Data Mining & Machine Learning

Yong Zheng

Illinois Institute of Technology
Chicago, IL, 60616, USA

ILLINOIS TECH
College of Computing

Intro

- Who am I
- Data Management and Data Science
- Topics in This Class
- Syllabus, Blackboard system and Policy
- Data & Data Types

Intro

- Who am I
- Data Management and Data Science
- Topics in This Class
- Syllabus, Blackboard system and Policy
- Data & Data Types

Who am I

Yong Zheng, PhD
2018 – Present, Assistant Professor at IIT, USA
2016 – 2018, Senior Lecturer at IIT, USA
2017 – Present, Consultant at NPAW, Barcelona, Spain
2016 – 2017, Adjunct Lecturer at DePaul University
2016, PhD in CIS, DePaul University, Chicago, USA
2015, Data Scientist at Pandora, Inc., Oakland, USA
Research: Data Science for Web Intelligence
Particular: Recommender Systems (RecSys)
Website: http://yongzheng.me/

Research Assistants

Qualifications

- Must: Good in math and data science
- Must: Good in programming for algorithm implementations
- Must complete ITMD 514/522/524/525 classes

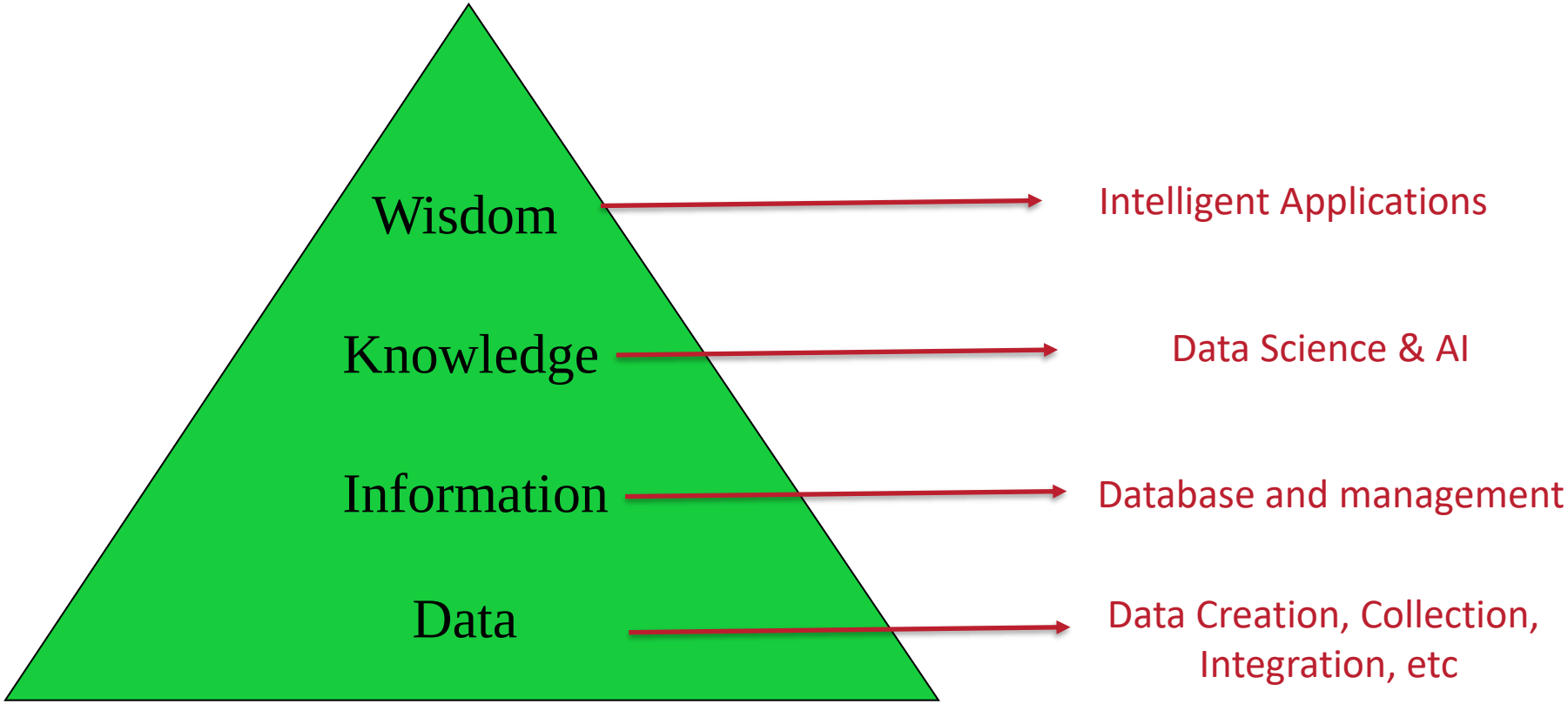
How to be RA

- We do not have funded RA positions in ITM department
- You can take ITMD/ITMT 597 – Independent Study, and work with me on research projects (1-3 credits)
- Or, you can work when you are available anytime

Intro

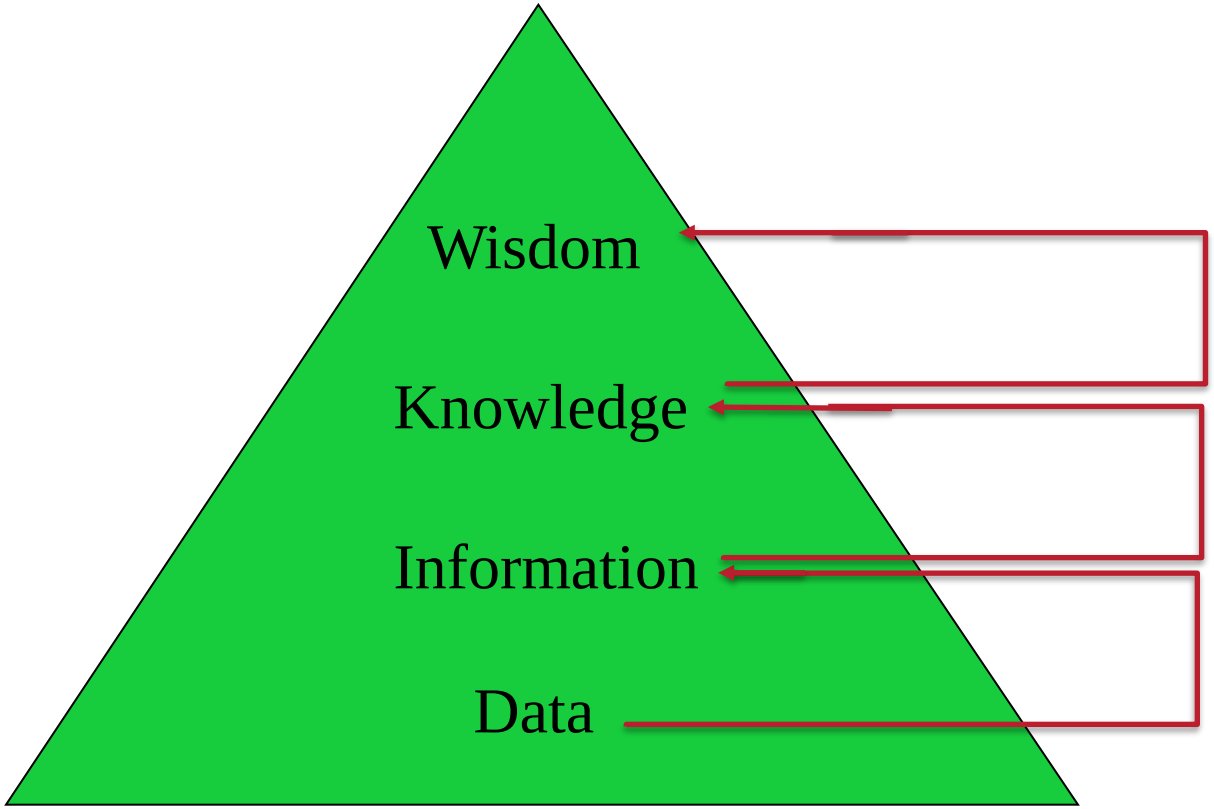
- Who am I
- Data Management and Data Science
- Topics in This Class
- Syllabus, Blackboard system and Policy
- Quick view of the topics

Intro: Data Management



The Information Hierarchy

Intro: Data Management



Business Intelligence

Data Science & AI
ITMD 522, 524/525, 514, 529, 521

Data Preprocessing & Storage
ITMD 522, 523

The Information Hierarchy

Relevant Courses in ITM

- ITMD 514 – Data Analytics
  Introduce and discuss statistical analysis & models
- ITMD 529 – Advanced Analytics
  Extend data analytics and introduce data mining
  It may have several overlaps with ITMD 525
- ITMD 522 – Data Mining & Machine Learning
  Introduce data science and its applications (such as IR, RecSys, Web mining)
- ITMD 521 – Client and Server Techniques
  A big data class with Hadoop and Spark
  But this course focuses more on deployment rather than data science or analysis
- ITMT 524 – Applied AI & Deep Learning
  An AI class which delivers knowledge in classical AI and modern AI (machine learning and deep learning)
- ITMT 525 – Topics in Web Intelligence
  A class for data science/AI applications which delivers knowledge and practical skills in natural language processing (NLP), text mining, sentimental analysis, information retrieval, recommender systems

