WELCOME TO IST707 AND WEEK 1

Jeremy Bolton

Thank you to
Dr. Ami Gates

COURSE NAVIGATION

1) The Wall

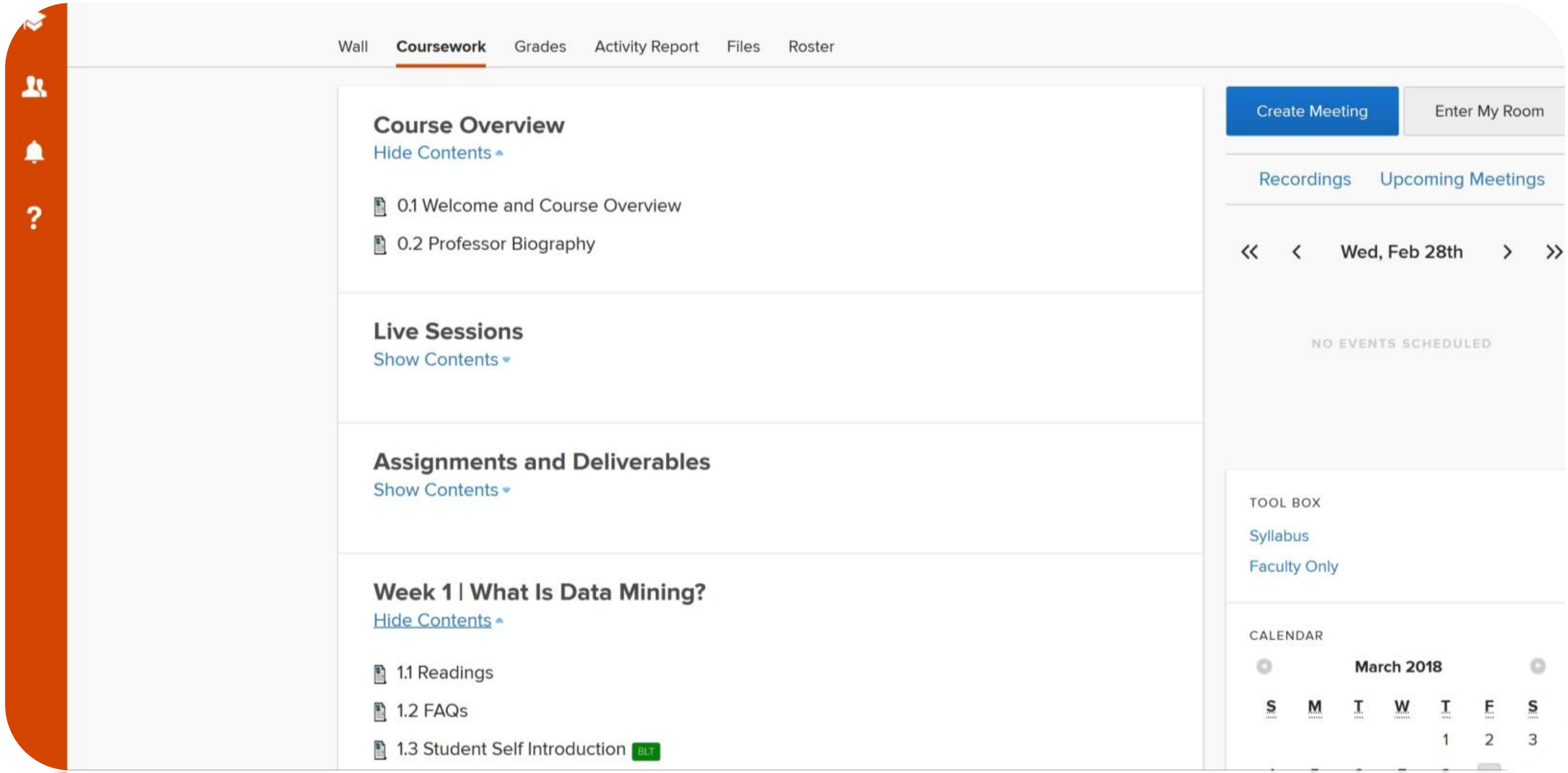
2) Assignments and Deliverables

3) Week 1 | What is Data Mining?

4) Grades

5) Files

6) Live Seminar Recordings



Wall | **Coursework** | Grades | Activity Report | Files | Roster

Create Meeting | Enter My Room

Recordings | Upcoming Meetings

« ‹ Wed, Feb 28th › »

NO EVENTS SCHEDULED

Course Overview

Hide Contents ▴

- 0.1 Welcome and Course Overview
- 0.2 Professor Biography

Live Sessions

Show Contents ▾

Assignments and Deliverables

Show Contents ▾

Week 1 | What Is Data Mining?

Hide Contents ▴

- 1.1 Readings
- 1.2 FAQs
- 1.3 Student Self Introduction BLT

TOOL BOX

Syllabus

Faculty Only

CALENDAR

March 2018

S	M	T	W	T	F	S
				1	2	3

THE WALL AND EMAIL

1) The Wall

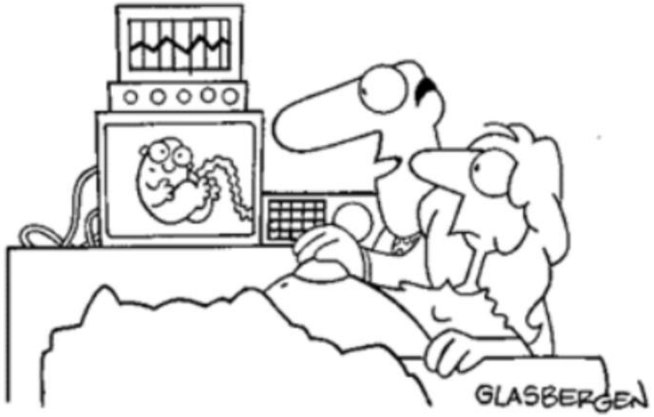
- is our Class Announcements location – I will post all important information in this area.
- The Wall also serves as a communication area for students.
- However – if you have a question for me – please email it to me – I do not want to miss it and I check email often.
- Be sure to check the the Wall once per day, at least 5 times per week.

2) Email

- Email is the best and fastest method to communicate with me.
- I will always respond to emails within 24 hours (unless you email me later on Friday).
- I would also appreciate if you respond to all of my emails (that require a response) within 24 hours.

3) Phone

- I do not use phone for communication because there is no ideal method for keeping records or logs. In addition, I cannot always answer my phone. I can (and do) always quickly answer email.
- Always use email to reach me.



"Your baby is developing very nicely. Would you like to send him an e-mail?"

REACHING YOUR PROF

1) I am online 6 days per week. I check email at least once – often more – per day.

2) I go **off-line from Friday around 12 noon through Sunday around 1 pm**. This works well as it means that I am only offline on Saturdays ☺

3) During the **Summer months,** I am often in a very different time-zone and/or hiking. I am still online each day, but it may be at times that are odd with respect to US time.

4) To avoid a crisis ☺ **never wait until the last day to do anything**.

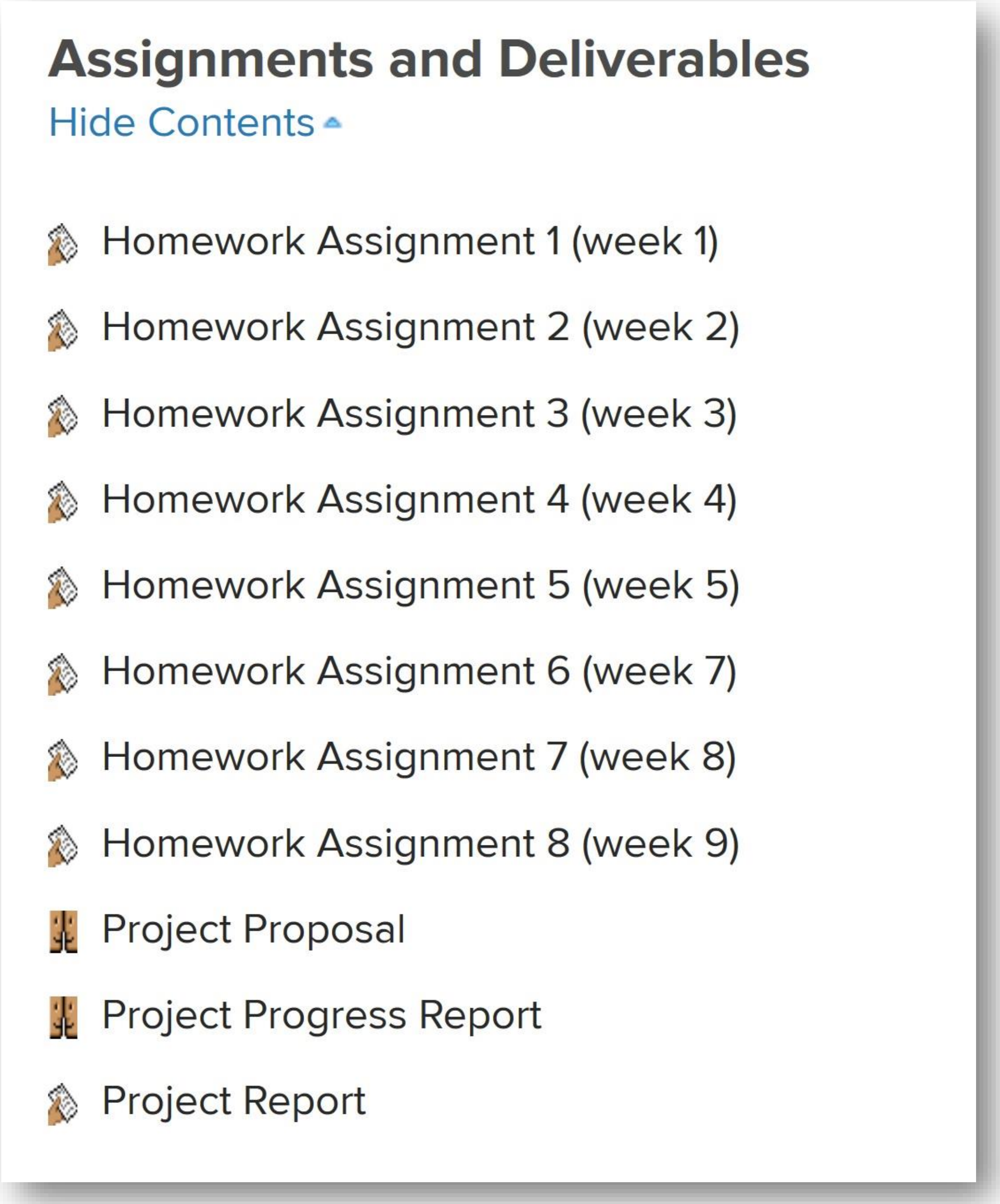
5) Always plan ahead, start work on day 1, and manage your time. This way, if you have a question, there is time to answer it.

