# 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: TBD

Office Location: ISC 3100

Office Hours: TBD

# Course Information

**Course Title:** Introduction to Data Science for Biology

**Credits: 1**

**Time: W 1:00-3:00**

**Location: Wheatley 02-030**

**Online?** no

**Course**

**Description:** This is the lab for An Introduction to Data Science for Biology. The lab will introduce students to how to implement conceptual material from the course using the computer language R.

**Context:** This lab will arm students with the skills they need to be successful data scientists in biological research and beyond. 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. This course will emphasize both good data science and programming skills throughout, with a particular focus on applications to biology.

**Corequisite:** BIOL 3XX, the lecture for the course

**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 acquire a variety of data types and bring them into a computational environment.
2. Be able to merge multiple different data sources to derive more meaningful answers to biological questions.
3. Be able to clean and visualize datasets before analysis.
4. Be able to extract summary properties from any data set.
5. Be able to perform basic analysis of datasets.

**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 scientific computing tools 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 turn-in weekly lab writeups. Each week we will introduce new skills to students, and then ask them to implement those skills on a novel data set for their write-up.

**Course Rubric:**

|  |  |  |
| --- | --- | --- |
| **Assignment/Deliverable** | **Number** | **Grade %** |
| **1. Weekly write-up** | **14** | **80** |
|  |  |  |
| **Attendance** (as defined above) |  | **20** |

**Course**

**Policies:**

* + 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:** Grolemund, G., and Wickham, W. 2016. R for Data Science. This book is available online for free at <http://r4ds.had.co.nz/>

Data Carpentry Lessons. Ongoing. http://www.datacarpentry.org/lesson

**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

*Note: G&W refers to Grolemund and Wickham. DC indicates a Data Carpentry lesson.*

Week 1. Data and Metadata.

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: Planning and collecting data. Meet on the 3rd floor of the ISC. We will use a survey of different architectural properties of the ISC (or the grounds around it, if open) as a means to discuss how to effectively design a data collection process for easy deposition into a data processing pipeline. Please bring a notebook!

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: Introduction to Excel. We will use the [Data Carpentry Spreadsheets in Ecology lesson](http://www.datacarpentry.org/spreadsheet-ecology-lesson/) and lab to provide a rich introduction to Microsoft Excel, Google Sheets, and general paradigms in data object creation. For writeup, students will enter and format their data and metadata from the previous week, explicitly noting traps and pitfalls found along the way.

Week 3 & 4. Visualization & Introduction to R

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: In week 3, students will continue the introduction to R begun in lecture. Students will create and manipulate the basic data types within R. We will also focus creating new R project and data importing. For lab write-ups, students will load and summarize the data they generated in the previous week.

In week 4, students will learn the basics of the *ggplot2* package using the Plum Island LTER plankton community composition dataset as a sample data set. We will emphasize both basic data visualization and introduce spatial data visualization with *ggmap.* As this is a large dataset, for lab write-up students will be asked to complete visualizations using the species of their choice.

Week 5&6. Data Reduction and Summarization

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: Students will be introduced to pipes, and the basics of *dplyr* for data summarization. Using data on sockeye salmon, we will work through techniques for summarizing data and visualizing summarizations. Students will then use the workflows to visualize data on human genome size. In week six, we will extend this work to building interactive visualizations of climate change over the last century with *gganimate* and *plotly*.

Week 7. Tidy Data and Data Cleaning

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). Understand how to manipulate string data

Lab: Using *tidyr* we will learn how to read and clean a dataset on Axoltl Limb Regeneration data created without thought for analysis and visualization in a scientific computing workflow. Students will then apply these techniques to a dataset of salt marsh sedimentation rates.

Week 8. Functions

Objectives: Learn the benefits of reusable code, Understand the structure of a function, Discover debugging and making functions fail usefully, Apply conditional logic to build flexible code, Derive principles to make functions that are easy to understand and apply to multiple data sets.

