https://mitxpro.mit.edu/

Syllabus for Data Science and Big Data Analytics

OUTLINE

Course Description | Time Requirement/Commitment | Who Should Participate? | Pedagogy | Course Staff | Course Requirements | Course Schedule

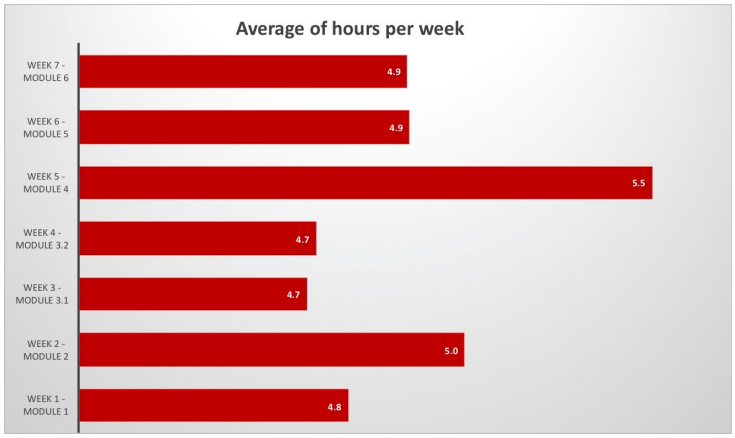
Course Description

Access the full course description on the Course About Page.

Time Requirement / Commitment

This course is accessible online 24/7; most of the course is self-paced. Lectures are pre-taped, and you can follow along at your convenience as long as you submit the two compulsory peer-reviewed case studies and graded activities before the due dates. You may complete all assignments at your own pace. 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 4.5—5 hours per week on watching videos and other 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 planning somewhere between 1 and 3 hours per case study.



*** 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 this time converter tool.

« 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, including:

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

« Back to Top

Pedagogy

Learning Objectives

After taking this course, participants will:

- Accelerate learning from research to industry dissemination and expose participants to the 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;
- Learn how to interpret model results and what questions you should be asking before you use the results to make business decisions;
- Identify the challenges and constraints associated with scaling big data algorithms.

Methodology

Course materials blend the following pedagogical strategies to best achieve the learning objectives of the course and individual modules.

- Instructivism: Teacher-centered learning where the instructors present relevant content (tutorial videos enhanced with animation and graphics). Participants test their knowledge through graded tests.
- Constructivism: Learning-by-doing approach. We encourage participants to construct their own understanding through solving the mandatory and optional case studies and learning through the practice activities.
- Social constructivism: Learning through social interactions and communication. You will discuss and interact with your peers in the discussion forum and evaluate and get reviews from your peers through two compulsory case studies.
- Connectivism: Connecting with others and extending your knowledge through communication. You will be able to expand on and share your knowledge with others through the discussion forum, Facebook, and an exclusive LinkedIn group.

Learning Activities Planned for the Program

- Optional participation in threaded discussions on designated forums
- Graded activities assessments
- 2 graded case studies
- Non-graded practice activities and case studies
- Video learning sequences
- Resources tab
- Optional knowledge sharing and networking with participants through LinkedIn, Facebook, and/or direct contact.

« Back to Top

Course Staff

- Meet your Instructors and TAs: Access to short bios

« Back to Top

Course Requirements

Participants must complete a mandatory entrance survey to gain access to the videos and other course materials. You will be able to access the survey on the course start date: **Monday**, **May 7, 2018 - 05:00 UTC**.

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; the course is pass/fail. To earn an MIT Professional Certificate, you must achieve **an overall completion of 60% of the required activities.** This information will be the "Total" column on the course progress screen. MIT xPRO will not track your video progress, but please note that your understanding of all course content is necessary to

complete the course's graded activities and case studies.

Participants who successfully complete all course requirements earn an MIT Certificate and receive 1.8 Continuing Education Units (1.8 CEUs).

« Back to Top

Course Schedule

This course is a 7-week program. Although most of the course is self-paced and asynchronous, the calendar below suggests a weekly schedule for the purpose of staying up-to-date with the discussion forums and submitting timely the two peer-review case studies and graded activities on time.

Please note that **no extensions will be granted**, and all required assessments and assignments must be completed and submitted on or before **Monday**, **June 25, 2018, at 23:59 UTC**. This date does **not** include the 2 graded case studies (see Modules 4 and 6), which have separate due dates that are clearly indicated below. Still, no extensions will be granted on either of the 2 graded case studies, per official course policy.

