# CSC 466:Knowledge Discovery from Data (KDD)
# Fall 2019
# Course Syllabus

September 10, 2019

**Instructor:** Alexander Dekhtyar
**email:** dekhtyar@calpoly.edu
**office:** 14-210

| What | When | | Where |
|---|---|---|---|
| Lecture | MWF | 11:10 am – 12:00pm | 20-139 |
| Lab | MWF | 12:10 – 1:00pm | 14-256 |
| **Final Exam** | December 13 (Friday) | 10:10 - 1:00pm | 20-139 |

*Note: the class will not have a written final exam, but we will most likely use the exam time for team project presentations.*

**Office Hours**

| | When | Where |
|---|---|---|
| Monday | 10:10pm - 11:00am | 14-210 |
| Tuesday | 9:10 - 11:00am | 14-20 |
| Wednesday | 10:10am - 11:00pm | 14-210 |

Additional appointments can be scheduled by emailing the instructor at *dekhtyar@calpoly.edu.*

## Description

This class is an overview of the field of knowledge discovery from data (KDD, also often referred to as "Data Mining" or "Machine Learning") and related technologies. The course is intended for senior students in Computer Science, Software Engineering and Computer Engineering majors, as well as for students completing the Cross-Disciplinary Studies Minor in Data Science. The course gives a broad overview of data mining (association rules mining, classification, clustering), information filtering and recommender systems, information retrieval and web search, and web mining.

## Learning Objectives

After taking the course the students are expected to be able to

1. **recognize** different types of KDD procedures and **identify** their uses;
2. **implement** algorithms/methods/techniques for KDD tasks to **solve** KDD problems;
3. **interpret** and **analyze** the *results* of KDD processes;
4. **recognize** and **evaluate** *societal impact* of KDD technology, **make informed choices** about use of KDD technology.

## Texbook

- Bing Liu. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, Springer, 2st ed. 2011. ISBN: 978-3642194597.

This book is **mandatory**. It contains almost all material studied in the course (and then some). While we rely on instructor's lecture notes as much as we rely on the content of the book, the book is extremely useful.

In addition, some other books may be of use. If you want an alternative take on most of the material covered in class, I recommmend (as an optional book)

- Mohammed J. Zaki, Wagner Meira Jr. *Data Mining and Analysis: Fundamental Concepts and Algorithms*, Cambridge University Press, 2014, ISBN: 978-0-521-76633-3.

This book was the textbook in Professor Khosmood's version of CSC 466.

## Topics

| No. | Topic | Duration | Liu | Zaki,Meira |
|---|---|---|---|---|
| 1. | Association Rules | 1 | Chapter 2 | Chapters 8, 9 |
| 2. | Supervised Learning (Classification) | 1.5 | Chapter 3 | Chapters 18, 19, 22 |
| 3. | Unsupervised Learning (Clustering) | 1.5 | Chapter 4 | Chapters 13, 14, 15, 17 |
| 4. | Collaborative Filtering | 1 | Chapter 11, 12 | |
| 5. | Information Retrieval | 2 | Chapter 6 | |
| 6. | Link Analysis | 2 | Chapter 7 | |
| 7. | Advanced Topics | 1 | | |

Please note that the order in which these topics are covered may be different than the order in which they are presented above.

# Grading

| | |
|---|---|
| **Labs and Homeworks** | 60% |
| **Lab Exam** | 10% |
| **Projects** | 30% |

# Course Policies

## Prerequisites

The official prerequistite for this course is CSC 349 (Algorithms). This prerequisite is enforced **strictly**. CSC 466 can be viewed as an advanced algorithms course for a certain important category of algorithms. Therefore, it is important that the craft of algorithm design is not a mystery to anyone in the class.

## Exams

The course will have **no written exams**. In their stead you will be offered three things:

1. A take-home group project. The project assignment will come out after Week 5-6 of the class, once the finish covering the core topics of the course (classification and clustering). The group assignment will ask you to apply the methods studied in the course to exploratory analysis of some real-life data.

2. A take-home individual assignment. This assignment will come out in the last 2-3 weeks of the class, and will most likely be a creative writing exercise.

3. An in-class lab exam. The lab exam will most likely take place during the lab periods of Week 10 of the course. You will be asked to write some code to implement some straightforward KDD tasks. We will be using Jupyter Notebooks for the exam, and you will be given some sample problems at least a week or two ahead of the time.

Most of lab assignments and other coursework in this course is done in either small teams or in pairs. The individual assignment and the lab exam were introduced into the class to allow me to properly assess individual abilities of students, that otherwise may have been obscured due to teamwork.

The reserved final examination time will be used for group project presentations.