Relevant Terms and Relationships

Artificial Intelligence [Intelligent Agent]

- Search, rank, logics, neural networks, robotics, etc

Data Mining [Exploration & Knowledge Discovery]

- Classification, clustering, association rules, regressions, outlier detections, etc

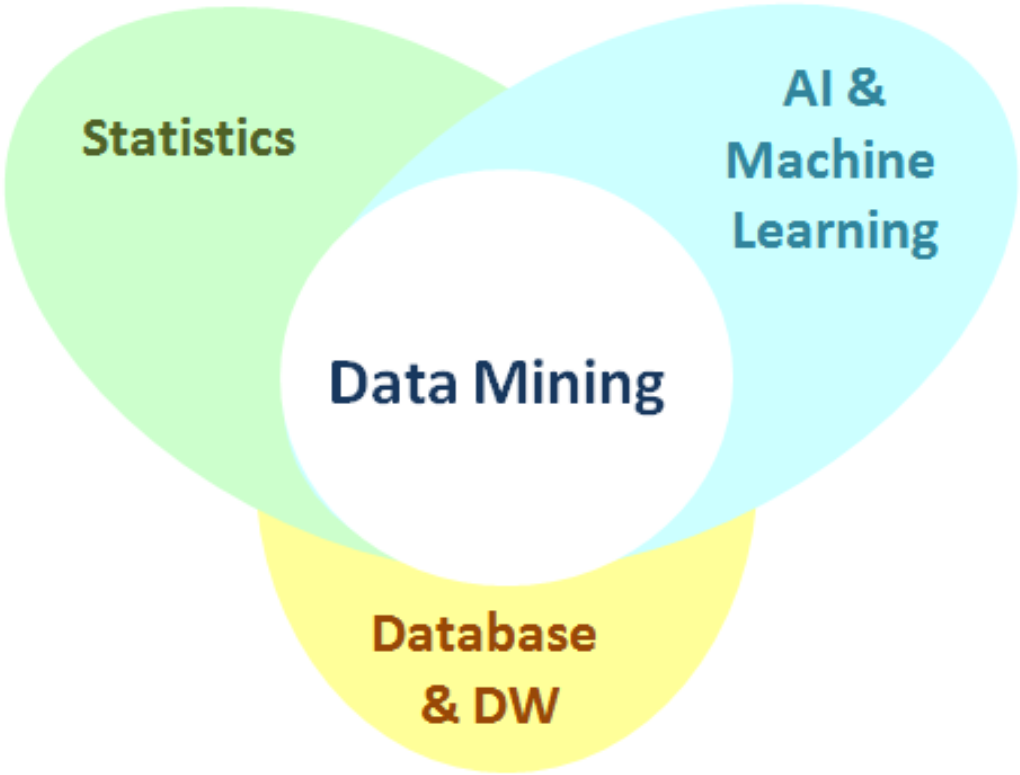
Machine Learning [Learning & Optimizations]

- Supervised and unsupervised leaning, reinforced learning, linear & non-linear learning, etc

Data Science [A combination of the above]

- DS = Statistics + DM + ML + App/Domain Knowledge

Relevant Terms and Relationships



1. Neural Networks: AI & ML
2. Classification: DM & ML
3. BI Analytics: DM & DW
4. Regression: Stat, AI, DM, ML

Data Mining & Machine Learning

- Data mining = Knowledge Discovery from Database (KDD)
  - Extract knowledge from the information
  - We may not know what knowledge can we obtain
- Machine Learning = learning & optimization
- It's necessary to <u>understand</u> the problems, figure out and evaluate potential solutions, finally explain the pattern and apply them in real-practice
- Given a real-world data, you are able to figure out research problems and solutions

Intro

- Who am I
- Data Management and Data Science
- Topics in This Class
- Syllabus, Blackboard system and Policy
- Data & Data Types

ITMD 522: Topics

- Data and Data Preprocessing
- Supervised Learning
  - Classifications
  - Regressions
- Unsupervised Learning
  - Clustering, Association Rules, Outlier Detections
- Semi-Supervised Learning
- Advanced Topics: Neural Networks

Learning Objectives

- Be expert in data preprocessing
- Understand and be able to perform the data mining tasks and use the corresponding methods/algorithms for each task
- Be able to use the knowledge and skills to deal with real-world data sets and applications
- Be able to use Python for data science

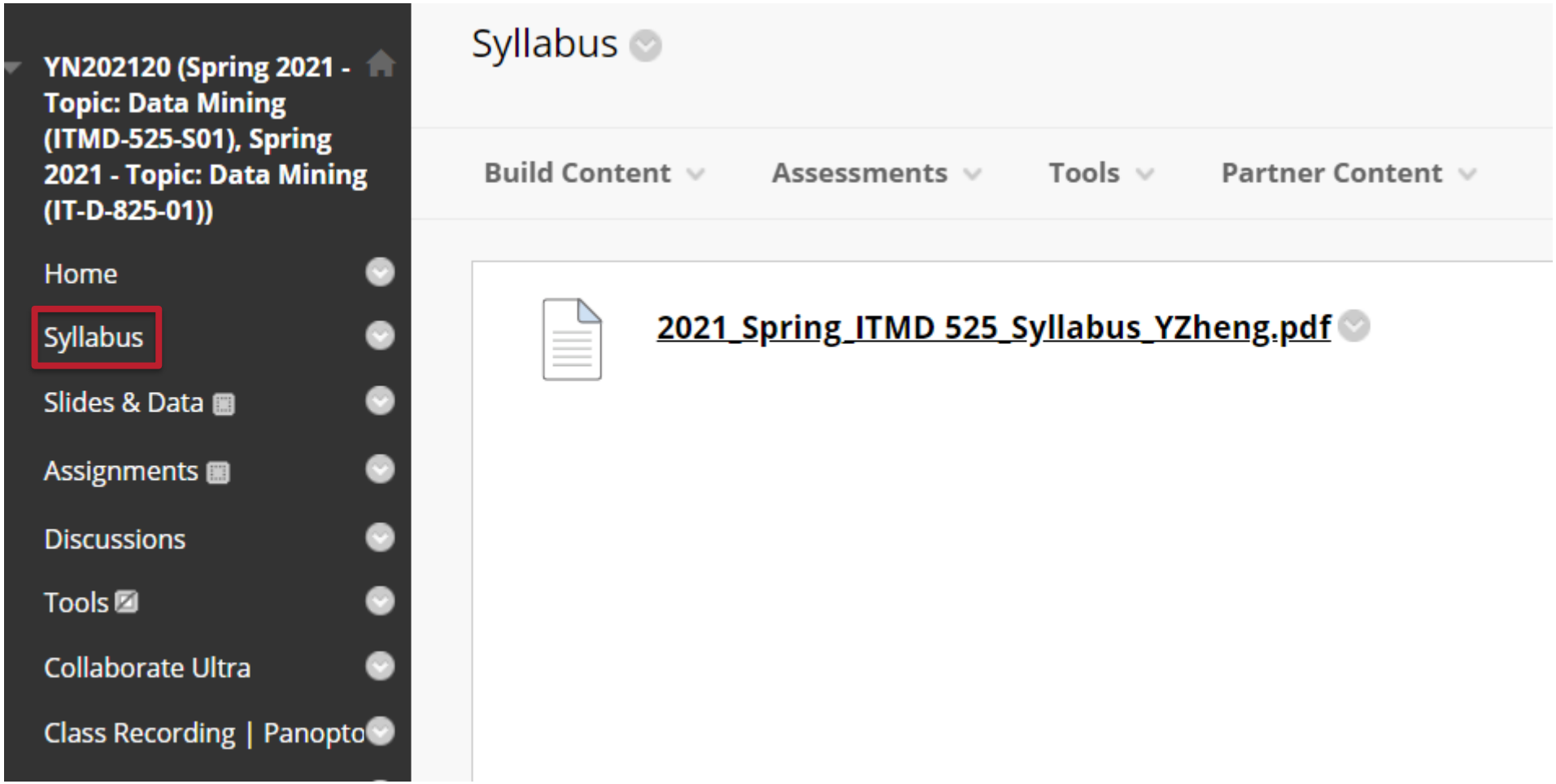
How to Learn Data Science

- First of all, you must focus on understandings
  Ask the following questions frequently
  - Do you know and understand this algorithm works? Try to remember how it works by closing notes/books
  - Do you know a method/algorithm should be used in which situations? And its pros & cons?
- Moreover, practical skills are also important
  - Do you know how to use data mining tools to run them? Do you know how to tune up the parameters?
  - You'd better know Python programming for data science. However, not every students are familiar with Python programming

Intro

- Who am I
- Data Management and Data Mining
- Topics in This Class
- Syllabus, Blackboard system and Policy
- Data & Data Types

