BIOL381/2 - Syllabus
Introduction to Data Science for Biology

Instructor Information

Jarrett Byrnes, Ph.D.
jarrett.byrnes@umb.edu
Phone (W): 617-867-3145
Office Location ISC 3140
Office Hours: Wednesdays 11-12:30

TA: Jillian Dunic
jillian.dunic@umb.edu
Office Location: ISC 3100
Office Hours: TBD

Course Information

Course Title: Introduction to Data Science for Biology
Credits: **4, Lecture and Lab**
Time: **T/Th 11-12:15 lecture, W 12:30-3:30**
Online? no
Course Description: This course will introduce undergraduates to the basic concepts of how we use data in the biological sciences. We will emphasize data creation, curation, manipulation, visualization, and some analysis. This course should prepare students for any data-intensive position or course in biology or other disciplines they might encounter in the future.

Context: Students interested in research in biology will need to take information from the lab bench or field site and translate that into meaningful inferences about the biological processes they are studying. This course will arm them with the skills they need to be successful biological researchers. It will enable them to take complex datasets and distill them into meaningful information from which they can draw reasoned conclusions. It will also introduce them to a suite of computational tools that are gaining popularity in biology and beyond for the integration and analysis of data.

Prerequisites: Two of BIOL 210, 252, 290
OR
Two of EEOS 210, 226, 261, 267L

Prerequisite Skills: Experience with programming is helpful, but not assumed.

Course Objectives: By fully participating in this course, you should be able to:
1. Learn how to create efficient understandable datasets for biological research.

2. Build a vocabulary of visualization tools that enable students to see what their data means.
3. Develop an understanding of how to manipulate data for the purposes of seeing useful patterns.
4. Understand how to unify data from disparate sources to build a larger picture of biological phenomena.
5. Learn basic analytical tools for deriving statistical inference from data.
6. Learn common programming languages associated with data science.

Core Competencies: The objectives for this course focus on the following core competencies:

1. Graduates should emerge with a broad understanding of how to use data to draw inferences about biological processes.
2. Graduates should have the confidence and skills to continue using R or other software for data manipulation, visualization, and analysis.
3. Graduates should have an appreciation for the ways that computational tools can improve the efficiency of their research.
4. Graduates should emerge as better data scientists.

Required Assignments: Students will have three forms of graded work. First, students will be expected to turn in write-ups based on lab exercises each week. Second, students are expected to turn in a weekly homework problem set. Last, students will be asked to write a short report at the end of the class where they show the different steps of working with a data set or data sets of their choice to demonstrate their ability to draw inferences from data.

Course Rubric:

Assignment/Deliverable	Number	Grade %
1. Weekly assignments	**14**	**30**
2. Lab Assignments	**14**	**30**
3.		
4.		
5.		
6.		
7.		
Final Project/Presentation	**1**	**35**
Group Work		
Participation (as defined above)		
Attendance (as defined above)		**5**

Course

Policies: Participation – Participation in the course includes completing all required readings, problem sets, and weekly lab exercises. Students are expected to create a thoughtful learning environment by asking questions and working together to help solve problems.

- Attendance - Students are expected to attend all classes and labs. Sick students are expected to bring a note from a doctor. Students who are otherwise prevented from coming to class are expected to bring a note from the relevant party (e.g., if your car is towed, a the relevant ticket). Two unexcused absences will reduce your maximum grade to a B. Three to a C. Four or more absences to an F.
- Group Work – Some problem sets and final projects can be done as a group. Students are expected to identify their contribution to group work honestly and understand that groups are graded together.
- Late Work – Late work loses 3% point per day late. Exceptions can only be made in the case of a documented emergency.

Grading

Grading: Grade type for the course is a whole or partial letter grade. (Please see table below) Note: the lowest passing grade for a graduate student is a "C". Grades lower than a "C" that are submitted by faculty will automatically be recorded as an "F". Please see the Course Catalog for more detailed information on the University's grading policy.