6) **It is never necessary to email me for lateness.** I will post a Late Policy that we will all stick to. Largely, you cannot be late in this class – so always plan to have items does well in advance as life happens.

7) If I email you, please also get back with me within the same time frame.

ASSIGNMENTS AND DELIVERABLES

Notice that we do not have HW in Week 6 so the naming gets odd

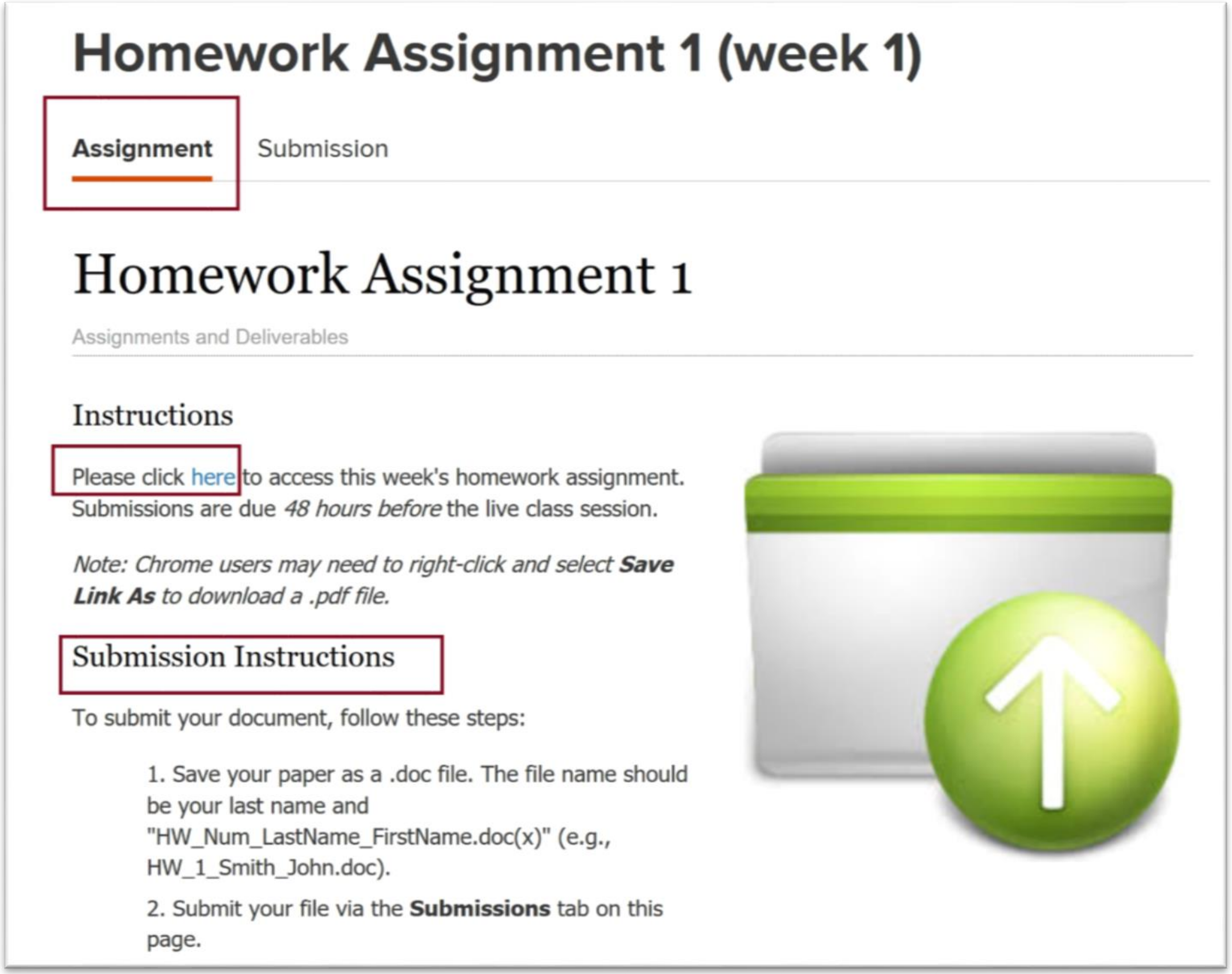


Assignments and Deliverables

Hide Contents ▲

- Homework Assignment 1 (week 1)
- Homework Assignment 2 (week 2)
- Homework Assignment 3 (week 3)
- Homework Assignment 4 (week 4)
- Homework Assignment 5 (week 5)
- Homework Assignment 6 (week 7)
- Homework Assignment 7 (week 8)
- Homework Assignment 8 (week 9)
- Project Proposal
- Project Progress Report
- Project Report

A LOOK AT THE WEEK 1 ASSIGNMENT



Homework Assignment 1 (week 1)

Homework Assignment 1

Assignments and Deliverables

Instructions

Please click here to access this week's homework assignment. Submissions are due *48 hours before* the live class session.

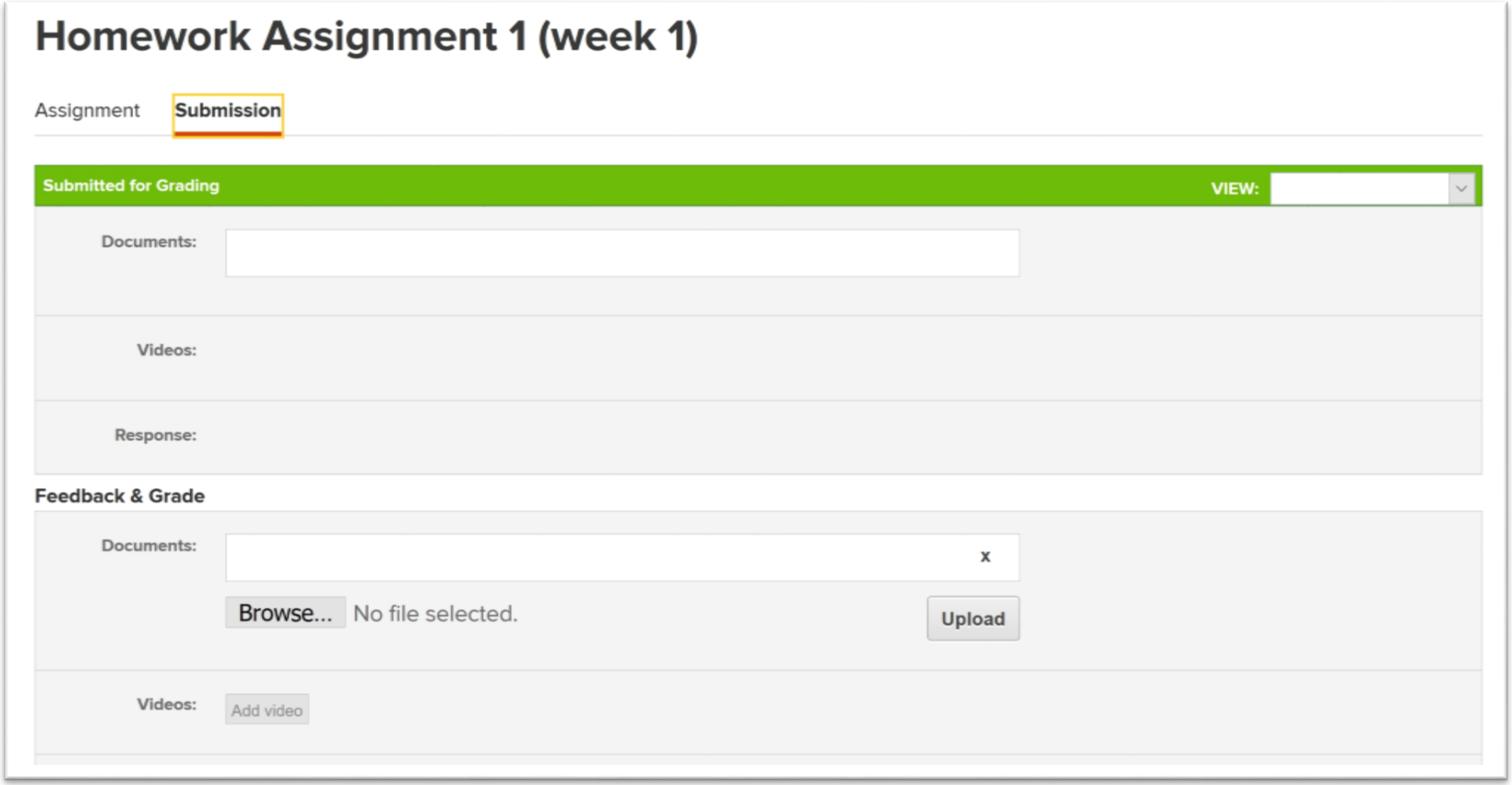
Note: Chrome users may need to right-click and select ***Save Link As*** *to download a .pdf file.*

Submission Instructions

To submit your document, follow these steps:

1. Save your paper as a .doc file. The file name should be your last name and "HW_Num_LastName_FirstName.doc(x)" (e.g., HW_1_Smith_John.doc).
2. Submit your file via the **Submissions** tab on this page.

