Course Syllabus for *Tackling the Challenges of Big Data*

https://mitprofessionalx.edx.org

**Important Course Dates:**
Course Officially Begins: February 3, 2015, 12:00 a.m. EST (5:00 a.m. UTC)
Course Assessments Due: March 16, 2015, 7:30 p.m. EDT (11:30 p.m. UTC)
Course Officially Ends: March 17, 2015, 7:30 p.m. EDT (11:30 p.m. UTC)
Certificates Posted: March 24, 2015
CEUs Awarded: April 7, 2015
Post-course access closes: June 15, 2015

**Course Description:**
Read the full course description [here](https://mitprofessionalx.edx.org).

**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 three hours of video every week. Most participants will spend about five hours a week on course-related activities.

Please note that for assessment due dates, the edX platform uses [Coordinated Universal Time (UTC)](http://www.timeanddate.com/worldclock/convert.html), which is 5 hours ahead of Eastern Standard Time (EST). To convert times to your local time zone, please use the following tool: [http://www.timeanddate.com/worldclock/convert.html](http://www.timeanddate.com/worldclock/convert.html)

Last revised on 1/14/15
WHO SHOULD PARTICIPATE?

**Prerequisite(s):** This course was designed to expose participants to the core concepts and topics of Big Data, to teach about some of the trends and evolution in the field, and to give pointers to software, relevant articles, and research material. It may not be appropriate for people already well versed in Big Data who want hands on exposure to new tools and techniques.

This course is designed to be suitable for anyone with a bachelor’s level education in computer science or equivalent work experience, such as working hands-on with IT / technology systems (programming, database administration, data analysis, actuarial work, etc.)

This is not a programming class. Programming experience or knowledge of programming languages is not required.

LEARNING OBJECTIVES

Participants will learn the state-of-the-art in Big Data. The course aims to reduce the time from research to industry dissemination and expose participants to some of the most recent ideas and techniques in Big Data.

After taking this course, participants will:

1. Understand the challenges posed by Big Data (volume, velocity, and variety,) its sources and its potential impact for your industry. Determine how and where Big Data challenges arise in a number of domains, including social media, transportation, finance, and medicine

2. Investigate multicore challenges and how to engineer around them

Last revised on 1/14/15
3. Explore the relational model, SQL, and capabilities of new relational systems in terms of scalability and performance

4. Understand the capabilities of NoSQL systems, their capabilities and pitfalls, and how the NewSQL movement addresses these issues

5. Learn how to maximize the MapReduce programming model: What are its benefits, how it compares to relational systems, and new developments that improve its performance and robustness

6. Learn why building secure Big Data systems is so hard and survey recent techniques that help; including learning direct processing on encrypted data, information flow control, auditing, and replay

7. Discover user interfaces for Big Data and what makes building them difficult

8. Manage the development of data compression algorithms

9. Formulate the “data integration problem”: semantic and schematic heterogeneity and discuss recent breakthroughs in solving this problem

10. Understand the benefits and challenges of open-linked data

11. Comprehend machine learning and algorithms for data analytics

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:

Daniela Rus, Professor, Electrical Engineering Computer Systems

Rus is Professor of Electrical Engineering and Computer Science and Director of the Computer Science and Artificial Intelligence Laboratory (CSAIL) at MIT. Rus’ research interests include distributed robotics, mobile computing, and programmable matter. At CSAIL, she has led numerous groundbreaking research projects in the areas of transportation, security, environmental modeling and monitoring, underwater exploration, and agriculture. Her research group, the Distributed Robotics Lab, has developed modular and self-reconfiguring robots, systems of self-organizing robots, networks of robots and sensors for first-responders, mobile sensor networks, techniques for cooperative underwater robotics, and new technology for desktop robotics. They have built robots that can tend a garden, bake cookies from scratch, cut birthday cake, fly in swarms without human aid to perform surveillance functions, and dance with humans.

Sam Madden, Professor, Electrical Engineering Computer Systems

Madden is a computer scientist specializing in database management systems. He is the faculty director of MIT’s Big Data Initiative at CSAIL and co-director of the Intel Science and Technology Center (ISTC) in Big Data at CSAIL. Recent projects include CarTel, a distributed wireless platform that monitors traffic and on-board diagnostic conditions in order to generate road surface reports, and Relational Cloud, a project investigating research issues in building a database-as-a-service. In 2005, Madden was named one of Technology Review’s “Top 35 Under 35.” He is also co-founder of Vertica (acquired by HP).

See the full list of faculty for this course.

« Back to Top

COURSE REQUIREMENTS

Last revised on 1/14/15
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 3, 2015.

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, you are required to watch all course videos and complete all assessments with an 80% or higher average score. 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 2.0 Continuing Education Units (2.0 CEUs). In order to earn CEUs, participants must complete the final course/CEU survey by March 31, 2015.

**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,
how familiar you are with big data 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 3, 2015. As soon as you complete the survey, you will be granted access to the videos, and can start the course.

**Week 1 - MODULE ONE: INTRODUCTION AND USE CASES**

**February 3 – February 9**

The introductory module aims to give a broad survey of Big Data challenges and opportunities and highlights applications as case studies.