Grading Policy			
	Letter Grade	**Percentage**	**Quality Points**
	A	93-100%	4.00
	A-	90-92%	3.75
	B+	87-89%	3.25
	B	83-86%	3.00
	B-	80-82%	2.75
	C+	77-79%	2.25
	C	73-76%	2.00
	F	0-72%	0.0
	INC	A grade of Incomplete (INC) is not automatically awarded when a student fails to complete a course. Incompletes are given at the discretion of the instructor. They are awarded when satisfactory work has been accomplished in the majority of the course work, but the student is unable to complete course requirements as a result of circumstances beyond his/her control. The student must negotiate with and receive the approval of the course instructor in order to receive a grade of incomplete	N/A
	IF	Received for failure to comply with contracted completion terms.	N/A
	W	Received if withdrawal occurs before the withdrawal deadline.	N/A
	AU	Audit (only permitted on space-available basis)	N/A
	NA	Not Attending (student appeared on roster, but never attended class. Student is still responsible for tuition and fee charges unless withdrawal form is submitted before deadline. NA has no effect on cumulative GPA.)	N/A

Required Text(s): Grolemund, G., and Wickham, W. 2016. R for Data Science. The book is in progress and can be found online at http://r4ds.had.co.nz/

Recommended Texts

Wickham, H. 2014. Advanced R. The book can be found online at http://adv-r.had.co.nz/

Technical Requirements: Access to a computer with the R programming language and Rstudio. This will be provided here at UMB.

Course Schedule

Week 1. Data and Metadata.

Readings: G&W Introductory chapters.

Objective(s): Introduce the students to the course; understand what is data, discuss how we preserve information about data, view different examples of datasets from different disciplines.

Lab: Introduction to Excel, How would you describe different biological observations as data?

Week 2. Data Creation

Objective(s): Compare poor versus good practice in creating data. Differentiate between data recording and data entry, Develop a practical familiarity with data quality control

Lab: Planning and collecting data. Meet on the 3rd floor of the ISC.

Week 3 & 4. Visualization & Introduction to R

Readings: G&W Chapter Import, Unwin 2008, Wickham 2010

Objective(s): Begin to learn the R computing language, develop understanding of graphical presentation best practices. Identify the syntax of an R function (name and arguments); Create an R project in RStudio; Read data into R using read.csv(); Use R as a basic calculator; Describe and create variables in R; Interpret the output of the str() function; Install packages in R; Create a scatterplot using ggplot();

Labs: Bringing data into R, visualization of Plankton data via ggplot2

Week 5&6. Data Reduction

Readings: Handout on Descriptive Statistics, Anderson 2014

Objective(s): Describe the meaning and identify applications of the following summary/descriptive statistics: mean, mode, median, standard deviation; Describe the split-apply-combine strategy of data reduction and summarization; Use group_by() and summarise() to calculate summary statistics for groupings within a dataset; Subset data using filter()

Lab: Descriptive statistics in R, Introduce vectors, dplyr for Microbial data aggregation

Week 7. Tidy Data

Readings: G&W Chapter on Tidy, Wickham 2014

Objective(s): Understand how to reshape and manipulate data. Describe the difference between the two fundamental forms of data – long versus wide, Use the tidyr package in R to convert between long and wide data; Use unite and separate to create tidy data (where each column is a variable)

Lab: Tidyr and data reshaping with Lizard Evo-Devo data

Week 8. Databases

Readings: TBD.

Objective(s): Define the term relational database; Draw a schema for a simple, 4 table relational database (including one-two-one and one-to-many- relationships); Explain to another person the relationships between tables/variables when given a database schema.

Lab: Create and navigate a 4 table relational database. Provide verbal descriptions of how to query the database.

Week 9. Data "Mashups"

Readings: Joins handout

Objective(s): Know when and where to use different types of joins, Understand how to merge survey data with geospatial information to get a geographic understanding of epidemiological patterns

Lab: Introduction to geospatial visualization, merging plankton data with maps, sp and Rvest libraries.