A LOOK AT THE WEEK 1 ASSIGNMENT



Homework Assignment 1 (week 1)

Documents:

Videos:

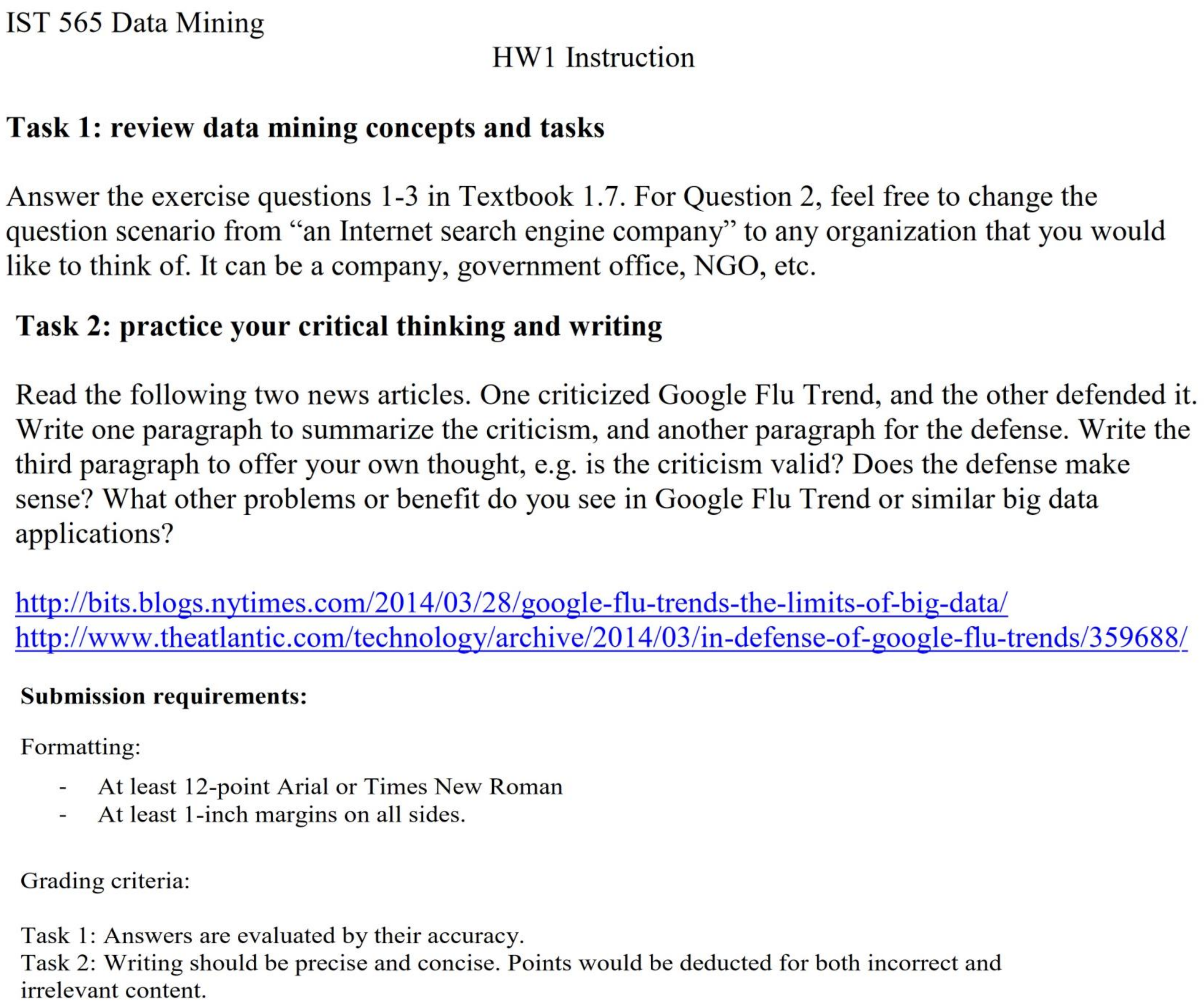
Response:

Feedback & Grade

Documents:

No file selected.

Videos:



IST 565 Data Mining

HW1 Instruction

Task 1: review data mining concepts and tasks

Answer the exercise questions 1-3 in Textbook 1.7. For Question 2, feel free to change the question scenario from “an Internet search engine company” to any organization that you would like to think of. It can be a company, government office, NGO, etc.

Task 2: practice your critical thinking and writing

Read the following two news articles. One criticized Google Flu Trend, and the other defended it. Write one paragraph to summarize the criticism, and another paragraph for the defense. Write the third paragraph to offer your own thought, e.g. is the criticism valid? Does the defense make sense? What other problems or benefit do you see in Google Flu Trend or similar big data applications?

http://bits.blogs.nytimes.com/2014/03/28/google-flu-trends-the-limits-of-big-data/
http://www.theatlantic.com/technology/archive/2014/03/in-defense-of-google-flu-trends/359688/

Submission requirements:

Formatting:

- At least 12-point Arial or Times New Roman
- At least 1-inch margins on all sides.

Grading criteria:

Task 1: Answers are evaluated by their accuracy.
Task 2: Writing should be precise and concise. Points would be deducted for both incorrect and irrelevant content.

AN IMPORTANT NOTE ABOUT WRITING SKILLS

1) Throughout your entire career, you will use your writing skills.

2) Your writing skills can display your excellence – or not.

3) A lack of writing skill can suggest a lack of ability or knowledge.

4) Writing is a key form of communication.

5) Technical writing is a form of writing that is concise, precise, and succinct.

6) Professional and academic papers, assignments, projects, etc. must be well written, written at the graduate level, and written as a technical work.

- Technical papers can and should explain ideas and points, make arguments, and offer creative anlaysis and thought.
- Technical papers are not "colorful" or overly filled with adjectives.
- Do not use words like, "I, we, you, us, our, your, etc. ".
- Technical writing is not in the first person and is generally in the third person.

7) **I *do* grade in part on writing skill**. It is recommended that you use a Writing Center and also proof-read your submissions out loud.

HEADINGS FOR ALL ASSIGNMENTS: PLEASE SEE THE WALL

1) Introduction

- This portion offers the reading an understanding of the topic, why it matters, its history, its background, its values, etc. This is a non-technical area. It does NOT discuss data or methods, but rather the area of interest and the goals.

2) Analysis

- This area should have a subheading: Data Preparation and Processing
  - The Data Prep area should discuss the data, the variables, their types and structure, how you cleaned it, why you cleaned it, transformations, binning, normalization, and any other preprocessing. Use visual EDA!
- The Analysis section focusses on the method(s) used to analyze the now cleaned and prepared data.
- Be very detailed so that a person could use only your paper to repeat exactly what you did and find the results that you found.
- Describe models used, offer background and basis for the method(s), etc.

3) Results

- For each model/analysis, create a technical set of results. Include limitations, what was found, what was not found, etc.

4) Conclusions – this is **non-technical**. All technical results and findings go into the Results area. In this area, discuss what was discovered in a way that anyone can understand. What did you find, what is its value, why it is important or interesting, etc.

A NOTE ABOUT COMMUNICATION

1) Effective communication via presentation and writing is critical.

2) In life, we are often judged and measured through these avenues.

3) When writing technical, academic, and/or professional papers (assignments and projects):

- Avoid speaking in the first person. Avoid “I”, “you”, “me”, etc.

- Include important R code and visualizations throughout the paper. This does not mean that you should show ever line of your code, but rather you can include critical lines that show models and methods.

- Include R results in your Assignments that support your models, methods, analysis, and results. Do this smartly.

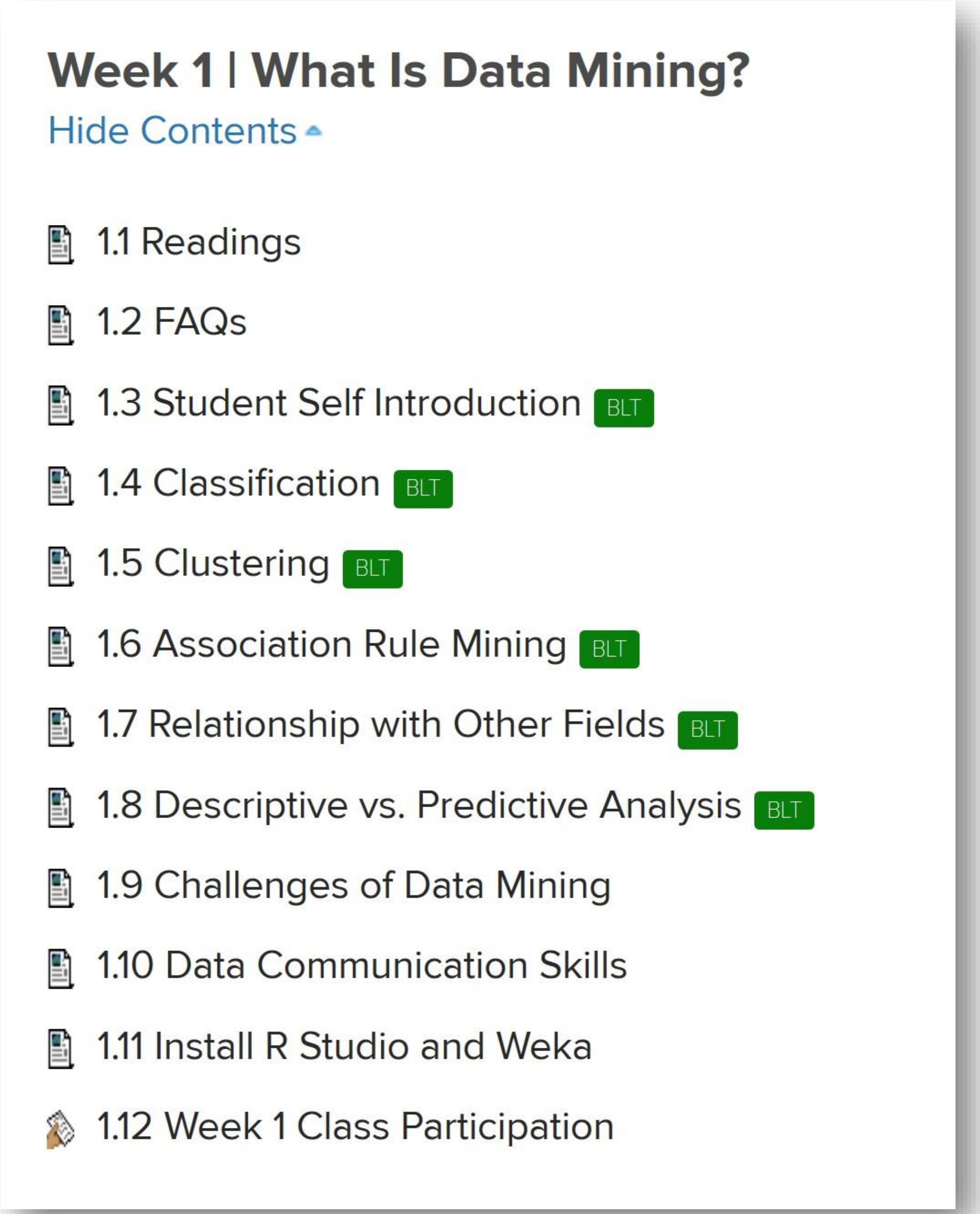
4) Proof-read all submission out loud first. Be sure your assignment papers have a clear flow and are easy to follow and understand.