Syllabus


YN202120 (Spring 2021 - Topic: Data Mining (ITMD-525-S01), Spring 2021 - Topic: Data Mining (IT-D-825-01))
Home
Syllabus
Slides & Data
Assignments
Discussions
Tools
Collaborate Ultra
Class Recording | Panopto
Syllabus
Build Content
Assessments
Tools
Partner Content
2021_Spring_ITMD 525_Syllabus_YZheng.pdf

ITMD 522

The following policies and introductions are applied to the sections below:

- ITMD 522-01, IT-D 870-01; Live Section
- ITMD 522-02, IT-D 870-02; Online Section
- ITMD 522-03; Remote Students from India
- ITMD 522-04; Remote Students from China

Prerequisite

- ITMD 514 / with Python coding

ITMD 522

Students should attend in-classroom lectures

- ITMD 522-01, IT-D 870-01; Live Section

Students watch recorded videos on Blackboard

- ITMD 522-02, IT-D 870-02; Online Section
- ITMD 522-03; Remote Students from India

Students watch recorded videos on Lumina

- ITMD 522-04; Remote Students from China

ITMD 522

Notes related to Covid-19 Pandemic

- Policy and updates: https://www.iit.edu/reopening
- Optional mask wearing on campus & classrooms
- If you are tested as positives
  - Notify me asap
  - Notify IIT asap (see URL above)

Syllabus and Blackboard

- Time and Place

  Time: Tuesdays & Thursdays, 11:25 – 12:40

  Location: SB-107

  Office Hours

  Thursdays 1:30 – 3:30 PM

  Location: Room 221, Perlstein Hall
  or zoom/google meet by appointment only

  It's better to reserve for office hours by sending emails;

  You can only stand by if you did not reserve;

Syllabus and Blackboard

- Email Rules

1) Clear title
   ITMD522 – I want to …
2) Clear descriptions
   clearly describe what questions you have.

Syllabus and Blackboard

- Example of Bad Emails

Meeting with you Inbox x

Hello Professor,
Are you available today after 3pm? I would like to meet you on certain clarifications. Do let me know you convenient time.
Thank you so much. Have a nice day.

Regards,

- Which course/class?
- What are your questions?
- Which pieces of work (lectures or assignments) you need clarifications?

Syllabus and Blackboard

- Advising Rules

  Questions on lectures

  Questions on class projects & assignments

  Questions on research projects & assistants

  Questions on jobs and careers

  Usually, I do not help you on debugging…

  You can also seek for helps on Discussion Forums

- TA

  TA will grade your assignments and answer your questions

  TA also have office hours for QAs

Syllabus and Blackboard

- Important Dates
  https://www.iit.edu/registrar/academic-calendar

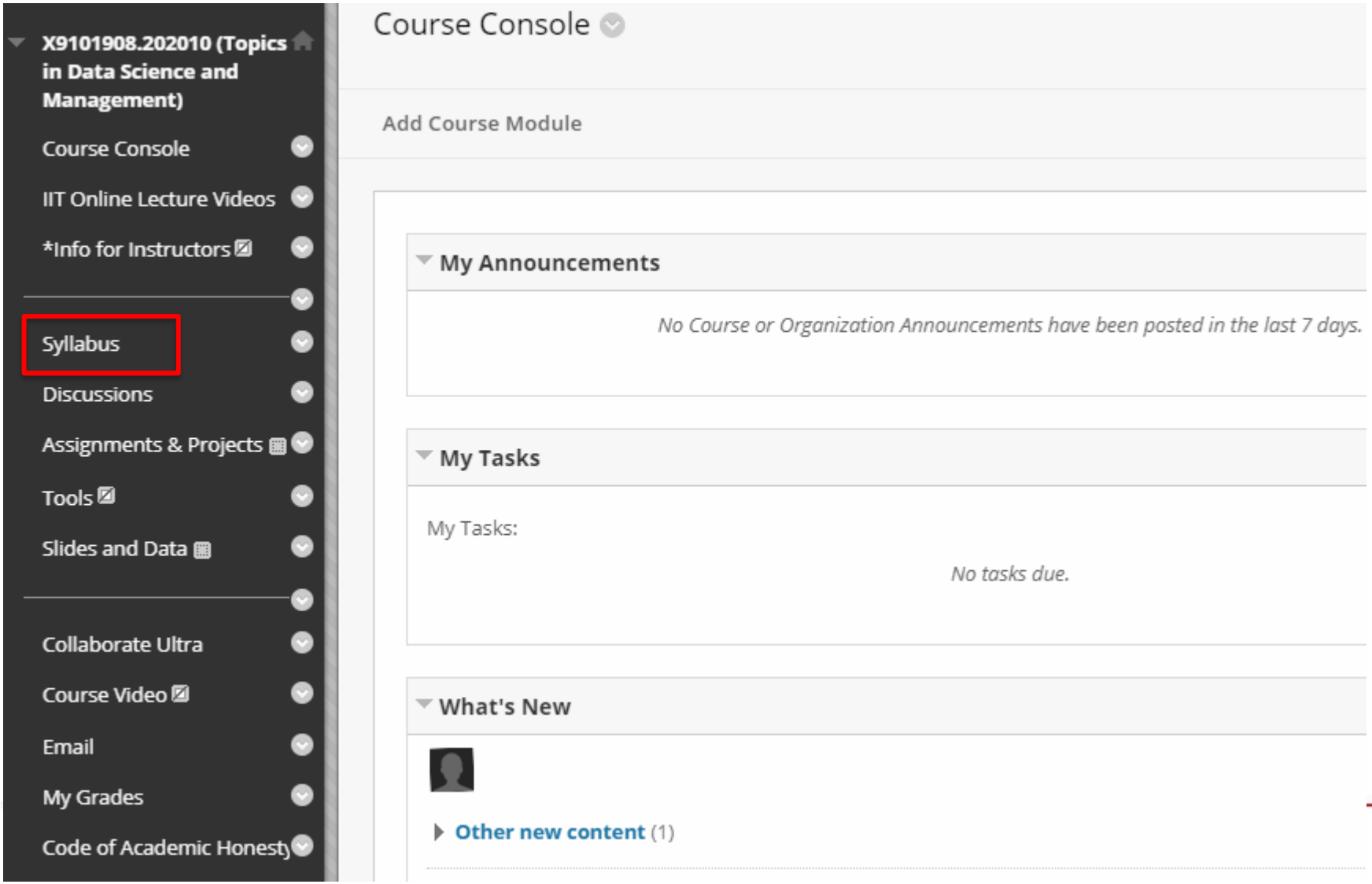
October 31	Last Day to Withdraw for Full Semester Courses
November 7	Spring and Summer Registration Begins
November 15	Spring Reinstatement Applications Due for Undergraduate Students
November 18	Last Day to Withdraw for *ID* B Session Courses
November 23–27	Thanksgiving Break—No Classes
December 3	Last Day of Fall Courses
December 4	Last Day to Request an Incomplete Grade
December 5–10	Final Exam Week/Final Grading Begins on Dec 5
December 14	Final Grades Due at Noon (CST)