Lab: Learn to write functions for multiple simple cleaning steps in parsing NOAA temperature buoy data. Students will learn how to write functions for easy use with *tidyverse* tools, and then develop a workflow to parse three decades worth of sea surface temperature data.

Week 9 & 10. Data “Mashups” and Geospatial Data

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: Students will learn how to merge geospatial data with standard data frames to visualize rates of change in kelp abundance across the Marine Ecoregions of the World. Students will learn the *sp, rgdal*, and *leaflet* libraries for geospatial visualization. Students will demonstrate skills using NIH data on cancer prevalence across different counties in the U.S.

Note: To install gdal on a mac, there are two steps  
1) Install Homebrew from <http://brew.sh/> (this is an awesome thing to have anyway)  
2) in Terminal type  
*brew install gdal*

To install on a Windows PC  
1) Install OSGEO4W <https://trac.osgeo.org/osgeo4w/wiki>

2) Use it to install gdal

Week 11 & 12. T-Tests, ANOVA, 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: We will work through the standard workflow for implementing, interpreting, and visualizing datasets with categorical predictors (t-tests and ANOVA) using data on Blackbird testosterone concentrations and circadian clock rhythms. We will focus on the analytic workflow, merging previous concepts with analytical techniques. Students will then work on faded examples of different elements of analytic workflows, and then choose one data set from those examples for a complete analysis each week.

Week 13 & 14. Linear Regression

Readings: G&W on [Model Basics](http://r4ds.had.co.nz/model-basics.html) and [Model Building](http://r4ds.had.co.nz/model-building.html)

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

Lab: We will work through first linear and then nonlinear regression models using data on seal morphometrics and survivorship. We will focus on the analytic workflow, merging previous concepts with analytical techniques. Students will then work on faded examples of different elements of analytic workflows using data on species richness and fire severity as well as naked mole rat life history. Students will choose one data set from those examples for a complete analysis during linear regression week and then revisit it with nonlinear techniques the following week.

# Methods of Instruction

**Methods:** This course will be a mixture of live-code demonstrations, faded examples, parsons problems, and free work on problems. Lab will be conducted in a computer lab.

# 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](http://www.umb.edu/life_on_campus/policies/code/)

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

Penalties for academic misconduct in the course, including plagiarism and cheating, are strictly enforced, and the penalties are very serious. Penalties include an F in the assignment or exam, an F in the course, or suspension from the University. If you have questions about what constitutes plagiarism or other forms of academic misconduct, see Prof. Byrnes **before** completing an assignment or exam.

**Ignorance of the rules does not excuse any academic conduct violation.**

**The University defines violations to include, but not be limited to, the following:**

* Submitting as one's own an author's published or unpublished work (e.g. material from a journal, Internet site, newspaper, encyclopedia), in whole, in part, or in paraphrase, without fully and properly crediting the author.
* Submitting as one's own work or materials obtained from another student, individual, or agency without full and proper attribution.
* Submitting as one's own work material that has been produced through unacknowledged or unauthorized collaboration with others.
* Submitting substantially the same work to more than one course (i.e., dual or multiple submission) without prior approval from all instructors involved.
* Using any unauthorized material during an examination, such as notes, tests, calculators, cell phones, or other electronic devices.
* Obtaining answers to examination questions from another person with or without that person's knowledge; furnishing answers to examination questions to another student; using or distributing unauthorized copies of or notes from an examination.
* Submitting as one's own an examination taken by another person; or taking an examination in another person's place.
* Interfering with an instructor's ability to evaluate accurately a student's competence or performance; misleading any person in connection with one's academic work.

**Plagiarism**

Plagiarism is defined by UMass Boston’s Code of Student Conduct (<http://www.umb.edu/life_on_campus/policies/code/> ). An act of academic dishonesty, plagiarism can include actions such as presenting another writer’s work as your own work; copying passages from print or internet sources without proper citation; taking ideas off the internet, modifying them, and presenting them as your own; or submitting the same work for more than one course.  If you plagiarize, you will fail this course.  Plagiarism cases will be referred to the Dean.  Plagiarism can result in further academic sanctions such as suspension from the University.