A LOOK AT ALL WEEK 1 REQUIREMENTS

Note: For Week 1, there are 6 Exercises and one Live Session. Together, these 7 items will determine your 1.12 Week 1 Class Participation Grade.

Note2: Item 1.11 (Install R and Weka) should be started ASAP.

It is required to have all Weekly items completed 3 days before the Live Session – for us – that's no later than Sat



Week 1 | What Is Data Mining?

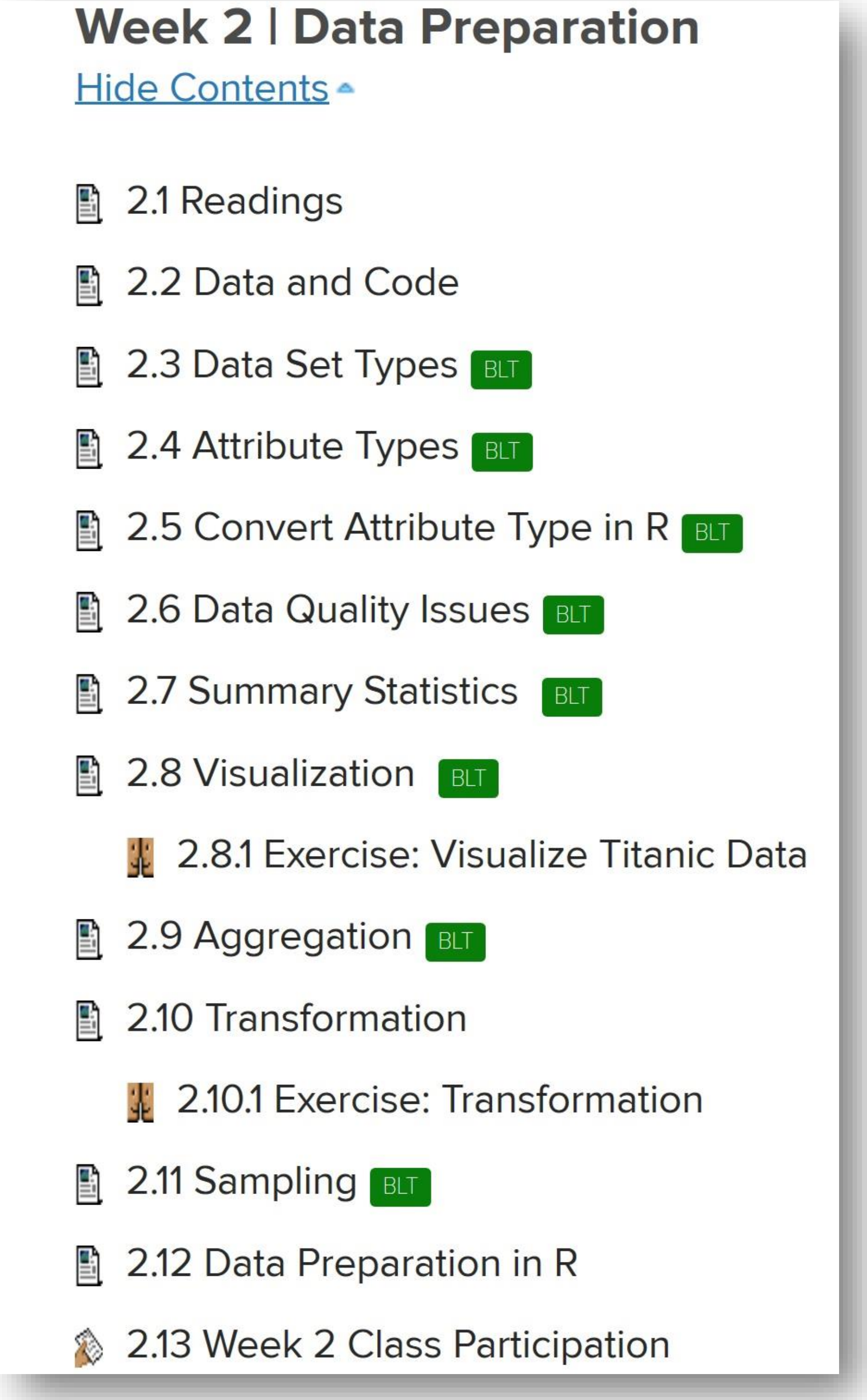
Hide Contents ▴

- 1.1 Readings
- 1.2 FAQs
- 1.3 Student Self Introduction BLT
- 1.4 Classification BLT
- 1.5 Clustering BLT
- 1.6 Association Rule Mining BLT
- 1.7 Relationship with Other Fields BLT
- 1.8 Descriptive vs. Predictive Analysis BLT
- 1.9 Challenges of Data Mining
- 1.10 Data Communication Skills
- 1.11 Install R Studio and Weka
- 1.12 Week 1 Class Participation

WEEK 2: REQUIRED BY FRIDAY AT 12 NOON *BEFORE* THE NEXT LIVE SESSION

I have posted due dates on the Wall and have emailed them out.

Lateness can get you into trouble in this class. Please manage your time and always plan ahead.



Week 2 | Data Preparation

Hide Contents

- 2.1 Readings
- 2.2 Data and Code
- 2.3 Data Set Types BLT
- 2.4 Attribute Types BLT
- 2.5 Convert Attribute Type in R BLT
- 2.6 Data Quality Issues BLT
- 2.7 Summary Statistics BLT
- 2.8 Visualization BLT
  - 2.8.1 Exercise: Visualize Titanic Data
- 2.9 Aggregation BLT
- 2.10 Transformation
  - 2.10.1 Exercise: Transformation
- 2.11 Sampling BLT
- 2.12 Data Preparation in R
- 2.13 Week 2 Class Participation

WEEKS 1 AND 2 READINGS

Required Reading: Week 1 (this should be completed before our Session)

Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining*. London and New York: Pearson. Chapter 1.

Required Reading | Week 2 | Data Preparation (Completed by the Friday 12 noon ET before out second Live Session)

Tan, P.-N., Steinbach, M., & Kumar, V. (2005). Introduction to data mining. London and New York: Pearson. **Chapters 2–3.**

Note: For Chapter 2, you may skip advanced content in 2.3.3-2.3.5 and 2.4.

For Chapter 3, you can skip 3.4.

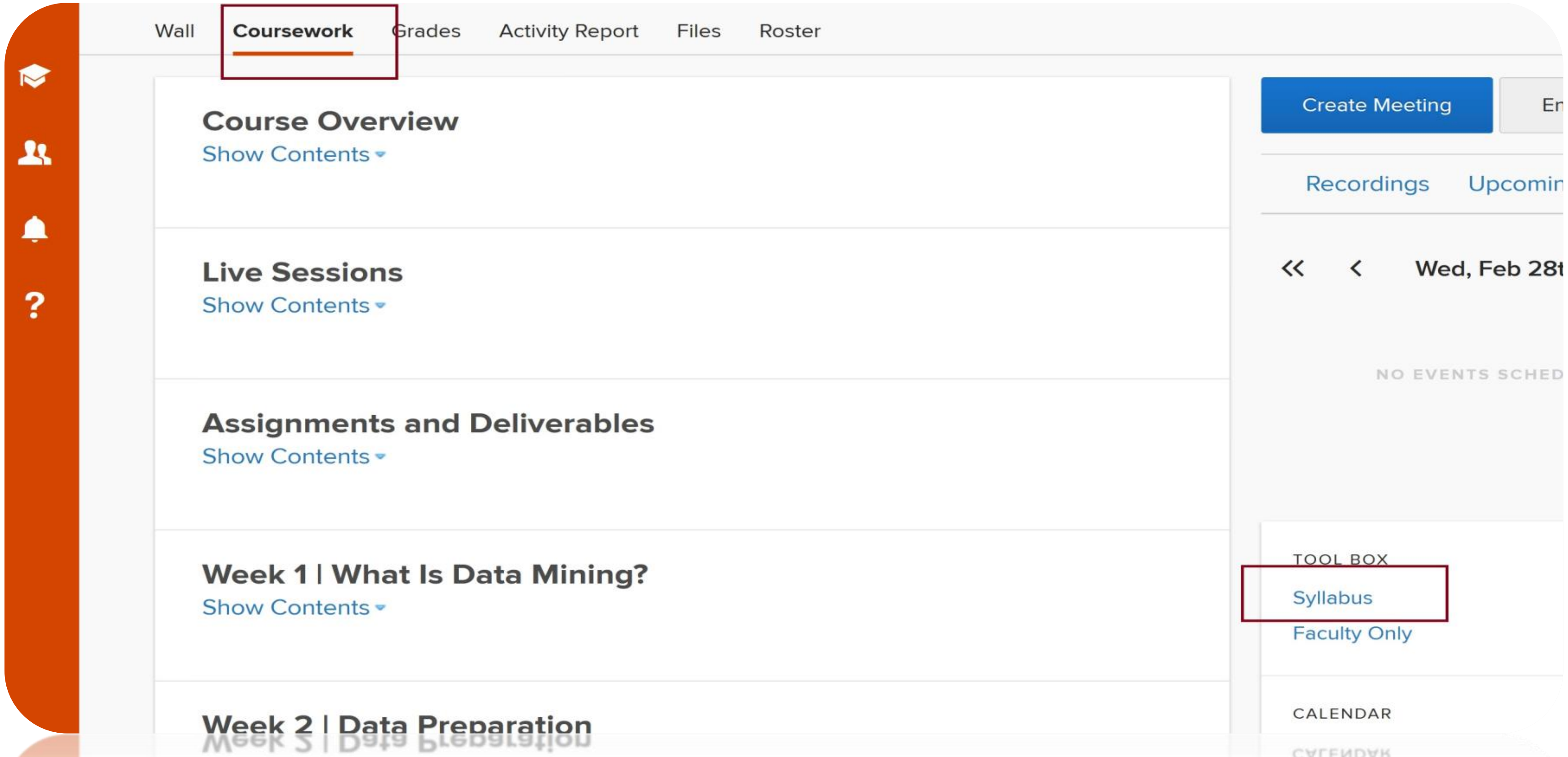
This is a great book!

