Oregon State University
Ecampus

Course Name: Data Science Tools and Programming
Course Number: CS 512
Credits: 4 Credits
Instructor name: Dr Michael Curry
Instructor email: michael.curry@oregonstate.edu
Teaching Assistant name and contact info: TBC

Course Description

Accessing and distributing data in the cloud; relational and non-relational databases; map reduction; cloud data processing; load balancing; types of data-stores used in the cloud. This course is only available to Department of Statistics graduate students working to MS and PhD degrees in Statistics

Incoming Expectations

This course is made for data science students who have completed CS 511 or have equivalent experience in python programming. You will be expected to use many third party programs and will need to be comfortable doing basic tasks using a command line. The emphasis of this course is on the data science tools and on programming. There should be many opportunities to use more advanced statistical methods, but they will not be covered in this course.

Communication

Please post all course-related questions in the Q&A Discussion Forum so that the whole class may benefit from our conversation. Please contact me privately for matters of a personal nature. I will reply to course-related questions within 24 hours. I will strive to return your assignments and grades for course activities to you within five days of the due date.

Time Expectations

On average this course combines approximately 150 hours of instruction, online activities and assignments for 4 credits. This roughly translates into 4-5 hours per week, with the majority of that time spent completing the activities and assignments. Assignments can be challenging, start early and if you have questions ask them on the class discussion page. Waiting until the last minute is not an effective strategy.

Technical Assistance

If you experience any errors or problems while in your online course, contact 24-7 Canvas Support through the Help link within Canvas. If you experience computer difficulties, need help downloading a browser or plug-in, or need assistance logging into a course, contact the IS Service Desk for assistance. You can call (541) 737-8787 or visit the IS Service Desk online.

Learning Resources

There is no course text, and instead you will be assigned to read documentation online. Its very important to read this material which has instructions on how to use the programming tools covered. Some of the material is very well documented such as the Conda documentation which is well established. However other toolsets are newer and less complete like the Google Cloud toolset and may need to be supplemented by students.

Note: each student will need a Google Cloud account and as a class we have credits that are usually sufficient to cover students computing costs for the term. However, if you use all your credits you will have to purchase additional ones.

Measurable Student Learning Outcomes

1. Use appropriate tools to access and store distributed data
2. Design programs to analyze distributed data
3. Evaluate various storage and access options for large data sets
4. Evaluate existing programs to find and remove bottlenecks in data processing
5. Understand differences between relational and non-relational databases
6. Use tools to clean and transform datasets for use in database systems
7. Create visualizations of large data sets

Evaluation of Student Performance

- Homework 70%
- Final Project 15%
- Midterm Project 15%

Letter Grade

Grade	**Percent Range**
A	93
A-	90
B+	87
B	83
B-	80
C+	77
C	73
C-	70
D+	67
D	63
D-	60
F	0

Course Content

Week	Topic	Readings	Learning Activities
1	Overview and Tools	Explorations: • Course Overview • The OSEMN Data Science Process • Python Review • Anaconda Overview • GCP Overview	Homework: • Environment Configuration • Python Review
2	SQL and Relational Databases	Explorations: • SQL Overview • SQL SELECT • Data Manipulation • SQL Relationships • Creating SQL Tables	Homework: • SQL Queries • Self-Signup Groups
3	Data Formats and Data Wrangling	Explorations: • Data Formats • A Data Wrangling Case Study	Homework: • Data Wrangling
4	Data Visualization	Explorations: • Graphical Excellence and Integrity • matplotlib	Homework: • Data Visualization Critiques • Data Visualization Workbook **Start Mid Term Project**
5	Project Week		**Complete Mid Term Project** **Reflections on Weeks 1 to 5 Learning**
6	Introduction to Big Data	Explorations: • Big Data Overview • Big Data Product Examples	Homework: • BigQuery and Bigish Data

Week	Topic	Readings	Learning Activities
		- Data Storage Walkthrough	
7	Non-Relational Databases and Spark	Explorations: - CAP Examples - Brewer's CAP Theorem - Hello Spark	Homework: - Spark Plane Distances - First Attempt
8	Spark Specifics	Explorations: - Web Interfaces - Spark Partitions - Spark IO	Homework: - Spark Plane Distances - Spark Activities
9	Spark Wrap Up	Explorations: - Spark Dataframes - Spark Rows - Views and Schemas	**Start Final Project**
10	Odds and Ends at the End	Explorations: - SparkR - Local Parallelization	**Reflections on Weeks 6 to 10 Learning**
11	Finals Week		**Complete Final Project**

Module Availability

Each module is generally available on beginning of Thursday, and two weeks ahead. So when the term starts, modules 1 and 2 are open, and then on Thursday of week 1, module 3 will open and so forth each week of the term.

