Intro to Data Science

Course Syllabus ~ Fall 2019

OVERVIEW

The purpose of this course is for students to learn how to engage in the scientific process using data-centric concepts and methods and to *think like a data scientist* by critically analyzing their own work and the work of others.

Learning Outcomes

It is my goal that after completing this course successfully, you will be able to:

1. Explore a data set to determine whether and how it might illuminate questions of interest.
2. Define and operationalize a research question such that a data analysis could produce meaningful knowledge.
3. Use best practices to carry out analyses in a documented, reproducible, and efficient fashion.
4. Present the results of a data analysis with appropriate visuals and written argument.
5. Identify weaknesses in a data analysis and assess their impact on the correctness and utility of the results.
6. Assess ethical implications of an analysis in terms of both classical human subject research ethics and contemporary concerns such as fairness and bias.
7. Understand the space of data science techniques and applications, and relate future learning to this framework.

COURSE LOGISTICS

Course Title	CS 533: Introduction to Data Science
Credits	3
Schedule	Tu/Th 9:00–10:15 in CCP 259
Course Website	Blackboard

INSTRUCTOR

Michael Ekstrand

Office	CCP 255
E-mail	michaelekstrand@boisestate.edu
Phone	(208) 426-5761
Office Hours	Tu/Th 10:30-11:30am or by appointment

RESOURCES AND READINGS

Textbook

Our primary textbook is:

> *Data Science Essentials in Python* by Dmitry Zinvoev (ISBN 9781680501841).

Online Readings

Throughout the semester, I will assign various readings from the Internet and research papers. These will be posted to Blackboard.

Supplemental Books

The following optional texts may be useful:

> *Think Like a Data Scientist* by Brian Godsey (Manning, ISBN 978-1633430273)

> *Python for Data Analysis* by Wes McKinney (O'Reilly, ISBN 978-1491957660)

Software

Throughout this class, we will be using Python with the PyData tools (Pandas, Numpy, Scipy, matplotlib, Seaborn, etc.). The easiest way to install the required software is to install Anaconda Python (https://www.anaconda.com/distribution/). Onyx already has Anaconda installed.

I will not provide support for debugging Python installations other than Anaconda.

The various Python libraries we use each have their own documentation:

- Python
- Pandas
- Seaborn

COURSE STRUCTURE

You are responsible for reading the textbook. We do not have time to discuss everything that you need to know in class; our time there is better spent diving deeper into the topics.

Course Components

Your final grade will be computed from these components as follows:

Category	*%*
Assignments	30
Labs	5
Project	30
Quizzes	20
Final	15

The standard 70/80/90 scale determines the minimum grade you will receive (that is, if you have 80 total course points, you will receive at least a B-).

Assignments and Labs

There will be 6 homework assignments and 4 smaller labs throughout the semester, practicing data science techniques in Python.

Each assignment is due at **midnight on Thursday** of the week in which it is due.

Unless otherwise stated in the assignment description, uou may work in groups of up to 2 on the assignments, but **not** the labs. If you work with a partner, only submit the assignment under **one group member**, and list your partner in your assignment submission.

Project

A crucial component of this class is a semester-long project working with an external non-profit or municipal partner on a data analysis or modeling project. I will be providing more details in the second week of class, and in the Project section of Blackboard. This project will be in groups of 3 or 4, and will involve coordinating with external partner representatives.

Quizzes and Final

There will be 3 midterm quizzes (each about ½ of a class period) and a final exam. Your lowest quiz score will be dropped.

COURSE POLICIES

Web Site and Announcements

I will make important class announcements in class and on Blackboard. You are responsible for these announcements (e.g. corrections to assignments).

Attendance

I strongly encourage you to attend all class sessions. If you need to be absent for some reason, such as conference travel or illness, please let me know as soon as you can.

Late Work

For the **lab assignments**, you have a budget of **4** late days to use throughout the semester, at your discretion. Each late day extends an assignment deadline by 24 hours with no penalty. When submitting an assignment using a late day, state with your submission the number of days you are using. Late days are indivisible; submitting an assignment 12 hours late requires an entire late day.

This policy is designed to accommodate most ordinary need for extensions or late submissions. Therefore, no other late work will be accepted. Exceptions to this policy will only be granted in extreme circumstances. Any requests for individual exceptions must be submitted by e-mail so that I have a record of the request and my response.

Project deliverables are due on stated dates, and no extensions will be granted.

Quizzes and the **final exam** will be at the published times.

Cheating and Academic Integrity

As both a scientist and a student, you are expected to do your own work, attribute sources, and respect the legal and moral rights of others with respect to their work; as a student, you are also required to abide by the Boise State University Student Code of Conduct. While I aim to allow you to make reasonable use of resources, cheating (including copying code, using unauthorized resources during tests, etc.) will not be tolerated. If you are found to be cheating, the penalty may range from an F on the assignment to an F on the course, and will also be reported to the university.

Conduct

I expect you to behave in a civil, respectful manner in all class interactions, both in official meetings such as lectures and out-of-classroom activities such as project group meetings and study sessions, and to contribute to a constructive learning environment.

The Recurse Center Social Rules are a good source of guidance on how to maintain a constructive and educational environment.

If you experience or witness harassment of any form, please let me know.

Disability Accommodations

If you need particular accommodations or support to be able to fully participate in this course, please talk with me as soon as possible. I may ask that you provide documentation from the Office of Disability Services, so if you have such documentation please bring it.

SCHEDULE

Following is a tentative schedule. I will likely adjust as we progress through the semester. **Bold** items are key dates for the project and exams.

Week	Date	Topic	Due
1	8/26	Intro & Asking Questions	L0
2	9/2	Understanding Data	L1
3	9/9	Modeling: From Data to Inference	A1
4	9/16	Probability and Information Theory	L2
		Quiz 1	
5	9/23	Sourcing, Cleaning, and Integrating Data	A2
		Initial meeting report due	
6	9/30	Visualization, Time, and Space	L3
7	10/7	Regression	A3
8	10/14	Classification and Evaluation	
9	10/21	**Quiz 2**	A4
10	10/28	Ethics, Bias, and Social Effects	
		Midterm status report due	
11	11/4	Text Processing and Latent Models	
12	11/11	Clustering and Graphs	A5
13	11/18	**Quiz 3**	
T	11/25	*Thanksgiving Break*	
14	12/2	Data Science in Production	A6
15	12/9	**Project Presentations**	
F	12/16	**Final Exam** (Wed. May 6 2:30-4:30 PM)	Reflection

Assignments and labs (A*n* or L*n*) are due **on Thursday at midnight** the week they are listed.