Week 10. Accessing Online Data

Readings: TBD by Todd Riley

Objective(s): List the major source of online data for bioinformatics and ecoinformatics, acquire the basic tools for scraping data from the web

Lab: Querying Genbank, Accessing NOAA buoy data, RCURL and web scraping

Week 11 & 12. T-Tests and P-Values

Readings: Cortina and Dunlop 1997

Objective(s): Describe the basics of probability and p-values, Compare groups of data using T-tests and its extensions

Lab: Implementing statistical tests in R for Microbial Abundance Data, Data Simulation and P-Values, Evaluation of assumptions

Week 13 & 14. Linear Regression

Readings: Handouts on linear regression

Objective(s): Fit a linear regression using lm() in R through a bivariate scatterplot, Describe when to use nonlinear models/curves

Lab: Fitting linear models in R, Testing Assumptions, Evaluating model outputs and generating predictions, using lizard evo-devo data

Methods of Instruction

Methods: This course will be a mixture of lecture, live-code demonstrations, and opportunities for in class work. Lecture days will have small exercises for students at the end of class. Labs will allow for a longer exploration of material from the week, with exercises designed to build skills and confidence and culminating in a weekly problem set. We will conduct lectures and labs in a computer lab in order for students to be able to follow along and try out new concepts once described and demonstrated in lecture, enabling rapid feedback between students and faculty.

Accommodations

The University of Massachusetts Boston is committed to providing reasonable academic accommodations for all students with disabilities. This syllabus is available in alternate format upon request. If you have a disability and feel you will need accommodations in this course, please contact the Ross Center for Disability Services, Campus Center, Upper Level, Room 211 at 617.287.7430. http://www.umb.edu/academics/vpass/disability/ After registration with the Ross Center, a student should present and discuss the accommodations with the professor. Although a student can request accommodations at any time, we recommend that students inform the professor of the need for accommodations by the end of the Drop/Add period to ensure that accommodations are available for the entirety of the course.

Academic Integrity and the Code of Student Conduct

Code of Conduct and Academic Integrity

It is the expressed policy of the University that every aspect of academic life--not only formal coursework situations, but all relationships and interactions connected to the educational process--shall be conducted in an absolutely and uncompromisingly honest manner. The University presupposes that any submission of work for academic credit is the student's own and is in compliance with University policies, including its policies on appropriate citation and plagiarism. These policies are spelled out in the Code of Student Conduct. Students are required to adhere to the Code of Student Conduct, including requirements for academic honesty, as delineated in the University of Massachusetts Boston Graduate Catalogue and relevant program student handbook(s).
UMB Code of Student Conduct

You are encouraged to visit and review the UMass website on *Correct Citation and Avoiding Plagiarism:* http://umb.libguides.com/citations

Other Pertinent and Important Information

Incomplete Policy: Students who must take an incomplete have one year to finish agreed upon work to get a grade for the course.
Coursework Difficulties: Please discuss all coursework matters with me or the TA sooner than later.
Withdrawing From This Course: Please refer to the written policies and procedures on formal withdrawal and add/change dates listed in the Graduate Studies Catalog.

You are advised to retain a copy of this syllabus in your personal files for use when applying for future degrees, certification, licensure, or transfer of credit.

Bibliography

Anderson, S.C. 2014. dplyr and pipes: the basics. http://seananderson.ca/2014/09/13/dplyr-intro.html

Cortina, J.M., Dunlap, W.P., 1997. On the logic and purpose of significance testing. Psychological Methods.

Wickham, H., 2010. A layered grammar of graphics. Journal of Computational and Graphical Statistics 19, 3–28.

Wickham, H., 2014. Tidy Data. J. Stat. Soft. 59, 1–23.

Unwin, A., 2008. Good Graphics? Handbook of data visualization.