Assignments

There will generally be an assignment every week. They are graded on correctness only. A program that fails to run will typically get no credit, even if all you had to do was add a : to make it run. If you want feedback on specific choices you made in your program you can post the code and the specific questions to the discussion board 48 hours after the assignment is due and the instructor and TAs can provide additional feedback there that you and other classmates can benefit from.

Assignment format

Unless otherwise specified all files should be be submitted as .py files if they are code files or as .pdf files if they are written documents. You should always submit all of the .py files you wrote for a project even if we will not be running them so we have a record of your program if we need to come back and look at it later. Files should not be zipped because that makes them hard to view in the Canvas grading tool. If we want a zipped file set we will ask for it specifically.

Final Project

This class will have a final project. This will require you to do calculations, visualization and analysis of a very large data set, ideally in the tens to hundreds of millions of records. More details will be released on this later in the course.

Incoming Expectations

This course is made for data science students who have completed CS 511 or have equivalent experience in python programming. You will be expected to use many third party programs and will need to be comfortable doing basic tasks using a command line. The emphasis of this course is on the data science tools and on programming. There should be many opportunities to use more advanced statistical methods, but they will not be covered in this course.

Tools

This course will primarily be about using Googles suite of cloud and big data tools. These tools are generally based on Python 2.7 however other tools will likely leverage Python 3.6. This isn't a fun place to be in but we can deal with it.

Ideally, if you can get Anaconda installed and running with two different environments, one running Python 2.7 and the other running Python 3.6 you will be ahead of the curve.

You may also find it useful to install the Google Cloud SDK but you should be aware this can be a little challenging.

Real World Stuff

The Google set of tools are professional tools used in industry. This is awesome because it means that employers will be excited you know how to use it. It also means that Google requires cash money to use it. As part of the class you will get $50 in Google cloud credit. This should be plenty to get through the course. But you need to be careful. If you leave a service running, get something stuck in an infinite loop or accidentally request to use a cluster of 1000 servers instead of 10 you might quickly burn through that budget.

The instructor will do what they can to help in that sort of situation, but if you do something wrong and it eats up all your credit you **may** need to pay for additional computation time if Google is unable to provide additional credit.

Getting Help With Tools

Speaking of challenges, when you run into trouble, please seek help. When it comes to software configuration issues like installing Anaconda or the Google SDK if you do the wrong thing to solve a problem it can make it harder to fix.

Try to follow the online documentation. If it does not work, don't improvise unless you are confident in your skills. It is better to come to the message board and ask for help. When you do please post:

- The steps you have taken
- The last thing you did that seemed to work correctly
- The full error message, this might be very long. If it is more than a page there are sites like Pastebin that are designed for posting and sharing code snippets and errors.
- Where else you have looked for solutions so I don't point you to something you already looked at.

Communications

There are several ways to communicate in an online class, they all have different purposes.

Message Board

We will use Piazza as a discussion platform. This is where you should post well prepared questions, both technical and theoretical about the content and assignments. You should follow similar steps to ***Getting Help With Tools*** when asking questions here.

You should not ask about any personal information on Piazza. For example it is not the place to ask about why you lost points on an assignment or to request an extension for an assignment.

Email

All emails to the instructor or the TA should have, as the first 7 characters in the email title [CS519] or [CS512]. This makes sure it goes to the right email box. If it does not have this tag at the beginning of an email it may get overlooked.

Instructor

You should email the instructor about general class concerns, extension requests or about issues you are unable to resolve with the TA. In general questions about the assignment or help with troubleshooting should go to Piazza where they can help other students. If you email the instructor these questions they might just ask you to post them to Piazza.

If you **did** post a question to Piazza and no one has answered it in 24 hours during weekdays then you can email the instructor to call attention to it and they will come answer it on Piazza. In general the instructor tries to answer questions before this but sometimes they are missed.

TA

You should email the TA grade related questions. If you do not understand why you got a deduction or think that they may have missed something while grading contact them and CC the instructor. The instructor will not look at or change grades until you have contacted the TA first. The TA will know better why they graded something the way they did than the instructor.

If, after emailing the TA, they are unable to resolve your issue then you should email the instructor, forwarding your exchange with the TA.

Office Hours

The TAs will hold regular office hours for general troubleshooting and questions. This is often a good place to get help working through error messages.

The instructor will meet online with students by appointment. The office hours are held online using Zoom.