Syllabus and Blackboard

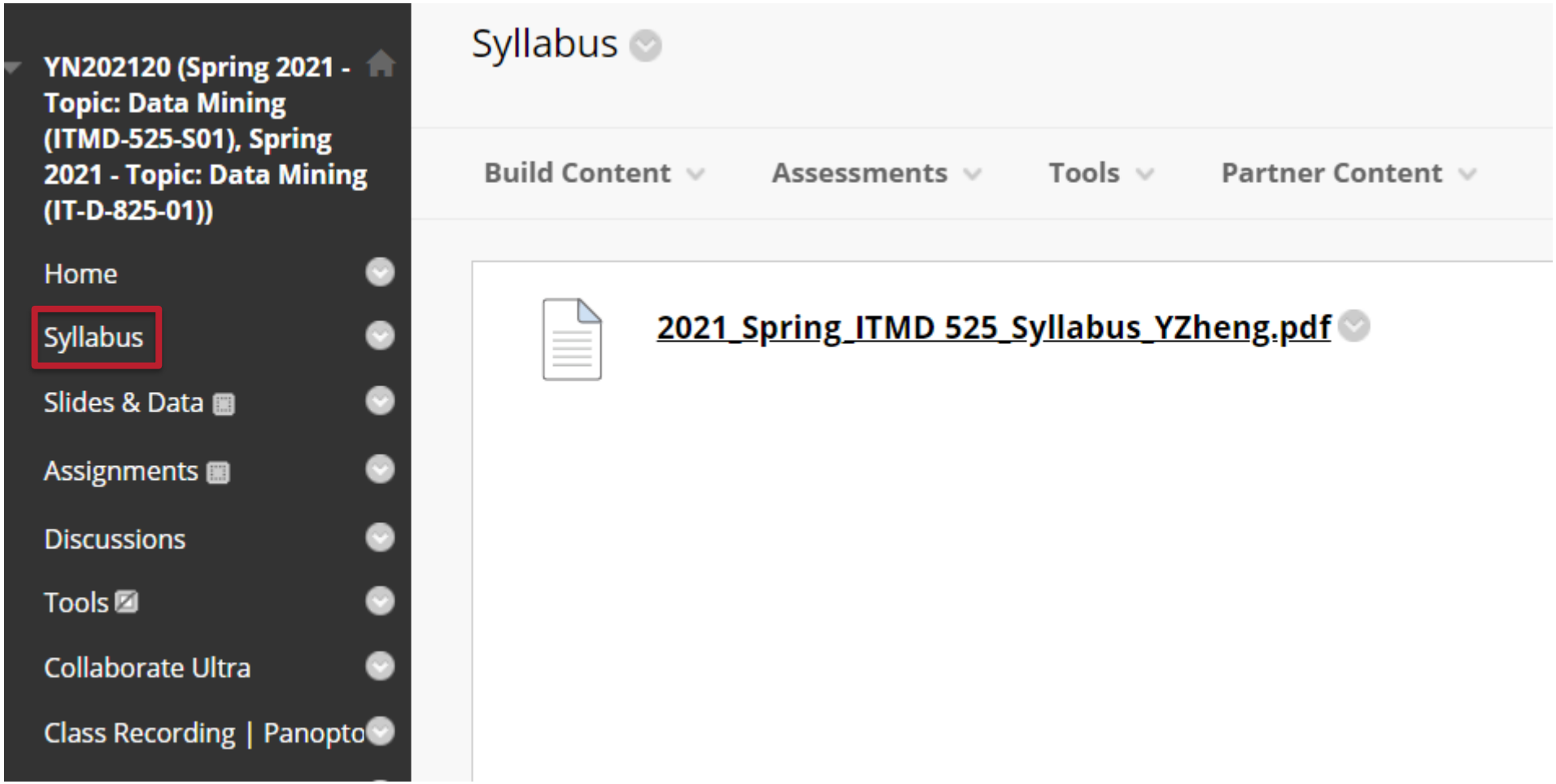
Laptops

- From Fall, 2016, ITM requires students to bring their own laptops to attend the classes.
- We will have in-class practice (almost everyweek), and you should bring your laptop to the class.

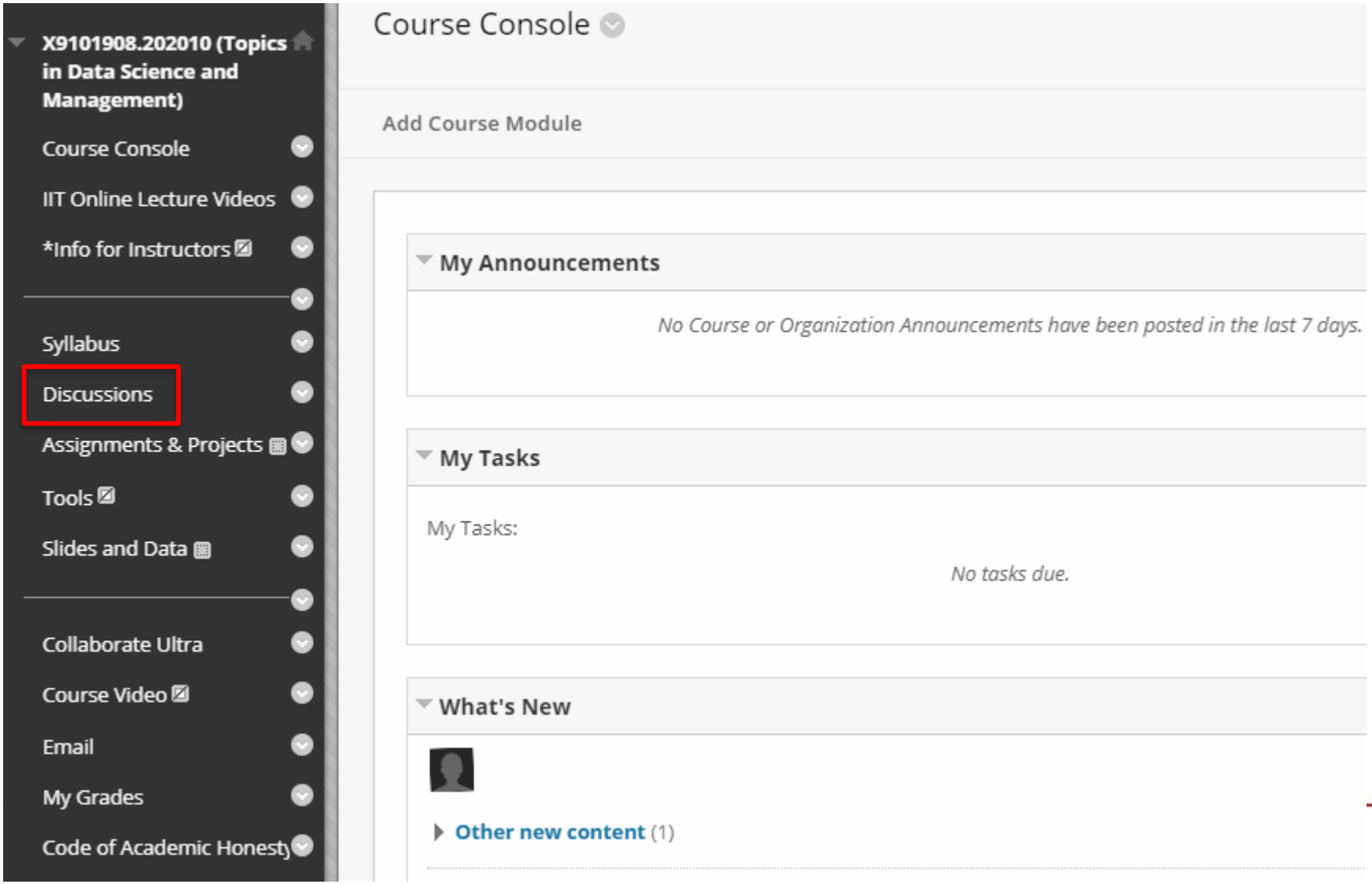
Blackboard


X9101908.202010 (Topics in Data Science and Management)
Course Console
IIT Online Lecture Videos
*Info for Instructors
Syllabus
Discussions
Assignments & Projects
Tools
Slides and Data
Collaborate Ultra
Course Video
Email
My Grades
Code of Academic Honesty
Course Console
Add Course Module
My Announcements
No Course or Organization Announcements have been posted in the last 7 days.
My Tasks
My Tasks:
No tasks due.
What's New
Other new content (1)