## Labs, Homeworks

Hands-on KDD-related activities are the core part of the course. Some activities will be set up as lab exercises, some other activities may be offered as purely take-home assignments (this will be determined by the pace of the course).

Each lab/homework assignment will involve some data analysis task, that may involve using existing software, software provided by the instructor as well as (and mostly) the software developed by you. The course concentrates on **basic algorithms for performing standard KDD tasks**: the labs/assignments give you an opportunity to cement the knoweldge of the algorithms covered in class.

Most of the lab assignments are pair programming assignmnets, although some exceptions may be made either in favor of small teams, or in favor of individual assignments. This will be announced ahead of each lab assignment.

**Note:** Machine learning, data mining and other KDD algorithms that we are going to study in the class are widely available, both as open source code and, in some cases, as methods/functions in popular KDD libraries. **One of the goals of this course is to have you implement these algorithms from scratch!** The assignments will specify when you can, and when you cannot third-party code/libraries to achieve the goals of the assignment.

**Note 2:** At the same time, in a lot of the assignments, the main deliverable will be not the code you write, but rather, the insight you obtain by running your code on the data provided to you. Please be aware of that, as this shift in what is the main deliverable, is perhaps one of the key unique features of CSC 466.

### Late Submissions

*Late* ***lab*** *and* ***assignment*** *submissions* are strongly discouraged. The course will run on a tight schedule, and not submitting on time will lead to time carved out of the next assignment. A penalty of 10 - 30% will be assessed for any submissions that are late by less than 24 hours. No credit will be given for any later submissions. You are encouraged to submit on time even if your submission it is not perfect. You can then resubmit a fixed version late, subject to the abovementioned rules. When more than one submission is present, we will independently grade two submissions: (i) the latest on-time submission and (ii) the latest late submission for which non-zero credit can be assessed. Your grade for the project will be the **maximum** of the two grades.

## Communication

The class will have an official mailing list. The email address for the mailing list is *csc-466-01-2198@calpoly.edu*. All students enrolled in the class are automatically subscribed to the mailing list (using the email addresss that the CS department has on file).

I encourage questions during classtime and questions via email. My answers to email questions may be broadcast to the entire class via the mailing list, if the answer may be relevant to everyone (e.g. a correction in a text of a handout, or a clarification of a homework problem), and may also appear on the web page. The questions can also be posted to the mailing list directly. The mailing list will also be used for all annoucements related to the course. It is your responsibility to read your class-related email. Failure to read email posted to the mailing list cannot be used as an excuse in the class.

## Web Page

Class web page can be found at

http://www.csc.calpoly.edu/∼dekhtyar/466-Fall2019

Through this page you will be able to access all class handouts including homeworks, lab assignments, project information, lab/project data and lecture notes.

Links to additional information, and notes and announcements will also be posted.

## Academic Integrity

### University Policies

Cal Poly's Academic Integrity policies are found at

http://www.academicprograms.calpoly.edu/academicpolicies/Cheating.htm

In particular, these policies define *cheating* as (684.1)

> "*. . . obtaining or attempting to obtain, or aiding another to obtain credit for work, or any improvement in evaluation of performance, by any dishonest or deceptive means. Cheating includes, but is not limited to: lying; copying from another's test or examination; discussion of answers or questions on an examination or test, unless such discussion is specifically authorized by the instructor; taking or receiving copies of an exam without the permission of the instructor; using or displaying notes, "cheat sheets," or other information devices inappropriate to the prescribed test conditions; allowing someone other than the officially enrolled student to represent same.*"

Plagiarism, per University policies is defined as (684.3)

> "... the act of using the ideas or work of another person or persons as if they were one's own without giving proper credit to the source. Such an act is not plagiarism if it is ascertained that the ideas were arrived through independent reasoning or logic or where the thought or idea is common knowledge. Acknowledgement of an original author or source must be made through appropriate references; i.e., quotation marks, footnotes, or commentary."

University policies state (684.2): "Cheating requires an "F" course grade and further attendance in the course is prohibited." (appeal process is also outlined, see the web site above for details.). Plagiarism, per university policies (684.4) can be treated as a form of cheating, although a level of discretion is given to the instructor, allowing the instuctor to determine the causes of plagiarism and effect other means of remedy. It is the obligation of the instructor to inform the student that a penalty is being assessed in such cases.

### Course Policies

All homeworks are to be completed by each student **individually**. Lab assignments are to be completed by the appropriate units (individual, pair, group), and no code/solution-sharing between units is permitted. Students are encouraged to discuss class content among themselves but NOT in a manner that constitutes plagiarism and cheating as defined above (e.g., you can solve together a problem from the textbook that had not been assigned in the homework, but you should solve assigned problems individually).