

## Introduction to Computing for Data Science

**Course Objective:** This course offers an introduction to the skills needed for data analysis and is intended for life science students with limited programming, scripting or data analytics experience.

**Instructors:**

Stephen Ficklin, Ph.D.

Assistant Professor of Systems Genetics and Bioinformatics

Dept. of Horticulture

Office: 153 Johnson Hall,

Phone: (509) 335-4295

Email: [stephen.ficklin@wsu.edu](mailto:stephen.ficklin@wsu.edu)

Tyler Biggs, Ph.D.

Postdoctoral Researcher

Dept. of Horticulture and Chemistry

Office: 158 Johnson Hall

Email: [biggst@wsu.edu](mailto:biggst@wsu.edu)

**Office Hours:**

Ficklin: 3:30-5:00pm Tues/Thurs; or by appointment.

Biggs: 9-10:30am Tues/Thurs; or by appointment.

*Note:* Office hours may change during the semester and if so, notification will be given.

**Class Times:** Pullman - TuTh 10:35am - 11:50am. Vogel Plant Sciences Bldg. Room #31

**Class Communication:** All class communication will occur using Slack (<https://slack.com/>), which is a popular project management software. Students are encouraged to post questions to slack for the instructor and other students to respond to. While it is possible for one-on-one communication within slack, communication regarding class performance should be conducted in person or via email. This class has a dedicated slack project and the subscription <https://join.slack.com/t/hort503-01-spring2018/signup>. You must use your WSU email address to sign up. Slack can be installed on a desktop or laptop (at <https://slack.com/downloads>), or on a tablet or smart phone (via Google Play Store or IOS App Store).

**Prerequisites:** A laptop is required.

**Required Text:**

1. Zed A. Shaw. Learn Python 3 the Hard Way. Purchase options:
  1. Directly from the author: <https://learncodethehardway.org/python/>
  2. Amazon: <https://www.amazon.com/gp/product/0134692888>
  3. **Note:** you will need to purchase a digital version so that we can immediately begin work.
2. During the course additional reading assignments will be provided from online tutorials.

**Expected Student Learning Outcomes:**

Gain experience in the basic skills needed for data science, including python programming; use of python libraries for data cleaning, wrangling, analysis and visualization; basics of machine learning; and utilization of computing infrastructure including high performance computing resources, bioinformatics software installation and an intro to relational databases and SQL.

**Course Schedule:** This course is meant to be flexible and adjust to the learning needs of students. Adjustments to the following schedule may occur. If changes do occur an updated course schedule will be provided.