**Civility**

An educational institution is a unique cultural space: here, the open sharing of ideas is not only possible, but valued above all else.  Intellectual exchange depends on showing respect for your instructor and peers, taking responsibility for your own course contributions, and demonstrating a mature understanding that learning can involve disagreement over ideas and assessment.  If you engage in uncivil behavior, such as making inappropriate comments to your professor or fellow students in the classroom, out of the classroom, or via email or social networking sites, you can be referred to the Dean of Students.”

# Other Pertinent and Important Information

**Incompletes:**

Incompletes are rarely offered, as they are reserved for students who are unable to complete a small portion of the course at the end of the term due to an extreme circumstance such as illness.  Incompletes are not allowed to replace a significant amount of coursework or absences.  If you are awarded an Incomplete, you must complete a formal Incomplete Contract with your instructor and have that contract approved by the Department and submitted to the Registrar. The contract outlines the work to be done and due dates.  An INC automatically turns into an F after a year if the work is not completed.

Incomplete policy: [http://www.umb.edu/registrar/academic\_policies/incomplete\_policy/](https://ch1prd0102.outlook.com/owa/redir.aspx?C=1Af-1kK4Uk65slJpkYslPOIzYNQA1s8IbunRr2GGHihn6r0BeqdEgCh_53AUX46TJvdTkOxWNlU.&URL=http%3a%2f%2fwww.umb.edu%2fregistrar%2facademic_policies%2fincomplete_policy%2f)

**Cell Phones: Cell phones must be POWERED OFF during class.**   
  Much of this class is discussion, and use of phones in class is disruptive and disrespectful to your fellow students to withdraw from the conversation. I will give you one warning inside or outside of class, and then ask you to please leave in any future classes if it happens again. That class will be counted as an un-explained absence.

**Attendance in WIMBA Sessions:**

* You are responsible for material covered in any class that you do not attend.
* If you miss a WIMBA session you must review the recorded class and write a 1-2-page summary of your understanding of what was covered.
* **Coursework Difficulties:** Please discuss all coursework matters with me 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.

**Additional Resources**

* Distressed and distressing students: Seek help from the Dean of Students:

[http://www.umb.edu/life\_on\_campus/dean\_of\_students/students\_in\_distress/](https://ch1prd0102.outlook.com/owa/redir.aspx?C=1Af-1kK4Uk65slJpkYslPOIzYNQA1s8IbunRr2GGHihn6r0BeqdEgCh_53AUX46TJvdTkOxWNlU.&URL=http%3a%2f%2fwww.umb.edu%2flife_on_campus%2fdean_of_students%2fstudents_in_distress%2f)

* Counseling services: Seek help from Health Services: <http://www.umb.edu/healthservices/counseling_center>
* Services for students experiencing extreme off-campus circumstances, such as homelessness or domestic violence: Seek help from the U-ACCESS Program <http://www.umb.edu/life_on_campus/uaccess>
* Services for students experiencing academic difficulties: Seek help from the CSM Student Success Center: https://www.umb.edu/academics/csm/student\_success\_center

or University Advising Center: [http://www.umb.edu/academics/vpass/uac](https://ch1prd0102.outlook.com/owa/redir.aspx?C=1Af-1kK4Uk65slJpkYslPOIzYNQA1s8IbunRr2GGHihn6r0BeqdEgCh_53AUX46TJvdTkOxWNlU.&URL=http%3a%2f%2fwww.umb.edu%2facademics%2fvpass%2fuac)

* Tutoring services, including the “Reading, Writing, and Study Strategies Center”: Seek help from Academic Support:

[http://www.umb.edu/academics/vpass/academic\_support/tutoring](https://ch1prd0102.outlook.com/owa/redir.aspx?C=1Af-1kK4Uk65slJpkYslPOIzYNQA1s8IbunRr2GGHihn6r0BeqdEgCh_53AUX46TJvdTkOxWNlU.&URL=http%3a%2f%2fwww.umb.edu%2facademics%2fvpass%2facademic_support%2ftutoring)

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

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.