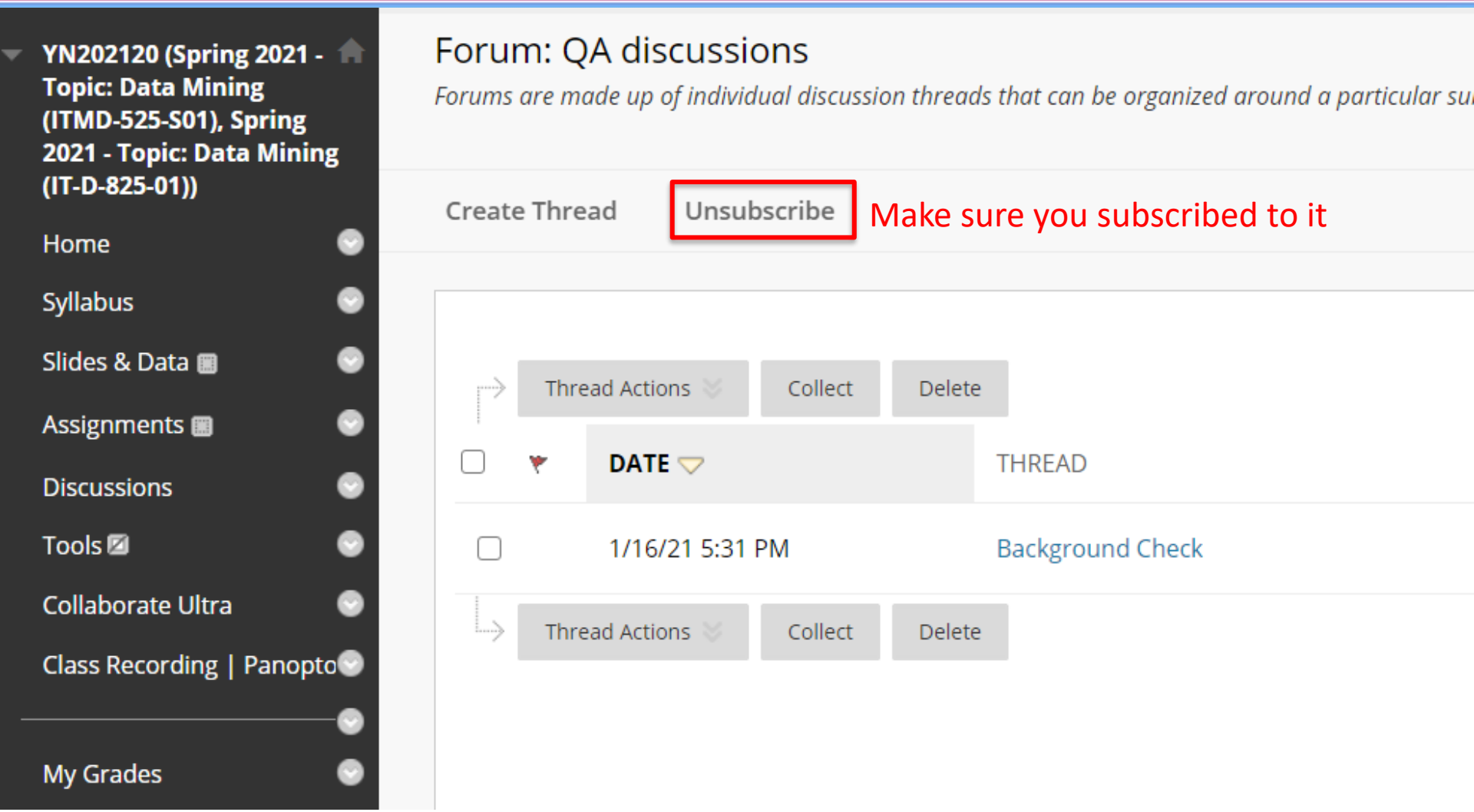
Syllabus


YN202120 (Spring 2021 - Topic: Data Mining (ITMD-525-S01), Spring 2021 - Topic: Data Mining (IT-D-825-01))
Home
Syllabus
Slides & Data
Assignments
Discussions
Tools
Collaborate Ultra
Class Recording | Panopto
Syllabus
Build Content
Assessments
Tools
Partner Content
2021_Spring_ITMD 525_Syllabus_YZheng.pdf

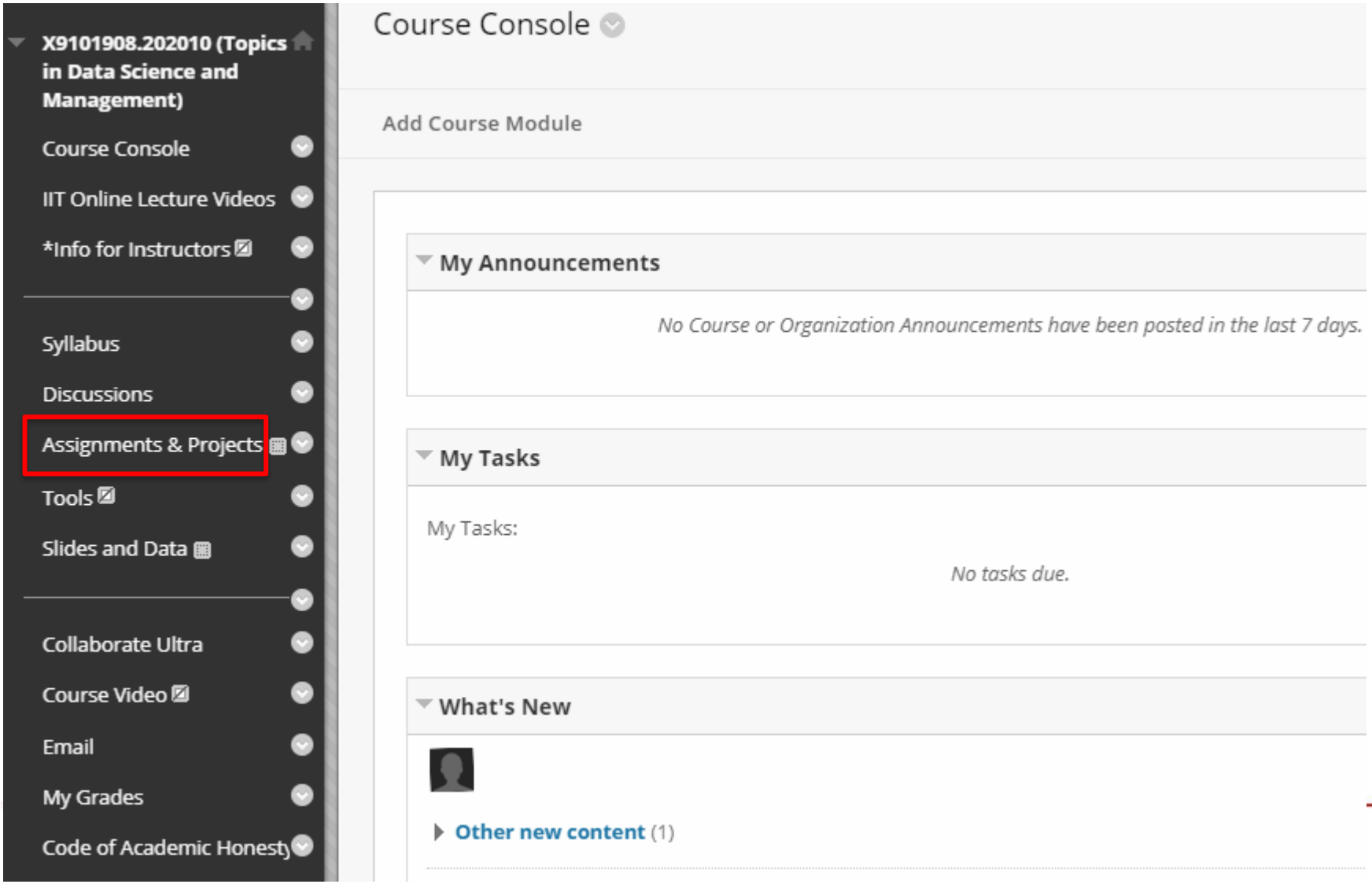
Blackboard


X9101908.202010 (Topics in Data Science and Management)
Course Console
IIT Online Lecture Videos
*Info for Instructors
Syllabus
Discussions
Assignments & Projects
Tools
Slides and Data
Collaborate Ultra
Course Video
Email
My Grades
Code of Academic Honesty
Course Console
Add Course Module
My Announcements
No Course or Organization Announcements have been posted in the last 7 days.
My Tasks
My Tasks:
No tasks due.
What's New
Other new content (1)

