problems. His recent research focuses on the statistical limitations of learning under computational constraints.
At the University of Paris VI, Dr. Rigollet earned a B.S. in statistics in 2001, a B.S. in applied mathematics in 2002, and a Ph.D. in mathematical statistics in 2006. He has held positions as a visiting assistant professor at the Georgia Institute of Technology, and as an assistant professor at Princeton University.

See the full list of faculty for this course.

« Back to Top

COURSE REQUIREMENTS

Students must complete a mandatory entrance survey in order to gain access to the videos and other course materials. You will be able to access the survey on the course start date, **February 21, 2017 05:00 UTC**.

In order to get the most out of this course, you are encouraged to watch all course videos, complete all weekly assessments, and actively participate in the discussion forums.

Grading:
Grades are not awarded for this program. However, to earn a Certificate of Completion from MIT Professional Education, you must achieve **an overall assessment average of 80%.** This information will be the "A Avg" column on the course progress screen. MIT Professional Education will not track your video progress, but please note that your understanding of all course content is necessary to complete the course assessments.

Participants who successfully complete all course requirements and earn a Certificate of Completion are eligible to receive 1.3 Continuing Education Units (1.3 CEUs). In order to earn CEUs, participants must complete the Final Course Survey by **April 19, 2017**.

« Back to Top

COURSE SCHEDULE

This course is structured into a 6-week program and is entirely asynchronous. Below is a suggested weekly schedule for the purpose of staying up-to-date with the discussion forums.

Please note that no extensions will be granted, and all required assessments and assignments must be completed and submitted on or before **April 3, 2017, 23:30 UTC**.

Pre-course Assignment: Participants are required to provide some information via a short course entrance survey. Your answers will help the course team and faculty better understand your goals for taking this course and how familiar you are with Data Science concepts, and they will ultimately be a guide to improving your experience and that of future courses.

This survey is your first assignment of the course. You will be able to access the survey on the course start date, **February 21, 2017**. As soon as you complete the survey, you will be granted access to the videos, and can start the course.

Week 1 - Module 1: Making sense of unstructured data
February 21 - February 27
Faculty Leads: Stefanie Jegelka & Tamara Broderick

Introduction

- What is unsupervised learning, and why is it challenging?
- Examples of unsupervised learning

Clustering (*Tamara Broderick*)

- What is clustering?
- When to use clustering
- K-means preliminaries
- The K-means algorithm
- How to evaluate clustering
- Beyond K-means: what really makes a cluster?
- Beyond K-means: other notions of distance
- Beyond K-means: data and pre-processing
- Beyond K-means: big data and nonparametric Bayes
- Beyond clustering

Case Studies:

- Case Study 1: Genetic Codes
- Case Study 2: Finding themes in Project Description

Spectral Clustering, Components and Embeddings *(Stefanie Jegelka)*

- What if we do not have features to describe the data, or not all are meaningful?
- Finding the principal components in data, and applications
- The magic of eigenvectors I
- Clustering in graphs and networks
- Features from graphs: the magic of eigenvectors II
- Spectral clustering
- Modularity Clustering
- Embeddings: new features and their meaning

Case studies:

- Case Study 3: PCA: Identifying Faces

- Case Study 4: Spectral Clustering: Grouping News Stories

Recommended weekly activities:

- Watch course videos for this week
- Try out optional case study activities
- Review and contribute to discussion forum, including module discussion questions (NOTE: Contributing to discussion forums is not required to earn a certificate or CEUs.)
- Review and contribute to Wiki

« Back to Top

Week 2 - Module 2: Regression and Prediction
February 28 - March 6
Faculty Leads: Victor Chernuzkov

Classical Linear and Nonlinear Regression and Extensions

- Linear regression with one and several variable
- Linear regression for prediction
- Linear regression for causal inference
- Logistic and other types of nonlinear regression

Case Studies:

- Case Study 1: Predicting Wages 1
- Case Study 2: Gender Wage Gap

Modern Regression with High-Dimensional Data

- Making good predictions with high-dimensional data; avoiding overfitting by validation and cross-validation
- Regularization by Lasso, Ridge, and their modifications
- Regression Trees, Random Forest, Boosted Trees
- Regression Trees, Random Forest, Boosted Trees

Case Study

- Case Study 3: Do poor countries grow faster than rich countries?