Expected Response Time

These times only apply to the weekday. The instructor will try to be available on the weekend but it should not be expected.

Medium	Time
Email	48 Hours
Message Board	24 Hours
Grades	1 Week

If you do not get a response in the above listed times, email again and include the tag [second attempt] after the course tag in the email title and it will get more urgent treatment. If you use this tag and the above listed times ***have not elapsed*** the instructor might be a little grumpy.

As an example if you email on Friday at 4pm, you might not get a response till Tuesday at 4pm. This is a upper bound. Responses will usually come more quickly but sometimes things happen and it takes awhile to get back to you.

Grades usually come back within a week but some of these assignments can be complex so you might see a couple assignments which take longer.

Difference with On Campus Classes

In an on campus class about ½ to ⅓ of the class would be the instructor lecturing at you. The rest of the time would be used to do in class examples and work as small groups. This is where you would test your understanding of the topics and ask your group mates for help. It is also where you would help your fellow students. If you were totally stuck the instructor would help.

In principle online learning should be about the same. But instead of a classroom you have a message board. You should be active on the message board to help and get help from your classmates. When everyone gets stuck the instructor can help get the class unstuck. But the value of the class comes in your ability to get help from other people learning the same thing at the same time and having an instructor to help guide that learning. If you are not active on the discussion board you will get much less out of the class.

Course Policies

Discussion Participation

I encourage you to participate a lot on the discussion boards to get help and get help from your classmates. When everyone gets stuck then the instructor can help the class get unstuck. But the value of the class comes in your ability to get help from other people learning the same thing at the same time and having an instructor to help guide that learning. If you are not active on the discussion board you will get much less out of the class. Students who do participate tend to do better in the class and have a more positive experience than those who don't.

Late Work Policy

Assignments are generally due on Monday and Thursday. Each assignment will remain open until three working days after the due date when it will be marked late. So if an assignment is due on Monday, you can still submit it until end of day on Thursday and marked late. Late assignments may be deducted up to 10% but check with your instructor.

Assignments maybe accepted late. If you need an extension for any reason the request must be submitted at least 48 hours in advance, typically a 72 hour is the maximum extension, but no documentation is required. If you need an **emergency** extension it must be requested within 48 hours after the assignment was due and must be accompanied by documentation supporting the request (eg. doctors note, visit summary from the ER, police report etc.). As long as there is supporting documentation these are generally granted for up to a 72 hour extension. Longer extensions are handled on a case by case basis.

Fixing Work

Sometimes you need to make changes to a program or an assignment to get it to run correctly. You will have 7 days from when an assignment is graded to resubmit the assignment for a maximum grade of 70%. These can only be minor changes to get a program to run. You can't write a program from scratch and still get 70% credit.

Statement Regarding Religious Accommodation

Oregon State University is required to provide reasonable accommodations for employee and student sincerely held religious beliefs. It is incumbent on the student making the request to make the faculty member aware of the request as soon as possible prior to the need for the accommodation. See the Religious Accommodation Process for Students.

Guidelines for a Productive and Effective Online Classroom

(Adapted from Dr. Susan Shaw, Oregon State University)

Students are expected to conduct themselves in the course (e.g., on discussion boards, email) in compliance with the university's regulations regarding civility. Civility is an essential ingredient for academic discourse. All communications for this course should be conducted constructively, civilly, and respectfully. Differences in beliefs, opinions, and approaches are to be expected. In all you say and do for this course, be professional. Please bring any communications you believe to be in violation of this class policy to the attention of your instructor.

Active interaction with peers and your instructor is essential to success in this online course, paying particular attention to the following:

- Unless indicated otherwise, please complete the readings and view other instructional materials for each week before participating in the discussion board.
- Read your posts carefully before submitting them.
- Be respectful of others and their opinions, valuing diversity in backgrounds, abilities, and experiences.
- Challenging the ideas held by others is an integral aspect of critical thinking and the academic process. Please word your responses carefully, and recognize that others are expected to challenge your ideas. A positive atmosphere of healthy debate is encouraged.

Expectations for Student Conduct

Student conduct is governed by the university's policies, as explained in the Student Conduct Code (https://beav.es/codeofconduct). Students are expected to conduct themselves in the course (e.g., on discussion boards, email postings) in compliance with the university's regulations regarding civility.

Academic Integrity

Integrity is a character-driven commitment to honesty, doing what is right, and guiding others to do what is right. Oregon State University Ecampus students and faculty have a responsibility to act with integrity in all of our educational work, and that integrity enables this community of learners to interact in the spirit of trust, honesty, and fairness across the globe.