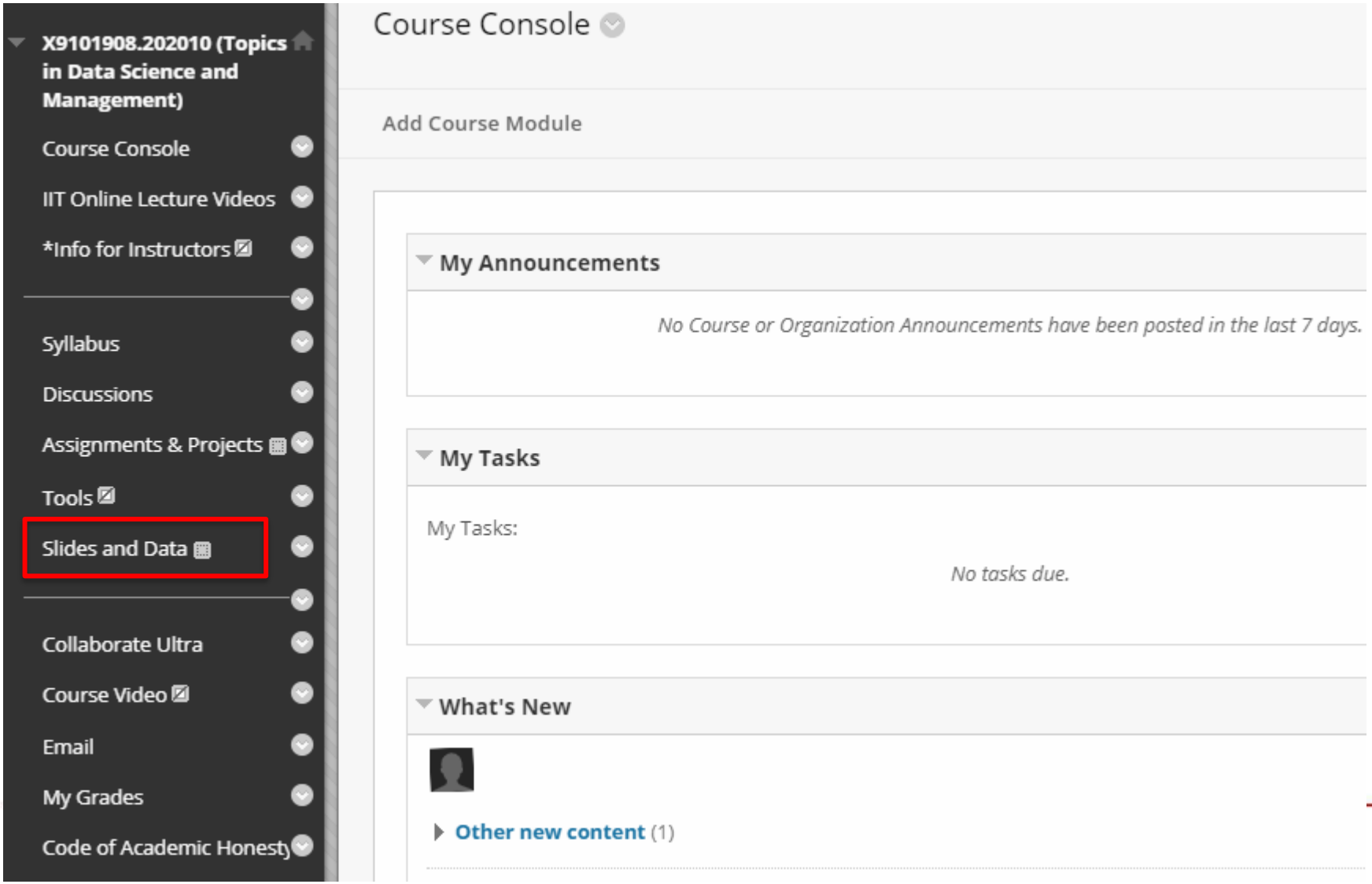
Blackboard


YN202120 (Spring 2021 - Topic: Data Mining (ITMD-525-S01), Spring 2021 - Topic: Data Mining (IT-D-825-01))
Home
Syllabus
Slides & Data
Assignments
Discussions
Tools
Collaborate Ultra
Class Recording | Panopto
My Grades
Forum: QA discussions
Forums are made up of individual discussion threads that can be organized around a particular su
Create Thread
Unsubscribe
Make sure you subscribed to it
Thread Actions
Collect
Delete
DATE
THREAD
1/16/21 5:31 PM
Background Check
Thread Actions
Collect
Delete

Blackboard


X9101908.202010 (Topics in Data Science and Management)
Course Console
IIT Online Lecture Videos
*Info for Instructors
Syllabus
Discussions
Assignments & Projects
Tools
Slides and Data
Collaborate Ultra
Course Video
Email
My Grades
Code of Academic Honesty
Course Console
Add Course Module
My Announcements
No Course or Organization Announcements have been posted in the last 7 days.
My Tasks
My Tasks:
No tasks due.
What's New
Other new content (1)

Blackboard


X9101908.202010 (Topics in Data Science and Management)
Course Console
IIT Online Lecture Videos
*Info for Instructors
Syllabus
Discussions
Assignments & Projects
Tools
Slides and Data
Collaborate Ultra
Course Video
Email
My Grades
Code of Academic Honesty
Course Console
Add Course Module
My Announcements
No Course or Organization Announcements have been posted in the last 7 days.
My Tasks
My Tasks:
No tasks due.
What's New
Other new content (1)

ITMD 522

Students should attend in-classroom lectures

- ITMD 522-01, IT-D 870-01; Live Section

Student should watch recorded videos by themselves

- ITMD 522-02, IT-D 870-02; Online Section
- ITMD 522-03; Remote Students from India

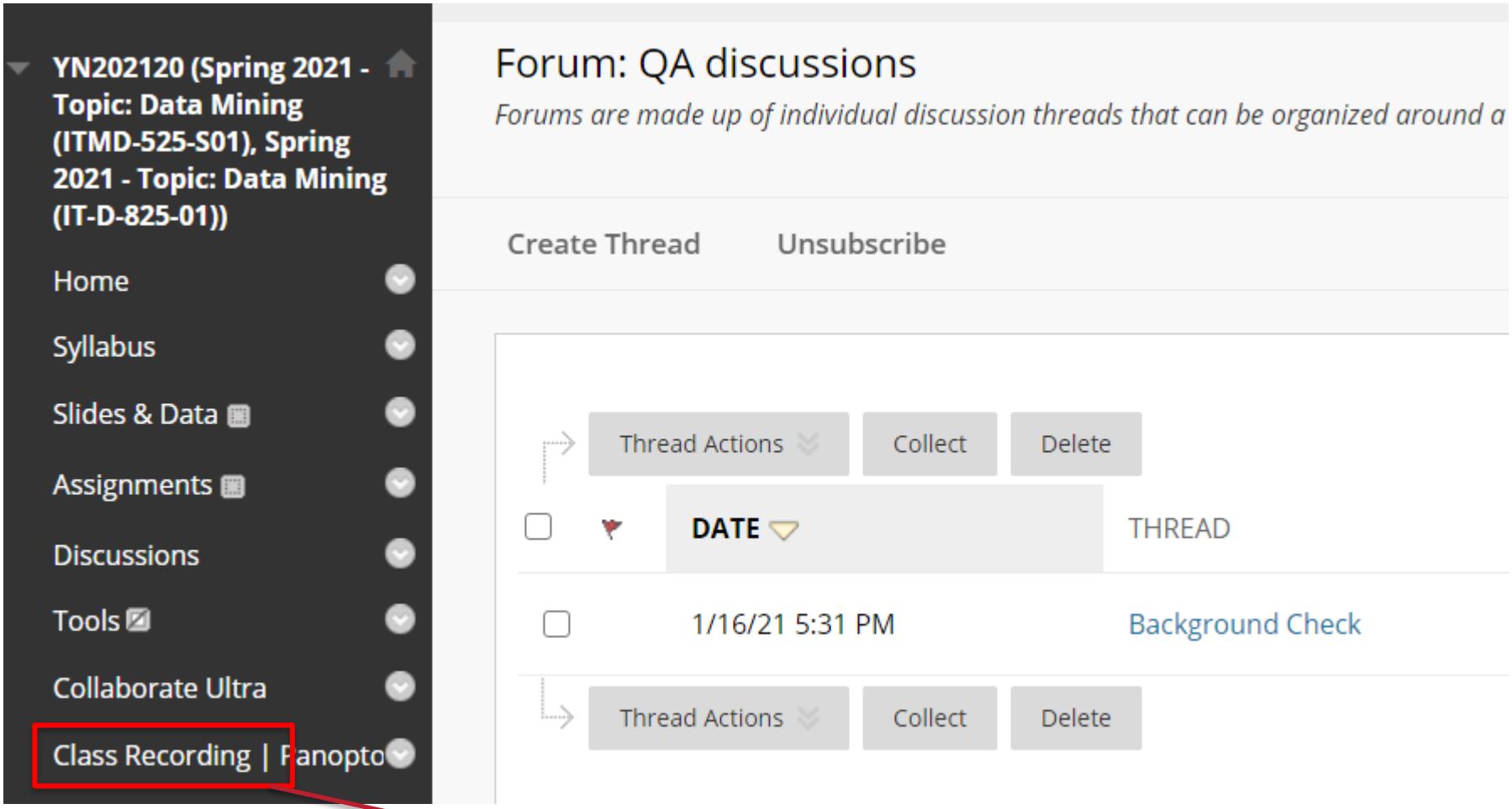
Student watch recorded videos on Lumina

- ITMD 522-04; Remote Students from China

Online Lectures, if needed

- On-Campus Lectures
  Lectures will be recorded by school, if IIT allows us to take classes in a classroom on campus
- Collaborative Ultra
  By default, we use it for online teaching.
  You will find recorded videos (see last page)
- Recording: Others
  I may pre-record lectures and share the video with you on Google drive, if I cannot give on-campus lectures