**Introduction: Big Data Challenges (Sam Madden)**

- Identify and understand the application of existing tools and new technologies needed to solve next generation data challenges
- Challenges posed by the ability to scale and the constraints of today's computing platforms and algorithms
- Addressing the universal issue of Big Data and how to use the data to align with a company’s mission and goals

**Case Study: Transportation (Daniela Rus)**

- Data driven models for transportation
- Coresets for Global Positioning System (GPS) data streams
- Congestion aware planning

**Case Study: Visualizing Twitter (Sam Madden)**

- Understand the power of geocoded Twitter data
- Learn how Graphic Processing Units (GPUs) can be used for extremely high throughput data processing
• Utilize MapD, a new GPU based database system for visualizing Twitter in action

**Recommended weekly activities:**

• Watch course videos for this week

• 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 Two: Big Data Collection**

**February 10 – February 16**

The data capture module surveys approaches to data collection, cleaning, and integration.

**Data Cleaning and Integration** ([Michael Stonebraker](#))

• Available tools and protocols for performing data integration

• Curation issues (cleaning, transforming, and consolidating data)

**Hosted Data Platforms and the Cloud** ([Matei Zaharia](#))

• How performance, scalability, and cost models are impacted by hosted data platforms in the cloud

• Internal and external platforms to store data

**Recommended weekly activities:**

• Watch course videos for this week

Last revised on 1/14/15
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

**Week 3 - MODULE THREE: BIG DATA STORAGE**

**February 17 – February 23**

The module on Big Data storage describes modern approaches to databases and computing platforms.

**Modern Databases** ([Michael Stonebraker](https://www.csail.mit.edu/michael))

- Survey data management solutions in today’s market place, including traditional RDBMS, NoSQL, NewSQL, and Hadoop
- Strategic aspects of database management

**Distributed Computing Platforms** ([Matei Zaharia](https://www.csail.mit.edu/matei))

- Parallel computing systems that enable distributed data processing on clusters, including MapReduce, Dryad, Spark
- Programming models for batch, interactive, and streaming applications
- Tradeoffs between programming models

**NoSQL, NewSQL** ([Sam Madden](https://www.csail.mit.edu/sam))

- Survey of new emerging database and storage systems for Big Data
- Tradeoffs between reduced consistency, performance, and availability

Last revised on 1/14/15
• Understanding how to rethink the design of database systems can lead to order of magnitude performance improvements

**Recommended weekly activities:**

• Watch course videos for this week

• 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 midcourse survey will be distributed by February 23, 2015 and is due March 1, 2015.

« Back to Top

---

**Week 4 - MODULE FOUR: BIG DATA SYSTEMS**

**February 24 – March 2**

The systems module discusses solutions to creating and deploying working Big Data systems and applications.

**Security** ([Nickolai Zeldovich](#))

• Protecting confidential data in a large database using encryption

• Techniques for executing database queries over encrypted data without decryption

**Multicore Scalability** ([Nickolai Zeldovich](#))

• Understanding what affects the scalability of concurrent programs on multicore systems

Last revised on 1/14/15
• Lock-free synchronization for data structures in cache-coherent shared memory

**User Interfaces for Data** (David Karger)

• Principles of and tools for data visualization and exploratory data analysis

• Research in data-oriented user interfaces

**Recommended weekly activities:**

• Watch course videos for this week

• 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 5 - MODULE FIVE, PART I: BIG DATA ANALYTICS**

**March 3 – March 9**

The analytics module covers state-of-the-art algorithms for very large data sets and streaming computation.

**Fast Algorithms I** (Ronitt Rubinfeld)

• Efficiency in data analysis

**Fast Algorithms II** (Piotr Indyk)

• Advanced applications of efficient algorithms Scale-up properties

**Data Compression** (Daniela Rus)

Last revised on 1/14/15
• Reducing the size of the Big Data file and its impact on storage and transmission capacity

• Design of data compression schemes such as coresets to apply to Big Data set

**Recommended weekly activities:**

• Watch course videos for this week

• 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 FIVE, PART II: BIG DATA ANALYTICS**

**March 10 – March 16**

The analytics module covers state-of-the-art algorithms for very large data sets and streaming computation.

**Machine Learning Tools** ([Tommi Jaakkola](#))

• Computational capabilities of the latest advances in machine learning

• Advanced machine learning algorithms and techniques for application to large data sets

**Case Study: Information Summarization** ([Regina Barzilay](#))

**Applications: Medicine** ([John Guttag](#))

• Utilize data to improve operational efficiency and reduce costs

Last revised on 1/14/15
• Analytics and tools to improve patient care and control risks
• Using Big Data to improve hospital performance and equipment management

**Applications: Finance** (*Andrew Lo*)

• Learn how big data and machine learning can be applied to financial forecasting and risk management
• Analyze the dynamics of the consumer credit card business of a major commercial bank
• Recognize and acquire intuition for business cases where big data is useful and where it isn't

**Recommended weekly activities:**

• Watch course videos for this week
• 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

**Completing the course:**

• In order to receive a Certificate of Completion, all end of topic assessments must be completed with a minimum of 80% success rate by March 16, 2015, 7:30 p.m. EDT (11:30 p.m. UTC).
• The course will close on March 17, 2015, 7:30 p.m. EDT (11:30 p.m. UTC).
• Certificates will be posted to your student dashboard on March 24, 2015.

Last revised on 1/14/15
• In order to receive 2.0 CEUs, you must earn a Certificate of Completion and complete the final course/CEU survey by March 31, 2015. A CEU award letter will be emailed to all participants that earn them by April 7, 2015.

**Post-course:**

• Course content will be accessible for an additional 90 days post program. There will be no exceptions or extensions. The site will officially close by June 15, 2015.

• An invitation will be sent out to join our restricted LinkedIn professional alumni group one week after certificates have been posted.

**Thank you for your participation in Tackling the Challenges of Big Data.**

MIT Professional Education

[http://web.mit.edu/professional/](http://web.mit.edu/professional/)