Tan, Steinbach, &Kumar is often thought of as one of the definitive Data Mining Books.

Do not skip the readings.

Note that you will need to get 2005 version as some assignments reference book pages.

WHERE IS THE SYLLABUS: BUT USE MY NOTES ON THE WALL – THEY ARE ALWAYS THE MOST UP TO DATE ☺



CODING – IT TAKES A LOT OF TIME AND PATIENCE

1) Learning to program in R is critical and essential.

2) Most, if not all data science, data analytics, and information science positions expect that applicants can program in R and Python (as well as others).

3) Programming can (and will!!) be frustrating, challenging, maddening, and time consuming ☺

4) Finding and **figuring our your own bugs and errors is 70% of the “learning to program”** process. It is important!

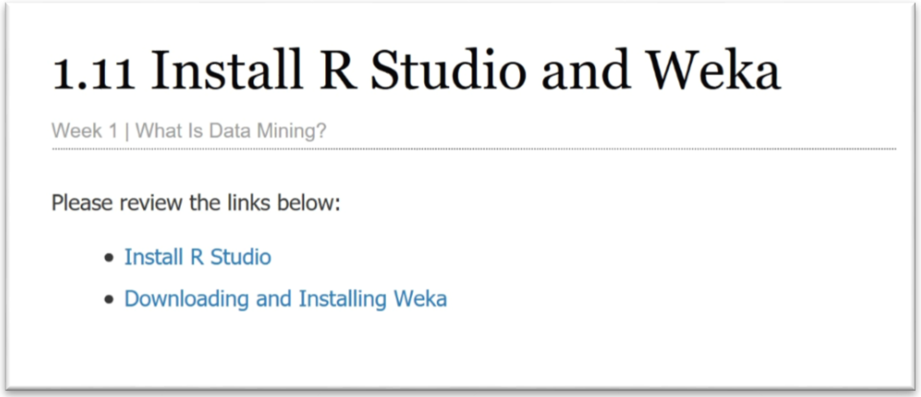
5) As such, **I will not debug or error-check your code** and I will not locate your bugs or errors for you. If I did, I would be doing you a great disservice.

6) I will talk about programming in R during our live meetings.

7) However, always start with something that works (like printing your name). Then add one or two lines of code at a time – save – test – add – save – test – add. If you run into a bug - *back-up* and find it.

8) Finding bugs and errors can take hours. Plan and manage your time wisely.

R AND WEKA



1.11 Install R Studio and Weka

Week 1 | What Is Data Mining?

Please review the links below:

- Install R Studio
- Downloading and Installing Weka

1) Make a point to learn R in this class.

2) It is well worth your time to go above and beyond and to spend the time learning R.

3) I also recommend learning Python3 if you can find spare time.

4) Placing R and Python3 on a Resume is *excellent* - especially for data science.

5) WEKA is more of an application. We will look at Weka as well, but I recommend doing as much as you can in R.

LINKS:

Weka: https://www.cs.waikato.ac.nz/ml/weka/

R:
https://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf

https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf

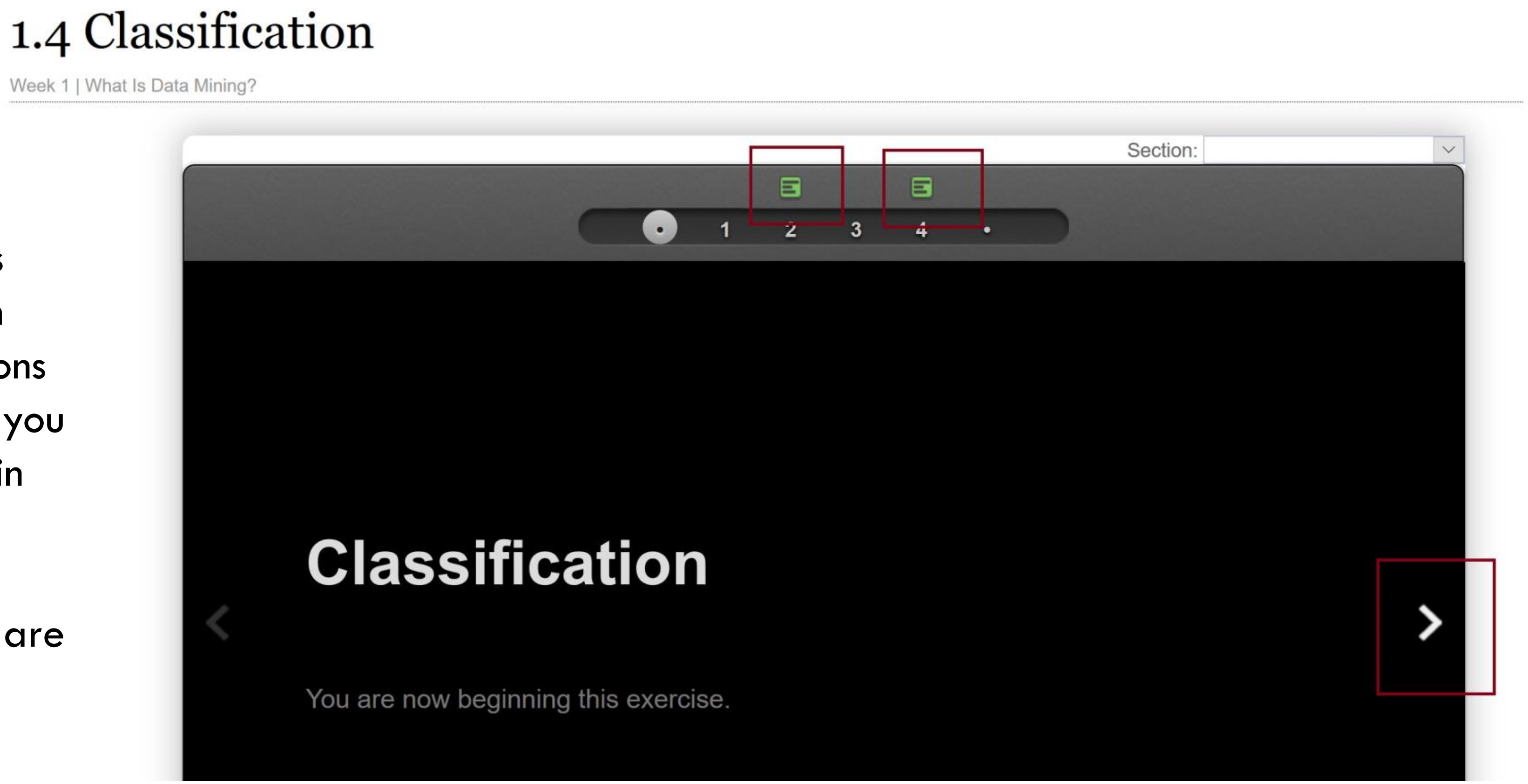
http://www.ievbras.ru/ecostat/Kiril/R/Biblio/R_eng/R%20dummies.pdf

Don’t forget about YouTube and Google ☺

WHAT ARE BLTS (BIDIRECTIONAL LEARNING TOOLS)

BLTs are where you will share your views and answer questions. As you go through the slide presentation, you will see portions (with the green box on top) that require you to answer a question and/or to interact in some way.

In this class, all BLTs and the Live Session are part of the Participation grade.