Blackboard

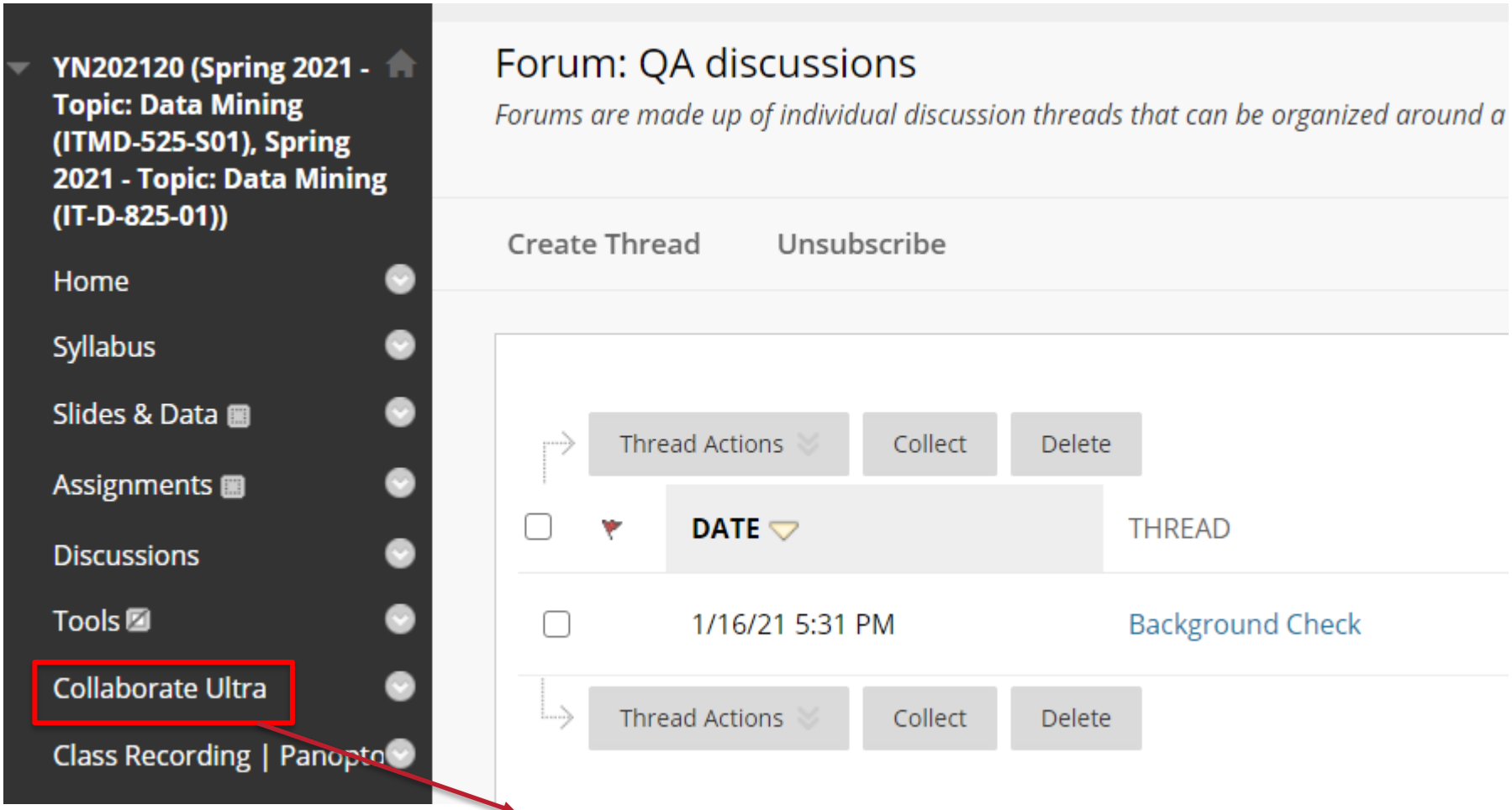


Recordings if it is recorded in the classroom on campus

Online Lectures, if needed

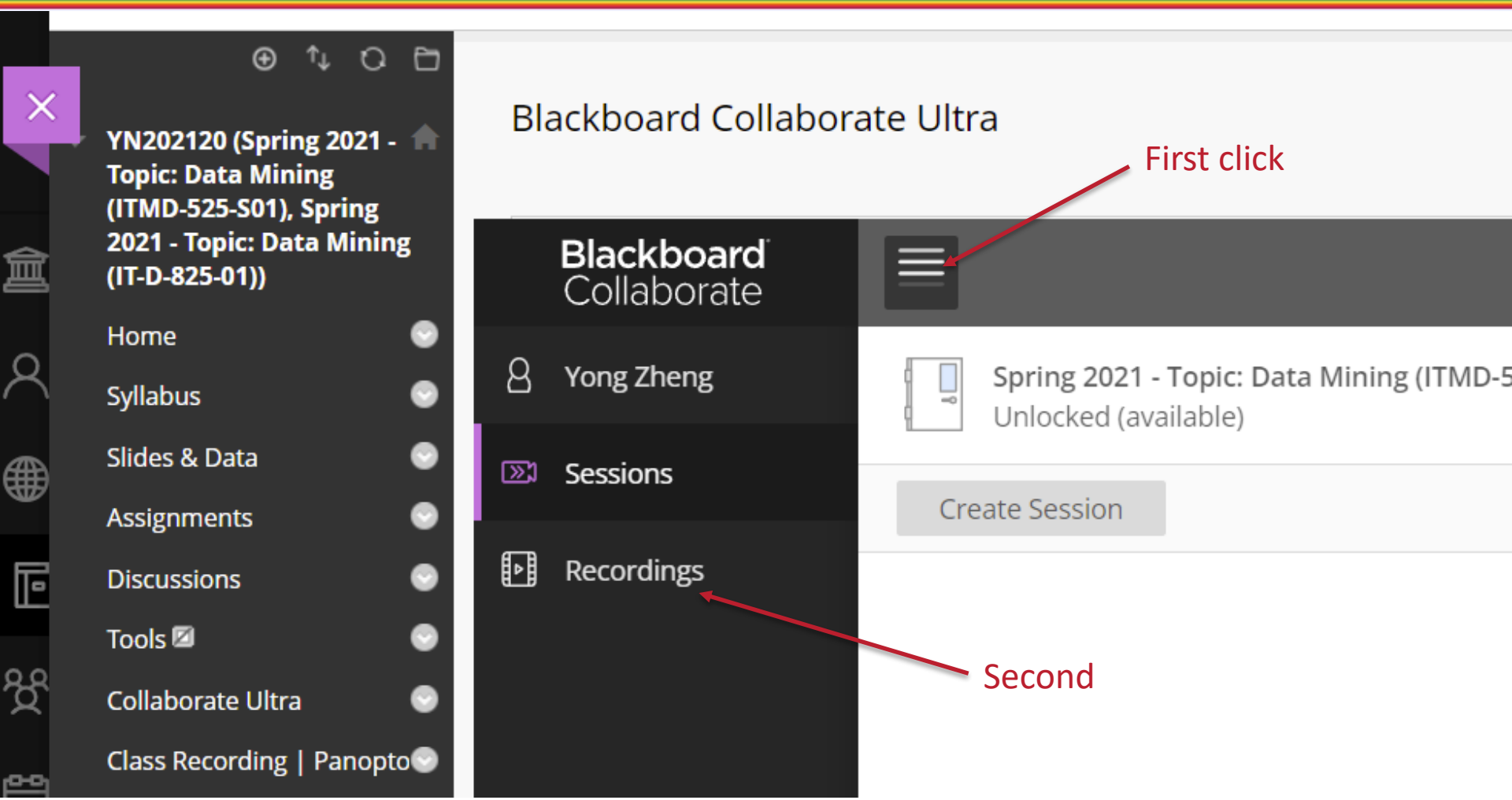
- On-Campus Lectures
  Lectures will be recorded by school, if IIT allows us to take classes in a classroom on campus
- Collaborative Ultra
  By default, we use it for online teaching.
  You will find recorded videos (see next page)
- Recording: Others
  I may pre-record lectures and share the video with you on Google drive, if I cannot give on-campus lectures

Blackboard



Online teaching and recordings by Blackboard

Blackboard


YN202120 (Spring 2021 - Topic: Data Mining (ITMD-525-S01), Spring 2021 - Topic: Data Mining (IT-D-825-01))
Home
Syllabus
Slides & Data
Assignments
Discussions
Tools
Collaborate Ultra
Class Recording | Panopto
Blackboard Collaborate Ultra
First click
Blackboard Collaborate
Yong Zheng
Sessions
Recordings
Spring 2021 - Topic: Data Mining (ITMD-5
Unlocked (available)
Create Session
Second

Online Lectures, if needed

- On-Campus Lectures
  Lectures will be recorded by school, if IIT allows us to take classes in a classroom on campus
- Collaborative Ultra
  By default, we use it for online teaching.
  You will find recorded videos (see last page)
- Recording: Others
  I may pre-record lectures and share the video with you on Google drive, if I cannot give on-campus lectures

Syllabus and Blackboard

Textbooks

- Peter Flach. "Machine Learning: The Art and Science of Algorithms that Make Sense of Data", Cambridge University Press; 1 edition; ISBN-10: 1107422221, ISBN-13: 978-1107422223
- Jake VanderPlas. "Python Data Science Handbook: Essential Tools for Working with Data 1st Edition", O'Reilly Media; 1 edition; ISBN-10: 1491912057, ISBN-13: 978-1491912058

Syllabus and Blackboard

Assignments & Exams

- Written Assignments
  Examine your understanding and skills in DS
  May ask you to process real-world data sets
  May ask you to use Python for practice
- Exams: open-book/notes exams
- Final Project: could be team project; more details will be given later

Syllabus and Blackboard

About Final Project

- Write a project proposal
  Introduce your research problems, which data you will use, what are the solutions and evaluations, what are the expected outcomes
- Work on the experiments
- Present your work and submit final project reports
  For remote students, you can record a video of your presentation and send it to me

Syllabus and Blackboard

Examinations

- Exam: open books/notes, and written exam.
- Final Project and presentations: could be individual or team project; you should present your project and submit project reports eventually. More details will be given later

Syllabus and Blackboard

Rules in Assignments and Examinations

- Assignments

  Usually, no late submission is allowed

  15% penalty will be applied for late submission (in 1 week)

  Late submission will be ignored if later than 1 week

  e.g., due date is March 1st 11:59 PM

  No penalty if submitted before due

  15% penalty if between due time and Mar 8th 11:59 PM

  A zero score if submitted later than Mar 8th 11:59 PM

Syllabus and Blackboard

Scales for your final grades

Grading: Grading criteria for this course will be as follows:

A	*Outstanding work reflecting substantial effort*	90-100%
B	*Adequate work fully meeting that expected of a graduate student*	75-89.99%
C	*Satisfactory work meeting minimum expectations*	60-74.99%
F	*Unsatisfactory work*	0-59.99%

The final grade for the class will be calculated as follows:

Regular Assignments	**28%**
Exam	**30%**
Final Project Presentations	**40%**
Class Attendance	**2%**

Syllabus and Blackboard

Attendance

- Live Sections: I will ask you to sign your name and student ID on a sheet
- Online Sections: I will ask TA to release Google forms on Tuesday, you need to sign by Friday

Syllabus and Blackboard

Academic Honesty

- Plagiarism is a very serious problem, and it is forbidden in all submissions, including assignments, paper reviews, midterm exams and final projects, as well codings/reports in these submissions.
- New rules from Fall, 2018 in ITM department ➔ you cannot share assignments/exams online, and some instructors may ask you not to share slides

Syllabus and Blackboard

So, what is plagiarism?