Academic misconduct, or violations of academic integrity, can fall into seven broad areas, including but not limited to: cheating; plagiarism; falsification; assisting; tampering; multiple submissions of work; and unauthorized recording and use.

It is important that you understand what student actions are defined as academic misconduct at Oregon State University. The OSU Libraries offer a tutorial on academic misconduct, and you can also refer to the OSU Student Code of Conduct and the Office of Student Conduct and Community Standard's website for more information. More importantly, if you are unsure if something will violate our academic integrity policy, ask your professors, GTAs, academic advisors, or academic integrity officers.

Plagiarism

Keeping with the idea of needing to communicate with your classmates, you will need to share code with each other to communicate ideas and to explain problems. That said, you should only share what is needed. This is good practice in general. Having to look through an entire file to find a single function is not helpful when you could have just posted the functions.

You will also get great help from your classmates. If you use a code fix or suggestion from them you need to do two things.

1. You need to cite who it came from, that means you need to, in your code, add a comment saying that the code was not your original code and say what student or 3rd party provided it.
2. You need to document what the code is doing. So you should write roughly a sentence per line of code explaining what the snippet is doing. So if you were to say, use an entire 100 line file provided by a student you would need to write something around a 2,000 word essay describing what the code is doing.

When giving help, don't give more help than is needed. Try to keep answers deep in depth but narrow in focus. Address the problem the student is asking about and address it well, but keep from giving away too much about topics they might not have run into yet.

TurnItIn

Your instructor may ask you to submit one or more of your writings to Turnitin, a plagiarism prevention service. Your assignment content will be checked for potential plagiarism against Internet sources, academic journal articles, and the papers of other OSU students, for common or borrowed content. Turnitin generates a report that highlights any potentially unoriginal text in your paper. The report may be submitted directly to your instructor or your instructor may elect to have you submit initial drafts through Turnitin, and you will receive the report allowing you the opportunity to make adjustments and ensure that all source material has been properly cited. Papers you submit through Turnitin for this or any class will be added to the OSU Turnitin database and may be checked against other OSU paper submissions. You will retain all rights to your written work. For further information, visit Academic Integrity for Students: Turnitin – What is it?

Statement Regarding Students with Disabilities

Accommodations for students with disabilities are determined and approved by Disability Access Services (DAS). If you, as a student, believe you are eligible for accommodations but have not obtained approval, please contact DAS immediately at 541-737-4098 or at http://ds.oregonstate.edu. DAS notifies students and faculty members of approved academic accommodations and coordinates implementation of those accommodations. While not required, students and faculty members are encouraged to discuss details of the implementation of individual accommodations.

Accessibility of Course Materials

All materials used in this course are accessible. If you require accommodations please contact Disability Access Services (DAS).

Additionally, Canvas, the learning management system through which this course is offered, provides a vendor statement certifying how the platform is accessible to students with disabilities.

Tutoring and Writing Assistance

TutorMe is a leading provider of online tutoring and learner support services fully staffed by experienced, trained and monitored tutors. Access TutorMe from within your Canvas course menu.

The Oregon State Online Writing Suite is also available for students enrolled in Ecampus courses.

Ecampus Reach Out for Success
University students encounter setbacks from time to time. If you encounter difficulties and need assistance, it's important to reach out. Consider discussing the situation with an instructor or academic advisor. Learn about resources that assist with wellness and academic success.

Ecampus students are always encouraged to discuss issues that impact your academic success with the Ecampus Success Team. Email ecampus.success@oregonstate.edu to identify strategies and resources that can support you in your educational goals.

- **For mental health:**
  Learn about counseling and psychological resources for Ecampus students. If you are in immediate crisis, please contact the Crisis Text Line by texting OREGON to 741-741 or call the National Suicide Prevention Lifeline at 1-800-273-TALK (8255).
- **For financial hardship:**
  Any student whose academic performance is impacted due to financial stress or the inability to afford groceries, housing, and other necessities for any reason is urged to contact the Director of Care for support (541-737-8748).

Student Evaluation of Courses
During Fall, Winter, and Spring term, the online Student Evaluation of Teaching system opens to students the Wednesday of week 8 and closes the Sunday before Finals Week.
Students will receive notification, instructions and the link through their ONID email. They may also log into the system via Online Services. Course evaluation results are extremely important and used to help improve courses and the learning experience of future students. Responses are anonymous (unless a student chooses to "sign" their comments, agreeing to relinquish anonymity) and unavailable to instructors until after grades have been posted. The results of scaled questions and signed comments go to both the instructor and their unit head/supervisor. Anonymous (unsigned) comments go to the instructor only.