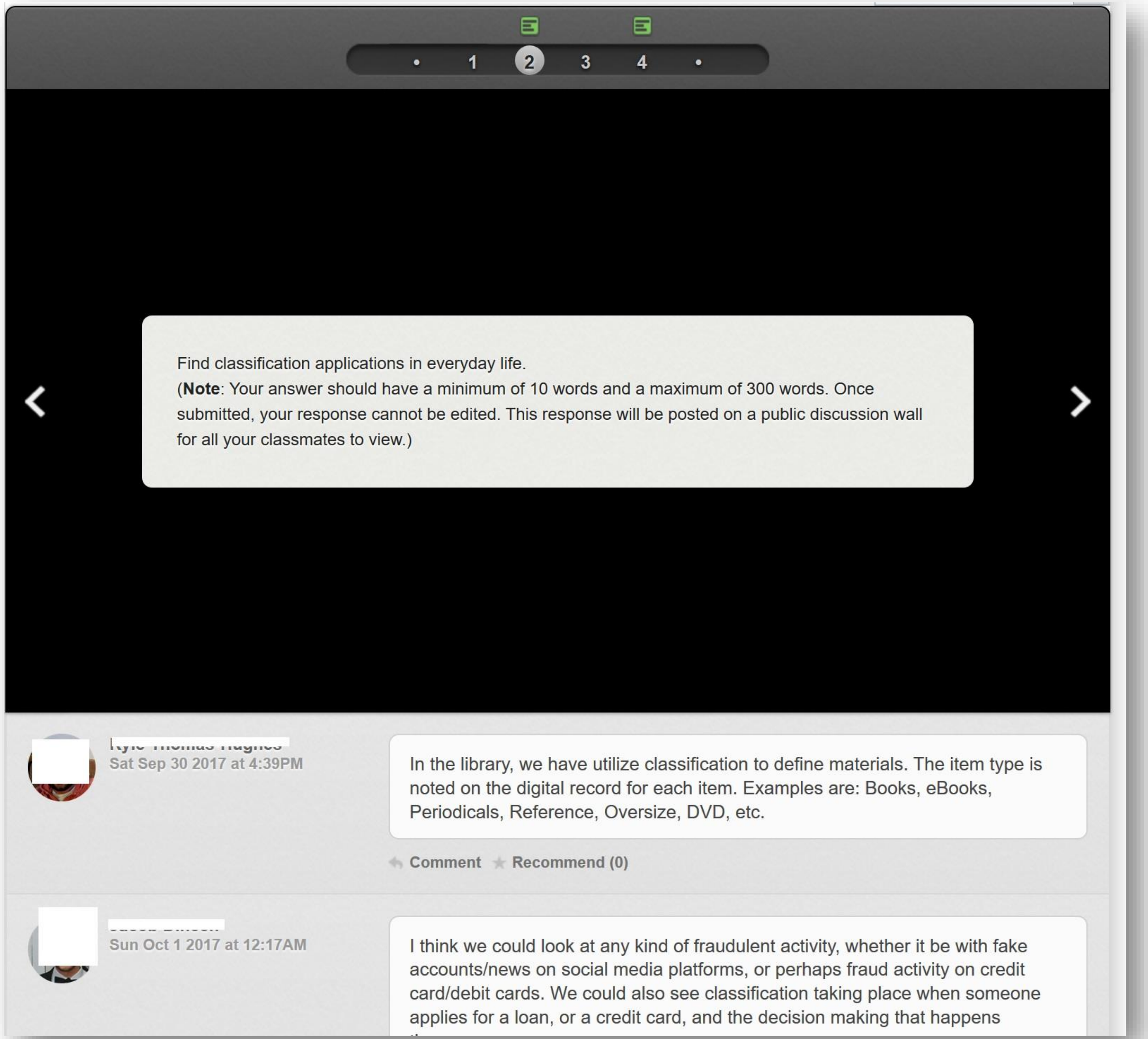
1.4 Classification

Week 1 | What Is Data Mining?

Classification

You are now beginning this exercise.

RESPONDING TO BLTS



Find classification applications in everyday life.
(**Note**: Your answer should have a minimum of 10 words and a maximum of 300 words. Once submitted, your response cannot be edited. This response will be posted on a public discussion wall for all your classmates to view.)

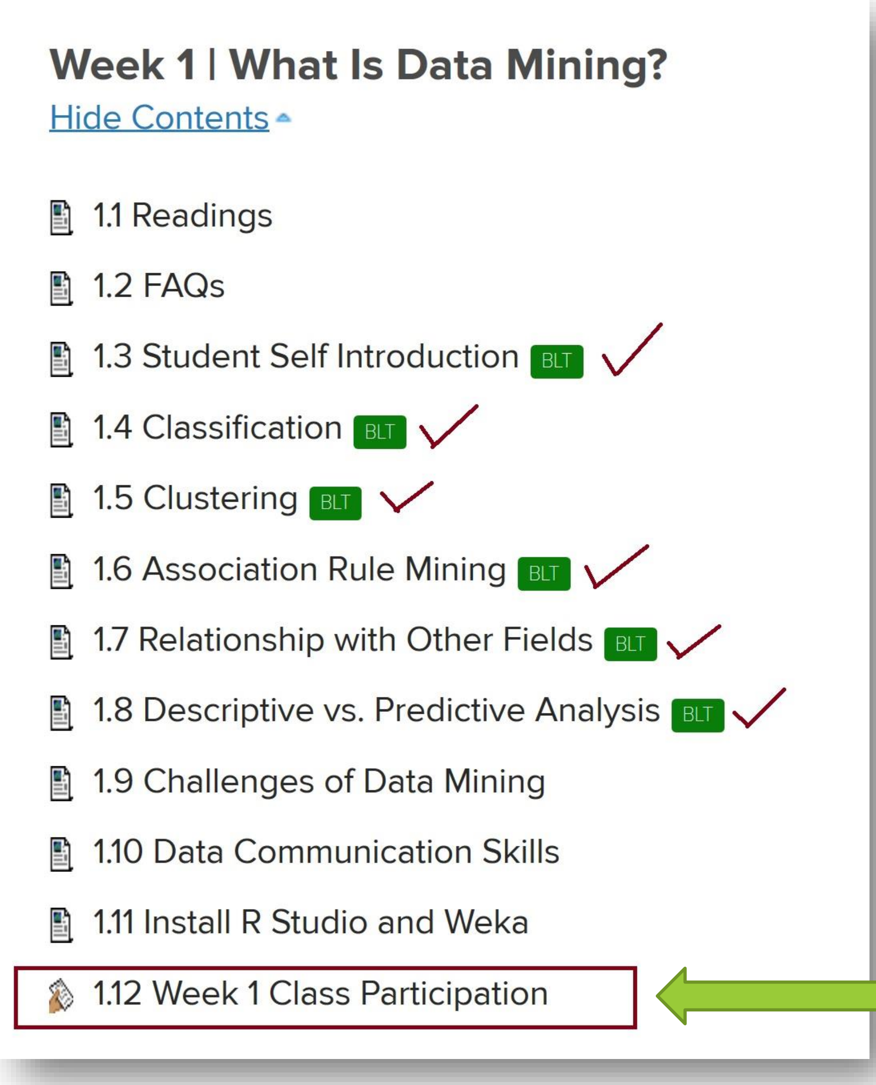
[redacted]
Sat Sep 30 2017 at 4:39PM

In the library, we have utilize classification to define materials. The item type is noted on the digital record for each item. Examples are: Books, eBooks, Periodicals, Reference, Oversize, DVD, etc.

[redacted]
Sun Oct 1 2017 at 12:17AM

I think we could look at any kind of fraudulent activity, whether it be with fake accounts/news on social media platforms, or perhaps fraud activity on credit card/debit cards. We could also see classification taking place when someone applies for a loan, or a credit card, and the decision making that happens

PARTICIPATION GRADE = EXERCISES/BLT + DISCUSSIONS + LIVE SESSION



Week 1 | What Is Data Mining?

Hide Contents ▲

- 1.1 Readings
- 1.2 FAQs
- 1.3 Student Self Introduction BLT ✓
- 1.4 Classification BLT ✓
- 1.5 Clustering BLT ✓
- 1.6 Association Rule Mining BLT ✓
- 1.7 Relationship with Other Fields BLT ✓
- 1.8 Descriptive vs. Predictive Analysis BLT ✓
- 1.9 Challenges of Data Mining
- 1.10 Data Communication Skills
- 1.11 Install R Studio and Weka
- 1.12 Week 1 Class Participation

1) All BLTs

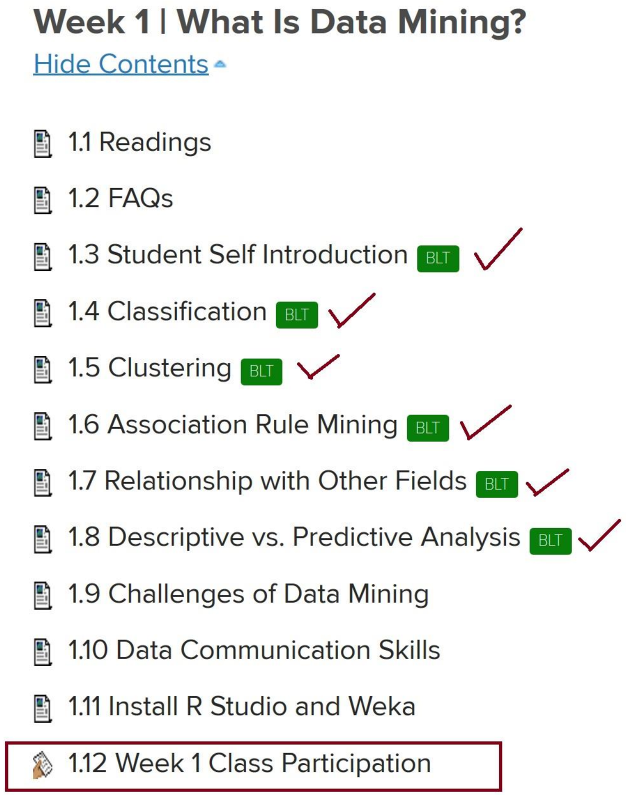
2) Attendance and participation in the Live Sessions.

To the left, all GREEN items are Exercises (BLTs).

The grade for participation is based on attendance, completion, and participation, but not on correctness.