- Cheating in the exams or final projects
- Cheating or copying answers from other students or other resources (such as online materials) in assignments, exams or final projects
- Cheating or copying texts from other resources without references in paper writing or reviews
- Share answers with others

Syllabus and Blackboard

Example of plagiarism in assignments

- For any concept questions in the assignments, such as “what is classification?. You can learn by searching answers from Internet, but you cannot simply copy the original texts online in your assignments. You should use your own language/texts as answers based on your understandings.

Syllabus and Blackboard

Policies of plagiarism

- 1st Time, you will get a zero score and warning, and a report to the department
- 2nd Time, you will get a zero score, and a fail (E) in this class, and an Academic Honesty Violation Report (AHVR) will be filed. You may be expelled from IIT

Note: only the final project could be a team project. Other assignments are individual home work

Examples of Unreasonable Requests

Reasonable requests can be accepted in some situations. Unreasonable requests will be ignored.

- Can you give me a second chance on the exam or the assignments?
- Can you give me extra practice or assignments so that I can improve my grade?
- Can you accept my super-late submissions (more than 1 week), because I have a medical issue this semester?

Important Notes

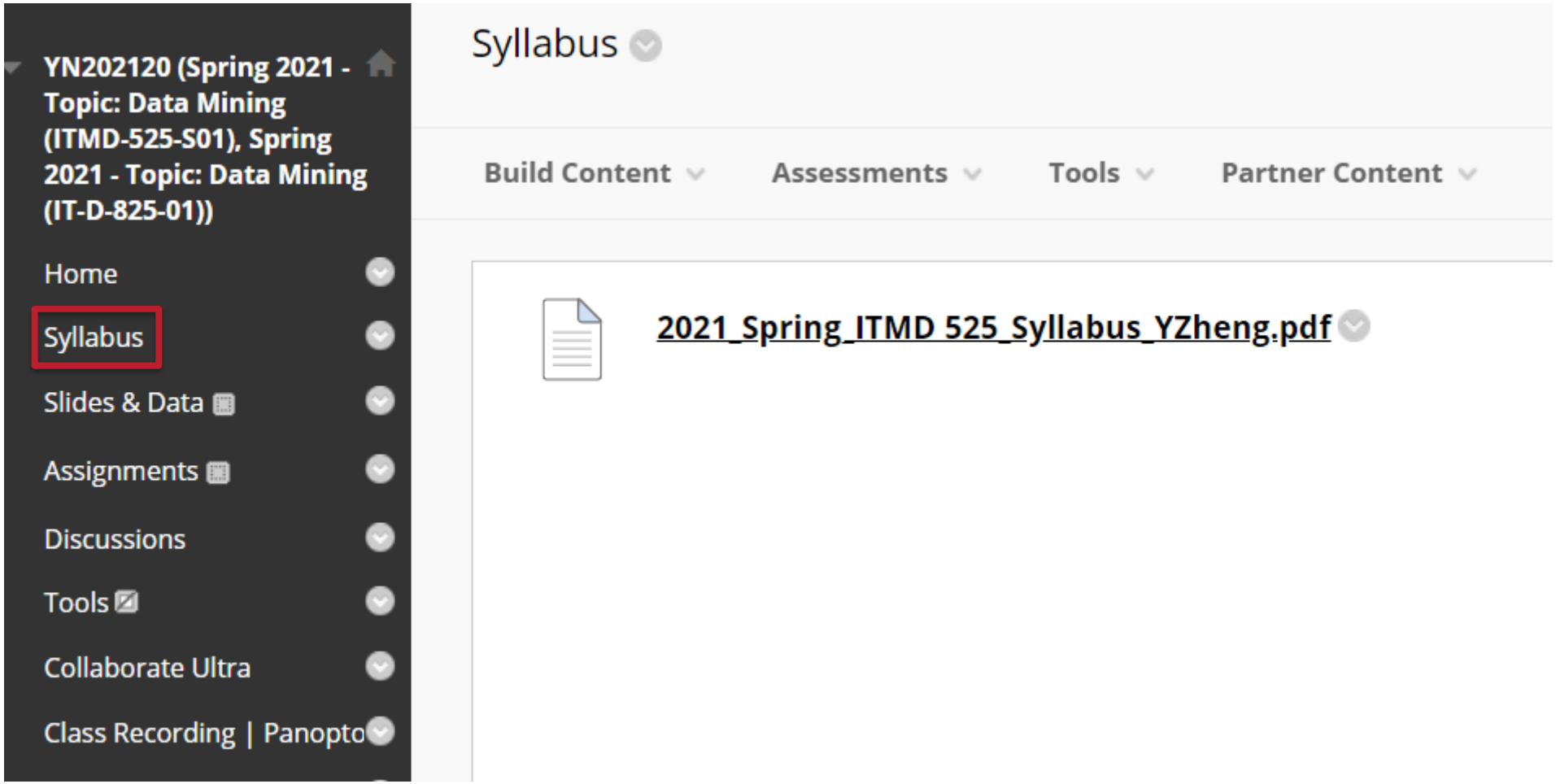
Syllabus and Blackboard

- It is your responsibility to read syllabus online and be familiar with the Blackboard system

Medical Issues and Disabilities

- You must introduce your medical situation as early as possible. You cannot use it as an excuse at the end of the semester
- Your medical issues or disabilities must be verified by the center for disability resources (CDR); telephone 312.567.5744 or disabilities@iit.edu

Syllabus


YN202120 (Spring 2021 - Topic: Data Mining (ITMD-525-S01), Spring 2021 - Topic: Data Mining (IT-D-825-01))
Home
Syllabus
Slides & Data
Assignments
Discussions
Tools
Collaborate Ultra
Class Recording | Panopto
Syllabus
Build Content
Assessments
Tools
Partner Content
2021_Spring_ITMD 525_Syllabus_YZheng.pdf

Practical Tools

In addition to the concepts and techniques, we will have practical experience in programming data science

- Python 3.0 + Data Science Package
  - Install Anaconda https://www.anaconda.com/
  - Coding and Running by Using Jupyter Notebook

Intro

- Who am I
- Data Management and Data Science
- Topics in This Class
- Syllabus, Blackboard system and Policy
- Data & Data Types

Getting to Know Data

- Types of the Data
  - Qualitative (Categorical/Nominal)
  - Quantitative (Numerical)

Getting to Know Data

- **Types of the Data**
  - **Qualitative (Categorical/Nominal)**
    - Nominal = Values are strings
    - Specials
      - **Binary Variable**: only two values in a variable
        e.g., gender (M or F)
      - **Ordinal Variable**: values have a meaningful rank
        e.g., letter grade (A, B, C, F)
        degree (BS, MS, PhD)
        size (S, M, L, XL, XXL)

Getting to Know Data

- **Types of the Data**
  - **Quantitative (Numerical)**
    - **Discrete**, we need to count objects to get values
      Example: number of students in the class
      They are usually integers
    - **Continuous**, we need to measure to get values
      Example: the length of the table
      They usually have decimals (not always)

Getting to Know Data

- If you observe that a column of numbers, it is not guaranteed that this variable is a numerical variable. You need to be careful about the identification of data types in a data set
- These numbers may be encoded for some reason, for example

ID	Nationality
1	2
2	2
3	3
4	1

ID	Nationality
1	India
2	India
3	Spain
4	China

Next

- Aug 25
  - Class is cancelled
  - Let's move to AWS Summit
    - Free, https://aws.amazon.com/events/summits/chicago/
    - Register and go to the event by yourself
    - Free gifts: T-shirts, pens, notebooks, etc.
    - Free talks: data science, AI, IoT, Security, etc.
    - Job opportunities: many companies rather than Amazon only

Next

- AWS Summit: Check-in

Event **Check-in**

Pre-event registration will be open on Wednesday, August 24, from 10:00 AM - 5:30 PM at McCormick Place – South for early badge pickup. Registration will be open on Thursday, August 25, from 8:00 AM - 6:00 PM. To expedite the pickup of your badge, follow the tips below:

- Have your registration QR code ready to scan (this can be found in your registration confirmation email from no-reply@awsevents.com)
- Have your government-issued photo ID in hand
- Have your record of COVID-19 vaccination or negative third-party COVID test result in hand