| Week  | Day | Date              | Topic   | Assignment  |
|---|-----|-------------------|---|---|
| 1   | 1   | 9-Jan             | Course Introduction, System Setup   | A01: Command-Line, (pp. 288-328). Due Jan 11 <sup>th</sup><br>A02: Shaw exercises #1-17, Due Jan 15 <sup>th</sup> . |
| 1   | 2   | 11-Jan            | Command-line crash course, GitHub<br><i>Jan 15<sup>th</sup>, Martin Luther King Jr. Holiday</i> |   |
| <b>MODULE #1: Introduction to Python Coding</b>           |     |                   |   |   |
| 2   | 3   | 16-Jan            | Variables, Strings, Printing, Files   | A03: Shaw exercises #18-23, 27-34, Due Jan 22 <sup>rd</sup> .   |
| 2   | 4   | 18-Jan            | Variables, Strings, Printing, Files   |   |
| 3   | 5   | 23-Jan            | Functions, Logic, Control Statement   | A04: Shaw exercises #35-42. Due Jan 29 <sup>th</sup> .  |
| 3   | 6   | 25-Jan            | Functions, Logic, Control Statement   |   |
| 4   | 7   | 30-Jan            | Lists, Dictionaries   | A05: Shaw exercises #43-44. Due Feb 5 <sup>th</sup> .   |
| 4   | 8   | 1-Feb             | Modules, Classes & Objects  | Module #1 Project given   |
| 5   | 9   | 6-Feb             | Modules, Classes & Objects  | A06: Jupyter, Numpy, & matplotlib.  |
| 5   | 10  | 8-Feb             | Inheritance, Composition  | Module #1 project due Friday 9 <sup>th</sup> at 5pm PST   |
| <b>MODULE #2: Python for Data Science</b>                 |     |                   |   |   |
| 6   | 11  | 13-Feb            | SciPy: NumPy, Jupyter Notebooks   | A07: Pandas Intro, etc. Due Feb 13 <sup>th</sup> .  |
| 6   | 12  | 15-Feb            | SciPy: Matplotlib visualization<br><i>Feb 19<sup>th</sup>, President's Day Holiday</i>          |   |
| 7   | 13  | 20-Feb            | SciPy: Pandas   | A08: RegEx & CL Data Wrangling  |
| 7   | 14  | 22-Feb            | SciPy: Pandas   |   |
| 8   | 15  | 27-Feb            | Regular Expressions   | A09: Python Data Cleaning   |
| 8   | 16  | 1-Mar             | Command-line data wrangling   |   |
| 9   | 17  | 6-Mar             | Data Cleaning with Tidy   | A10: Interactive visualization with Bokeh   |
| 9   | 18  | 8-Mar             | Data Cleaning with Pandas<br><i>March 12-16, Spring Break</i>                                   |   |
| 10  | 19  | 20-Mar            | Interactive Visualization: Bokeh  | A11: Basics of Machine Learning   |
| 10  | 20  | 22-Mar            | Interactive Visualization: Bokeh  | Module #2 project given   |
| 11  | 21  | 27-Mar            | Basics of Machine Learning  | A12: HPC -- Kamiak  |
| 11  | 22  | 29-Mar            | Basics of Machine Learning  | Module #2 project due Friday 30 <sup>th</sup> at 5pm EDT.   |
| <b>Module #3: Utilization of Computing Infrastructure</b> |     |                   |   |   |
| 12  | 23  | 3-Apr             | High Performance Computing  | Final Project Given<br>A13: Workflow development using Nextflow   |
| 12  | 24  | 5-Apr             | High Performance Computing  |   |
| 13  | 25  | 10-Apr            | Workflow Development  | A14: Bioinformatics Software installation   |
| 13  | 26  | 12-Apr            | Workflow Development  |   |
| 14  | 27  | 17-Apr            | Bioinformatic Software Installation   | A15: Intro to relational databases and SQL.   |
| 14  | 28  | 19-Apr            | Bioinformatic Software Installation   |   |
| 15  | 29  | 24-Apr            | Relational Databases & SQL  |   |
| 15  | 30  | 26-Apr            | Relational Databases & SQL  |   |
| 16  |     | Apr 30 -<br>May 4 | Finals Week   | Final Project Due Friday May 4 <sup>th</sup> at 5pm PDT.  |

**Assessment:** A final grade will be assigned at the end of the course in accordance with the following percentages.

| Assessment Item     | Percentage  | Points                                     |
|---------------------|-------------|--|
| Projects            | 60%         | 200 points per project: 600 total points   |
| Assignments         | 35%         | 30 points per assignment: 420 total points |
| Group Participation | 5%          | 50 points, assigned by peers.              |
| <b>Total</b>        | <b>100%</b> | <b>1000</b>                                |

Changes during the course may increase or reduce the total number of points that can be earned. However, the percentage for each type of assessment item will remain unchanged.

*Projects:* Because this course is primarily a hands-on experience there will be no traditional exams. Instead a project will be assigned for each of the three modules described in the course schedule. The project description will be provided at least 1.5 weeks before it is due. Each student is to work on the project independently. Students should not consult with one another to work on projects unless they do so using Slack. This is because the communication via Slack is preserved. However, example solutions or code for projects cannot be posted to Slack. The final project will be given two weeks before the last day of finals and may be cumulative.

*Assignments:* On the first-class day of each week (Tuesdays) a homework assignment is given. The assignment is meant to introduce the topic that will be covered in class the following week. Completion of the assignments is critical to ensure maximal understanding of the material. Students have the full week and weekend to work on the assignment. Assignments are not graded except for completeness. If the assignment is fully completed and turned in on time, then full credit is given. If an assignment is late then 5 points are subtracted for each day late. It is highly recommended that students begin assignments as soon as possible and not wait until the end of the week to start the assignment.