Nothing to do here ☺

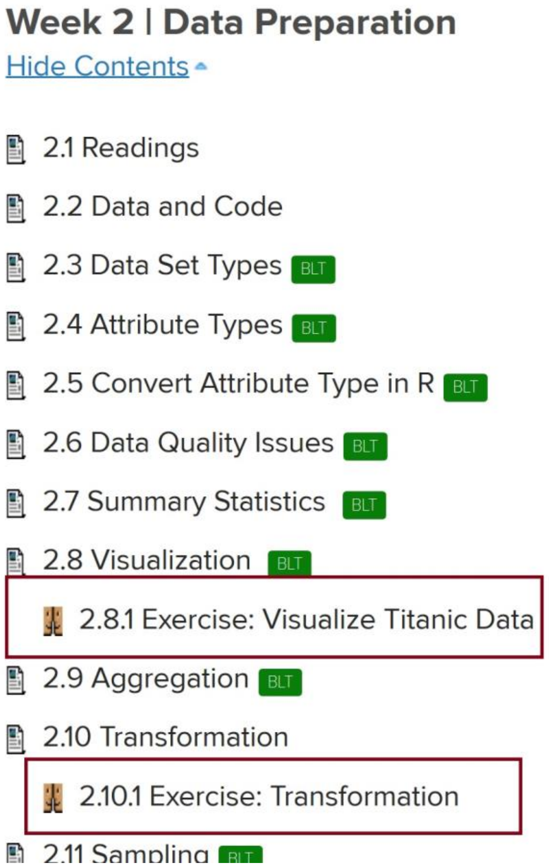
EXERCISES/BLT, DISCUSSIONS, FORUMS, ETC ALL PART OF THE PARTICIPATION GRADE…



Week 1 | What Is Data Mining?

Hide Contents

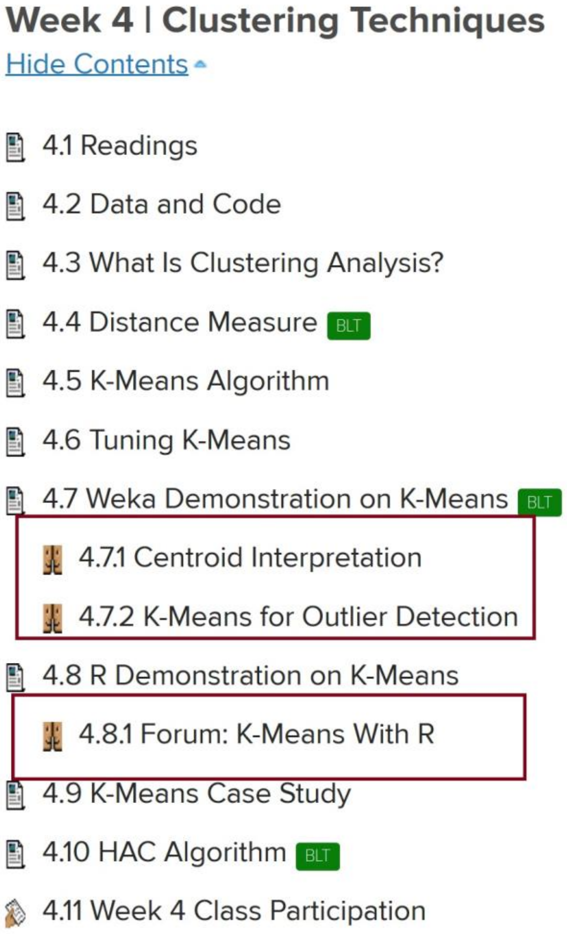
- 1.1 Readings
- 1.2 FAQs
- 1.3 Student Self Introduction BLT ✓
- 1.4 Classification BLT ✓
- 1.5 Clustering BLT ✓
- 1.6 Association Rule Mining BLT ✓
- 1.7 Relationship with Other Fields BLT ✓
- 1.8 Descriptive vs. Predictive Analysis BLT ✓
- 1.9 Challenges of Data Mining
- 1.10 Data Communication Skills
- 1.11 Install R Studio and Weka
- 1.12 Week 1 Class Participation



Week 2 | Data Preparation

Hide Contents

- 2.1 Readings
- 2.2 Data and Code
- 2.3 Data Set Types BLT
- 2.4 Attribute Types BLT
- 2.5 Convert Attribute Type in R BLT
- 2.6 Data Quality Issues BLT
- 2.7 Summary Statistics BLT
- 2.8 Visualization BLT
  - 2.8.1 Exercise: Visualize Titanic Data
- 2.9 Aggregation BLT
- 2.10 Transformation
  - 2.10.1 Exercise: Transformation
- 2.11 Sampling BLT



Week 4 | Clustering Techniques

Hide Contents

- 4.1 Readings
- 4.2 Data and Code
- 4.3 What Is Clustering Analysis?
- 4.4 Distance Measure BLT
- 4.5 K-Means Algorithm
- 4.6 Tuning K-Means
- 4.7 Weka Demonstration on K-Means BLT
  - 4.7.1 Centroid Interpretation
  - 4.7.2 K-Means for Outlier Detection
- 4.8 R Demonstration on K-Means
  - 4.8.1 Forum: K-Means With R
- 4.9 K-Means Case Study
- 4.10 HAC Algorithm BLT
- 4.11 Week 4 Class Participation

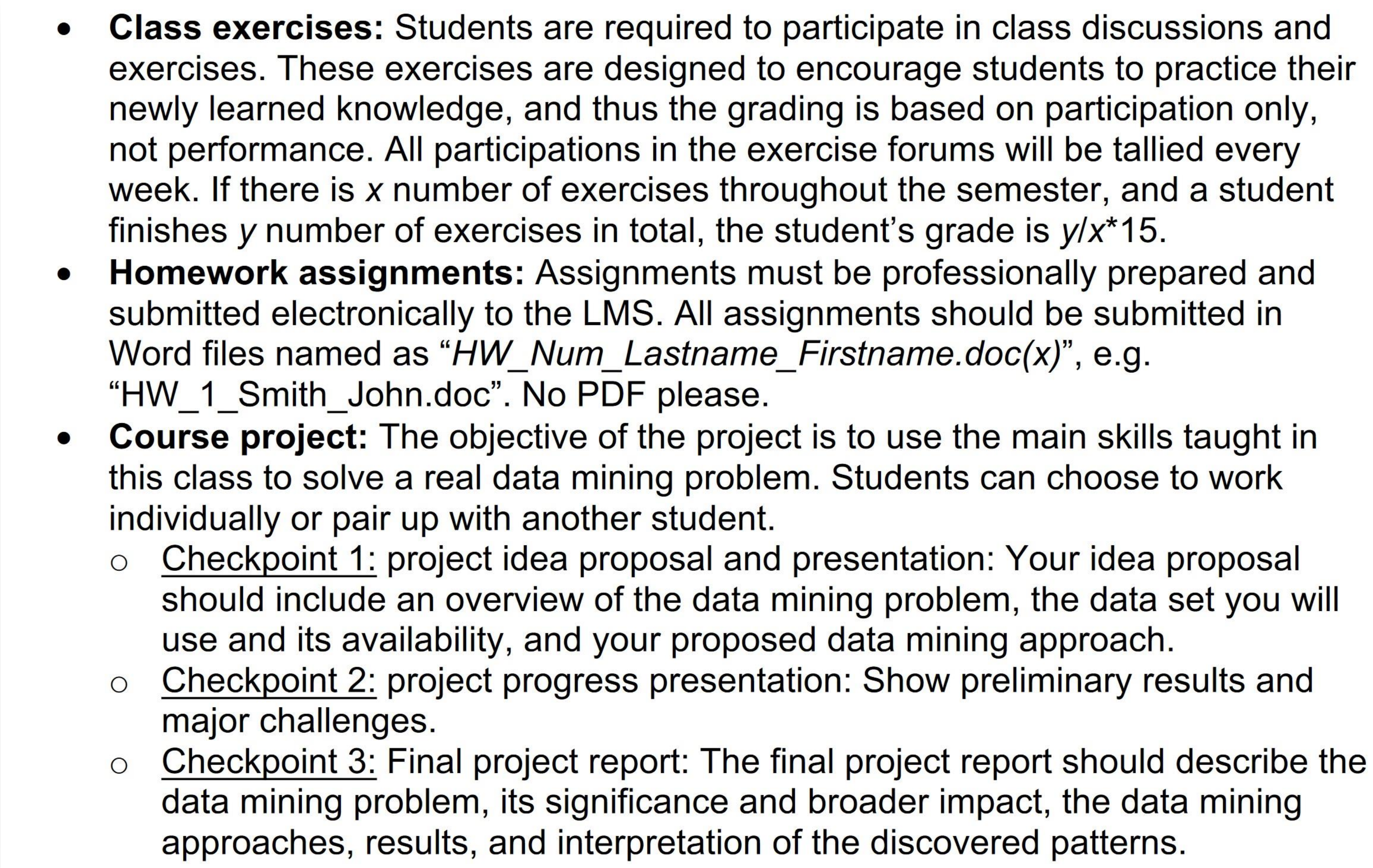
HOW IS YOUR CLASS GRADE DETERMINED?

1) Class Exercises: 15%

All BLT and Live Sessions and Discussion/Forums, etc. are required and part of this grade.