The Use of Modern Regression for Causal Inference

- Randomized Control Trials
- Observational Studies with Confounding

Case Studies:

- Case Study 4: Predicting Wages 2
- Case Study 5: The Effect of Gun Ownership on Homicide Rates

Recommended weekly activities:

- Watch course videos for this week
- Try out optional case study activities
- Review and contribute to discussion forum, including module discussion questions (NOTE: Contributing to discussion forums is not required to earn a certificate or CEUs.)
- Review and contribute to Wiki

« Back to Top

Week 3 - MODULE 3.1: Classification and Hypothesis Testing
March 7 - March 13
Faculty Leads: David Gamarnik & Jonathan Kelner

Hypothesis Testing and Classification:

- What are anomalies? What is fraud? Spams?
- Binary Classification: False Positive/Negative, Precision / Recall, F1-Score
- Logistic and Probit regression: statistical binary classification
- Hypothesis testing: Ratio Test and Neyman-Pearson
- p-values: confidence
- Support vector machine: non-statistical classifier
- Perceptron: simple classifier with elegant interpretation

Case Study

- Case-study 1: Logistic Regression: The Challenger Disaster

Recommended weekly activities:

- Watch course videos for this week
- Try out optional case study activities
- Review and contribute to discussion forum, including module discussion questions (NOTE: Contributing to discussion forums is not required to earn a certificate or CEUs.)
- Review and contribute to Wiki

« Back to Top

Week 4 - MODULE 3.2: Deep Learning
March 14 - March 20
Faculty Leads: Ankur Moitra

Deep Learning

- What is image classification? Introduce ImageNet and show examples
- Classification using a single linear threshold (perceptron)
- Hierarchical representations
- Fitting parameters using back-propagation
- Non-convex functions
- How interpret-able are its features?
- Manipulating deep nets (ostrich example)
- Transfer learning
- Other applications I: Speech recognition
- Other applications II: Natural language processing

Case Study

- Case Study 2: Decision boundary of a deep neural network

Recommended weekly activities:

- Watch course videos for this week
- Try out optional case study activities
- Review and contribute to discussion forum, including module discussion questions (NOTE: Contributing to discussion forums is not required to earn a certificate or CEUs.)
- Review and contribute to Wiki
- Optional mid-course survey will be distributed by **March 14, 2017** and is due **March 22, 2017**

« Back to Top

Week 5 - MODULE 4: Recommendation Systems
March 21 - March 27
Faculty Leads: Devavrat Shah & Philippe Rigollet

Recommendations and ranking

- What does a recommendation system do?
- So what is the recommendation prediction problem? and what data do we have?
- Using population averages
- Using population comparisons and ranking

Collaborative filtering

- Personalization using collaborative filtering using similar users
- Personalization using collaborative filtering using similar items
- Personalization using collaborative filtering using similar users and items

Personalized Recommendations

- Personalization using comparisons, rankings and users-items
- Hidden Markov Model / Neural Nets, Bipartite graph and graphical model
- Using side-information
- 20 questions and active learning
- Building a system: algorithmic and system challenges

Case Studies

- Movies
- Songs
- Products

Wrap-up

- Guidelines on building system
- Parting remarks and challenges

Recommended weekly activities:

- Watch course videos for this week
- Try out optional case study activities

- Review and contribute to discussion forum, including module discussion questions (NOTE: Contributing to discussion forums is not required to earn a certificate or CEUs.)
- Review and contribute to Wiki

« Back to Top

Week 6 - MODULE 5: Networks and Graphical Models
March 28 - April 3
Faculty Leads: Caroline Uhler & Guy Bresler

Introduction
- Introduction to networks
- Examples of networks
- Representation of networks

Networks

- Centrality measures: degree, eigenvector, and page-rank
- Closeness and betweenness centrality
- Degree distribution, clustering, and small world
- Network models: Erdos-Renyi, configuration model, preferential attachment
- Stochastic models on networks for spread of viruses or ideas
- Influence maximization

Graphical models

- Undirected graphical models
- Ising and Gaussian models
- Learning graphical models from data
- Directed graphical models
- V-structures, “explaining away”, and learning directed graphical models
- Inference in graphical models: marginals and message passing
- Hidden Markov Model (HMM)
- Kalman filter

Case Studies

- Case study 1: Navigation / GPS
- 1.1: Kalman Filtering: Tracking the 2D Position of an Object when moving with Constant Velocity
- 1.2: Kalman Filtering: Tracking the 3D Position of an Object falling due to gravity.

- Case study 2: Identifying New Genes that cause Autism

Recommended weekly activities:

- Review all course videos
- Try out optional case study activities
- Complete all assessments by **November 14, 2016 23:30 UTC**
- Final Course Survey will be distributed by **November 16, 2016** and is due by **November 30, 2016**

« Back to Top

Completing the course:

- **April 3, 2017 -** Must complete all end of topic assessments with a minimum of 80% success rate by **23:30 UTC** in order to receive your Certificate of Completion**.**
- **April 4, 2017 -** The course will close on **23:30 UTC**. Course content will be accessible for an additional 90 days post program.
- **April 12, 2017** - Certificates will be posted to your student dashboard
- **April 19, 2017** - Final Course Survey due, which is a requirement to earn CEUs. In order to receive 1.3 CEUs, you must
    - Earn a Certificate of Completion and
    - Complete the final course survey.
- **April 27, 2017** - A CEU award letter will be emailed to all participants that earn them.

POST-COURSE :

- **April 12, 2017** - An invitation will be sent out to join our restricted LinkedIn professional alumni group.
- **July 3, 2017** - The site will officially close. There will be no exceptions or extensions.

Thank you for your participation in
Data Science: Data to Insights
MIT Professional Education
http://professional.mit.edu/