*Group Participation:* This course is hands-on and in-class activities may include programming or data analysis. Often students should help each other especially if the instructors are busy assisting other students. When a student is helped by another student, the student being helped can award 2 points to the helper student. These points can be awarded by sending a message via slack to the instructor informing of the points. The instructors will remind students to give points at the end of each class period. Help does not need to result in success. Points can be awarded for simply attempting to help. Additionally, the instructors can award points to students who provide answers to other student’s questions on slack. A maximum of 50 points can be awarded for group participation.

**Mid-Term Grade:** The mid-term grade will be composed from scores of the assignments that have been turned in to that point and the same percentage based scores will be used as at the end of the semester.

**Grading scale:**

|               |               |               |              |
|---------------|---------------|---------------|--------------|
| 100 - 93% = A | 86 - 83% = B  | 76 - 73% = C  | 65 - 60% = D |
| 92 - 90% = A- | 82 - 80% = B- | 72 - 70% = C- | ≤59 = F      |
| 89 - 87% = B+ | 79 - 77% = C+ | 69 - 66% = D+ |              |

**Expectations:** Assignments are geared towards the expectation that student will spend 1-3 hours per credit hour outside of the class meetings (3-9 hours per week). Please see the instructor if your efforts require more (or less) than this. It is important that the course meet learning expectations for the topic but also not be over burdensome. Please bring laptops for in class exercises, but silence cell phones and turn off social media during class (e.g. skype, facebook, twitter, etc.).

**Attendance:** Attendance of this class is highly encouraged for student success. Absences because of illness, personal and/or family crises, mandated court appearances, university approved events, or similar reasons will be accommodated if such absences are not excessive, and notification is provided to the instructor in advance. Excused absences should be arranged prior to any known or planned event. Required University activities will be excused absences if an official Class Absence Request form signed by the sponsoring faculty or organization is given to the instructor before the event. If a student is ill and wishes to attend remotely and can notify the instructor with enough advanced warning, the instructor will attempt to coordinate use of remote conference services to help the student attend.

**Disability Statement:** Reasonable accommodations are available for students with a documented disability. If you have a disability and need accommodations to fully participate in this class, please either visit or call the Access Center (Washington Building 217; 509-335-3417) to schedule an appointment with an Access Advisor. All accommodations MUST be approved through the Access Center. For more information contact a Disability Specialist: 509-335-3417, <http://accesscenter.wsu.edu>, [Access.Center@wsu.edu](mailto:Access.Center@wsu.edu)

**Academic Integrity:** “As an institution of higher education, Washington State University is committed to principles of truth and academic honesty. All members of the University community share the responsibility for maintaining and supporting these principles. When a student enrolls in Washington State University, the student assumes an obligation to pursue academic endeavors in a manner consistent with the standards of academic integrity adopted by the University. To maintain the academic integrity of the community, the University cannot tolerate acts of academic dishonesty including any forms of cheating, plagiarism, or fabrication. Washington State University reserves the right and the power to discipline or to exclude students who engage in academic dishonesty.”

Students found responsible for academic integrity violations may receive an F on the particular assignment or exam, as well as an F for the course. Repeated and/or serious offenses may result in referral to the conduct board and expulsion from WSU. For graduate students, academic integrity violations may also result in the loss of teaching and/or research assistantships.

Academic Integrity Statement and link to WSU’s policy:

- <http://www.wsulibs.wsu.edu/plagiarism/main.html>
- <http://conduct.wsu.edu/academic-integrity-policies-and-resources/>

**Safety:** Washington State University is committed to enhancing the safety of the students, faculty, staff, and visitors. It is highly recommended that you review the Campus Safety Plan (<http://safetyplan.wsu.edu/>) and visit the Office of Emergency Management web site (<http://oem.wsu.edu/>) for a comprehensive listing of university policies, procedures, statistics, and information related to campus safety, emergency management, and the health and welfare of the campus community.