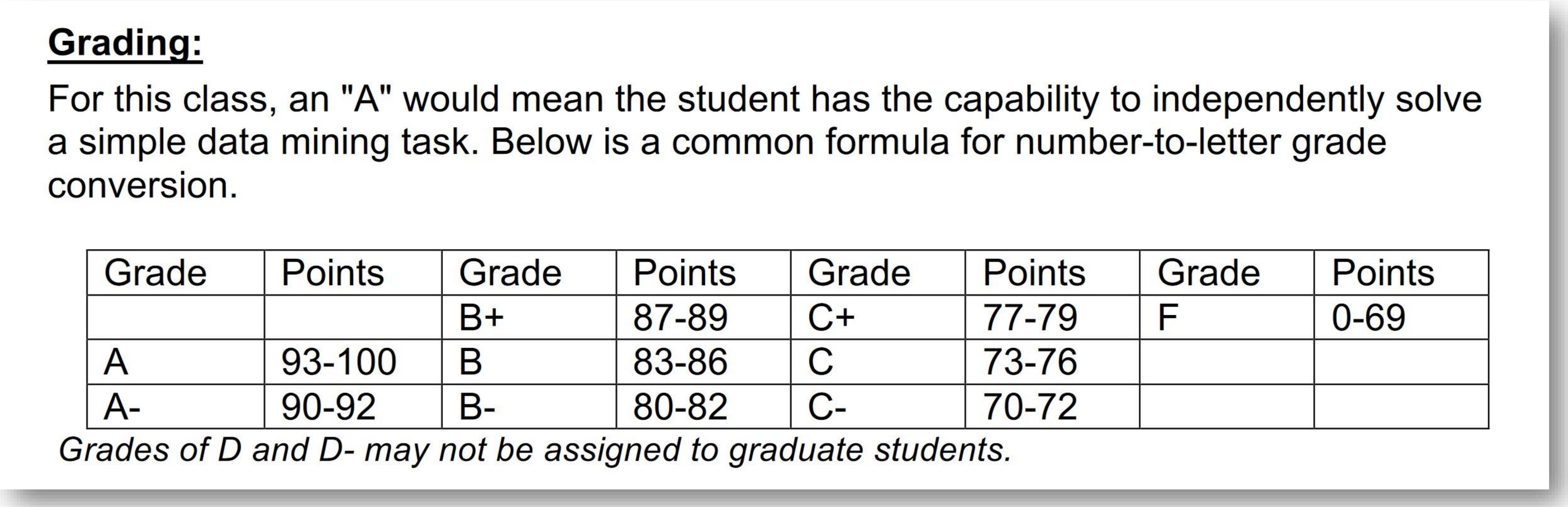
2) Homework Assignments 60%

3) Project: 25%



- **Class exercises:** Students are required to participate in class discussions and exercises. These exercises are designed to encourage students to practice their newly learned knowledge, and thus the grading is based on participation only, not performance. All participations in the exercise forums will be tallied every week. If there is *x* number of exercises throughout the semester, and a student finishes *y* number of exercises in total, the student's grade is y/x*15.
- **Homework assignments:** Assignments must be professionally prepared and submitted electronically to the LMS. All assignments should be submitted in Word files named as "*HW_Num_Lastname_Firstname.doc(x)*", e.g. "HW_1_Smith_John.doc". No PDF please.
- **Course project:** The objective of the project is to use the main skills taught in this class to solve a real data mining problem. Students can choose to work individually or pair up with another student.
  - Checkpoint 1: project idea proposal and presentation: Your idea proposal should include an overview of the data mining problem, the data set you will use and its availability, and your proposed data mining approach.
  - Checkpoint 2: project progress presentation: Show preliminary results and major challenges.
  - Checkpoint 3: Final project report: The final project report should describe the data mining problem, its significance and broader impact, the data mining approaches, results, and interpretation of the discovered patterns.

GRADING SCALE



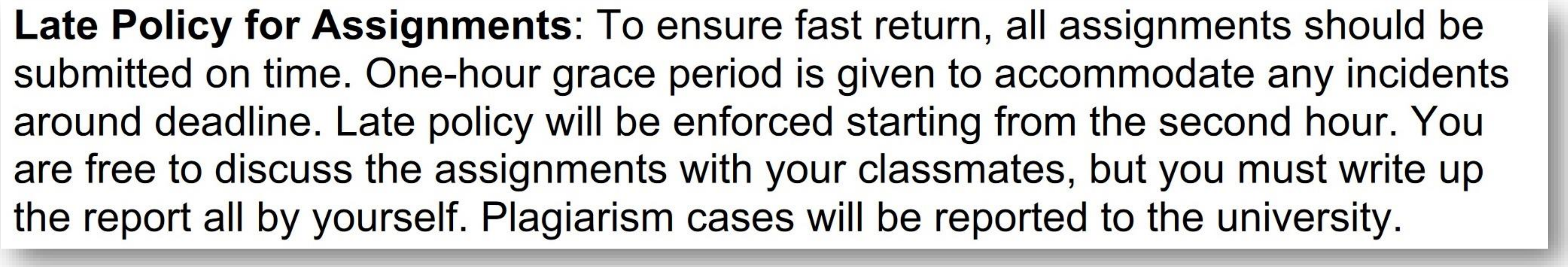
Grading:

For this class, an "A" would mean the student has the capability to independently solve a simple data mining task. Below is a common formula for number-to-letter grade conversion.

Grade	Points	Grade	Points	Grade	Points	Grade	Points
		B+	87-89	C+	77-79	F	0-69
A	93-100	B	83-86	C	73-76		
A-	90-92	B-	80-82	C-	70-72		

Grades of D and D- may not be assigned to graduate students.

SYRACUSE LATENESS AND LATE POLICY



Late Policy for Assignments: To ensure fast return, all assignments should be submitted on time. One-hour grace period is given to accommodate any incidents around deadline. Late policy will be enforced starting from the second hour. You are free to discuss the assignments with your classmates, but you must write up the report all by yourself. Plagiarism cases will be reported to the university.

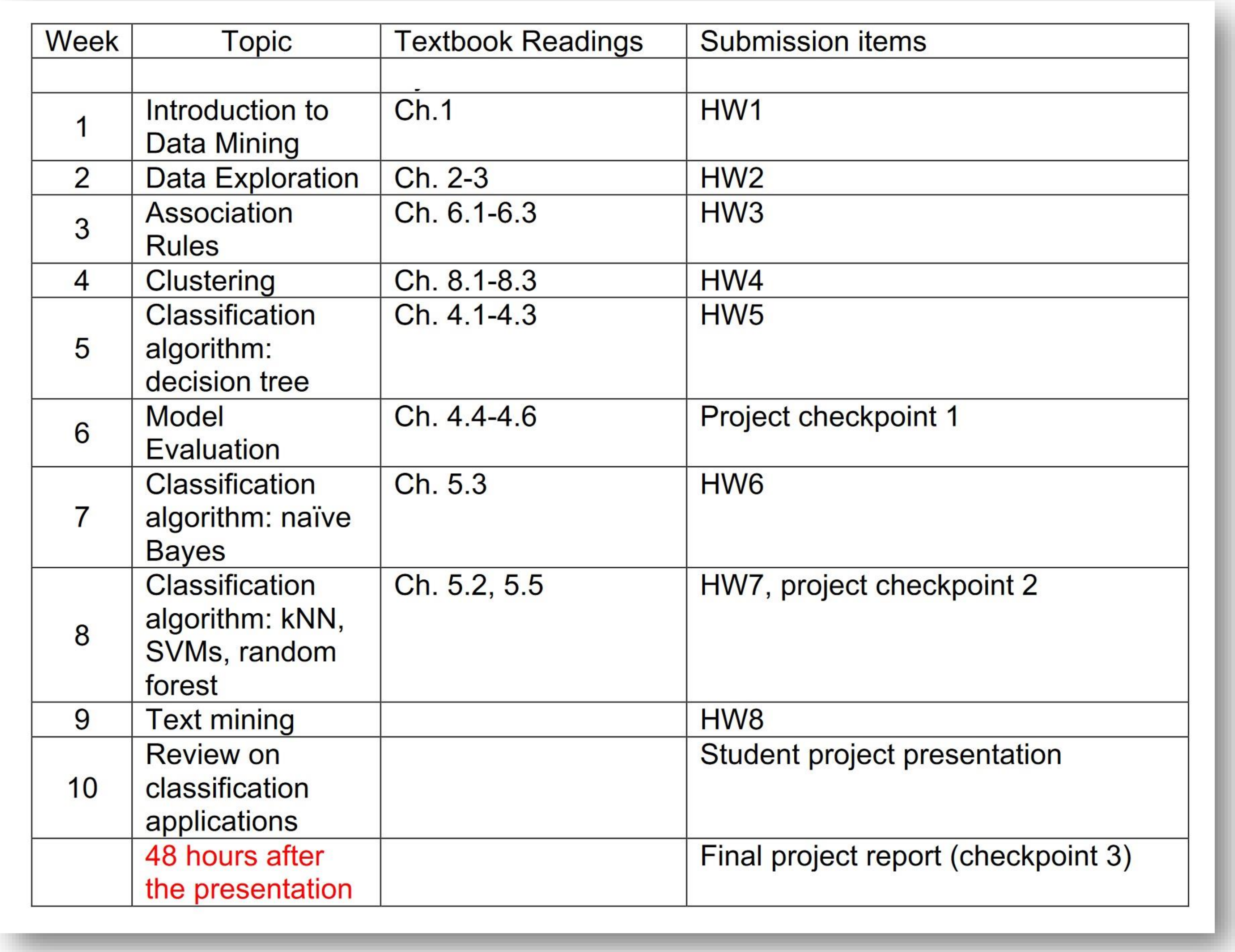
Our class use a **flipped design**.

1) This means that **<u>all requirements must be completed *before* each Live Session</u>**.
2) The Live Sessions are not lectures. They are discussions and involved student presentation and group work.

MY LATE POLICY:

1) Please see the Wall ☺

COURSE SCHEDULE BY WEEK



Week	Topic	Textbook Readings	Submission items
1	Introduction to Data Mining	Ch.1	HW1
2	Data Exploration	Ch. 2-3	HW2
3	Association Rules	Ch. 6.1-6.3	HW3
4	Clustering	Ch. 8.1-8.3	HW4
5	Classification algorithm: decision tree	Ch. 4.1-4.3	HW5
6	Model Evaluation	Ch. 4.4-4.6	Project checkpoint 1
7	Classification algorithm: naïve Bayes	Ch. 5.3	HW6
8	Classification algorithm: kNN, SVMs, random forest	Ch. 5.2, 5.5	HW7, project checkpoint 2
9	Text mining		HW8
10	Review on classification applications		Student project presentation
	48 hours after the presentation		Final project report (checkpoint 3)