Entrance Survey: 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. You will be able to access the survey on the course start date, **Monday, May 7, 2018, 05.00 UTC**. As soon as you complete the survey, you will be granted access to the videos and may start the course.

Module Menu
Module 1 | Module 2 | Module 3.1 | Module 3.2 | Module 4 | Module 5 | Module 6

MODULE	CONTENT

WEEK 1 **Module 1: Making Sense of Unstructured Data** **Dates** May 7 - May 13 **Faculty Leads** Stefanie Jegelka & Tamara Broderick	Compulsory *non-graded Entrance Survey* (complete this survey in order to view the course content). **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

- 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
- ★ Complete Module 1.1 graded activity assessment ***(DUE June 25 — 23.59 UTC)***

Recommended Weekly Activities

- Watch the course videos for this week
- Solve practice activities
- Try out optional case study activities
- Review and contribute to the Discussion Forum

Graded activities

- ★ Complete Module 1.2 graded activity assessment *(DUE June 25— 23.59 UTC)*

← Back to Module Menu

WEEK 2
Module 2: Regression and Prediction

Dates
May 14 - May 20

Faculty Leads
Victor Chernuzkov

Classical Linear and Nonlinear Regression and Extensions

- Linear regression with one and several variables
- 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

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 the course videos for this week
- Solve practice activities
- Try out optional case study activities

	Graded activities • ★ Complete Module 2 graded activity assessment ***(DUE June 25 — 23.59 UTC)*** ← Back to Module Menu
WEEK 3 **Module 3.1 Classification and Hypothesis Testing** **Dates** May 21 - May 27 **Faculty Leads** David Gamarnik & Johnathan 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 the course videos for this week • Solve practice activities • Try out optional case study activities • Review and contribute to the Discussion Forum **Graded activities** • ★ Complete Module 3.1 graded activity assessment ***(DUE June 25 — 23.59 UTC)*** ← Back to Module Menu

WEEK 4
Module 3.2 Deep Learning

Dates
May 28 - June 3

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 the course videos for this week
- Solve practice activities
- Try out optional case study activities
- Review and contribute to the Discussion Forum

Graded activities

- ★ Complete Module 3.2 graded activity assessment ***(DUE June 25 — 23.59 UTC)***

← Back to Module Menu

WEEK 5
Module 4
Recommendation Systems

Dates
June 4 - June 10

Faculty Lead
Devavrat Shah & Phillipe 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

- Solve practice activities
- Try out optional case study activities
- Review and contribute to the Discussion Forum

Graded activities

- ★ Complete Module 4 graded activity assessment ***(DUE June 25 — 23.59 UTC)***
- ★ Graded Case Study
  - Solve and Submit your Case Study *(DUE June 10 — 23.59 UTC)*
  - Review and submit the work of your peer *(DUE June 11 — 23.59 UTC)*

← Back to Module Menu

WEEK 6
Module 5: Networking and Graphical Models

Dates
June 11 - June 17

Faculty Lead
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

Recommended Weekly Activities

- Watch the course videos for this week
- Solve practice activities
- Try out optional case study activities
- Review and contribute to the Discussion Forum

Graded activities

- ★ Complete Module 5 graded activity assessment ***(DUE June 25 — 23.59 UTC)***

← Back to Module Menu

WEEK 7 MODULE 6: Predictive Analytics

Dates
June 18 - June 25

Faculty Lead
Kalyan Veeramachaneni

Predictive Modeling for Temporal Data

- Prediction Engineering

Feature Engineering

- Introduction
- Feature Types
- Deep Feature Synthesis: Primitives and Algorithms
- Deep Feature Synthesis: Stacking

Case Studies

- Case Study 6.1: NYC Taxi Tripts
- Case Study 6.2: UK Retail Dataset

Recommended Weekly Activities

- Watch the course videos for this week
- Solve practice activities
- Try out optional case study activities
- Review and contribute to the Discussion Forum

Graded activities

- ★ Complete ALL graded activity assessment activity ***(DUE June 25 — 23.59 UTC)***
- ★ Graded Case Study

- Retrieve your MIT Course Certificate and 1.8 CEUs *(Download it from your Dashboard — released on June 26 — by 23.59 UTC)*

Course Access

- +180 days of course access (June 25, 2018 — December 22, 2018)
- Course content is archived (no longer access to the course after December 22, 2018)

← Back to Module Menu

« Back to Top

Thank you for your participation in
Data Science and Big Data Analytics: Making Data-Driven Decisions