PHRM 7210
Biomedical Data Literacy & Applications in R
Fall 2025 (CRN 61436)

Class Schedule:	Class will meet for twice a week during the Fall semester of 2025. **9-10am Monday WP 337** **9-11am Friday WP 337** 3 Credit hours
Instructor:	Eugene Douglass, Rm 353, 706-542-4017 Office Hours: TBD
Course Objectives:	Develop practical skills (and intuition) with data-parsing, visualization and statistical analysis for biomedical datasets.
Course Materials:	Provided by Instructor. Lecture slides, lecture code, and datasets in excel and csv format.

Grading: S/U

Discussion Boards	30%
Class Participation & Quizes	30%
Class Project	40%

Difference between concurrent Graduate (7210) and Undergraduate (5210) courses:
The graduate students are expected to prepare brief presentations on their assigned homeworks and present their final projects in the last courses of the semester. Undergraduates are not expected to present in front of the whole class. *Graduate projects are based **thesis-relevant datasets*** and therefore have a higher bar for grading. *Undergraduate projects is focused on **preparation for senior research*** and therefore are circumscribed and focused on basic scientific knowledge.

- **Undergraduate Pre-requisites:** BIOL1107, BIOL1108

Course Description:
Overall, this is an introductorily course that is designed for undergraduate and graduate students with *zero computational training*. This course has 3-part structure which we believe to be generalizable to any discipline-specific data-science course:

1. **Data-Science in Excel (33%):**
    a. reviews on essential statistical foundations leveraging students prior experience in Excel.
2. **Data-Science in R (33%):**
    a. Introduces the R statistical language to students with no coding experience
    b. Leverages Excel-based procedures to streamline introduction
3. **Biomedical Data-Science (33%):**
    a. Introduces the world of Biomedical Data
    b. Teaches standard procedures in scRNAseq and spatial transcriptomics

Figure 1. Three part structure of course emphasizes: (1) underlying statistical concepts + manual statistical procedures in excel **(2)** R-coding ability to automate manual excel procedures **(3)** standard biomedical work flows in bulk-, single-cell RNAseq and spatial transcriptomics

Data-Science in Excel (33%)

Parts 1 of the course revisit basic statistical knowledge and fundamental data-analytics procedures by leveraging a sandbox of Excel datasets for which most students have some pre-existing intuitive understanding:

- **All dogs dataset:** matrix of 200 of the most popular dog breeds.
  The matrix consists of 194 dog breeds x 22 measurements of characteristics (trainability, friendliness, intelligence, etc.) compiled by the American Kennel Club (AKC)
- **All music dataset:** matrix of the 20,000 most popular songs on Spotify.
  The matrix consists of music characteristics 19,947 songs x 15 characteristics (e.g. loudness, vocals, tempo, etc.) which are used by Spotify to personalize music recommendations for users.
- **All food dataset:** matrix of 9,000 of the most popular foods compiles by the USDA. The matrix consists of 8,889 foods x 70 nutritional categories (e.g. carbs, vitamins, minerals) that are used to inform ingredient labels on commercial food products.
- **All beer dataset:** matrix of 3,000 of the most popular beers in the United States. The matrix consists of 3,197 beers x 20 taste measurements (bitterness, sweetness, ABV, etc.)
- **All movies dataset:** matrix of 5,000 of the most popular movies from 1920-2020.
  The matrix consists of 4,822 movies x 27 measurements (e.g. box office, average rating, genres, etc.) which was obtained from IMDB and is used for recommendations of movies by streaming surfaces and multiple websites.
- **Additional datasets are welcome** but need to be approved by the instructor.

Each student will choose a dataset, which they will use to review and solidify their training in the most fundamental statistical procedures. The class is organized into 5 teams who will collaboratively analyze the above datasets using group message boards on eLC. The purpose of this section, is to review fundamental statistical procedures and build students confidence applying these procedures to large data matrices.

Figure 2. Sandbox of Intuitively accessible data-matrices to streamline statical, coding, machine-learning training in the first 2/3 of the course.

Data-Science in R (33%)

Parts 2 of this class, introduces the R-programming language as a means to streamline and automate the statistical procedures learned in Part 1. The same teams, datasets, and procedures are used in this section to streamline the transition from excel to R-code based analyses.

Biomedical Data-Science (33%)

Part 3 of this class, focuses on giving an overview of all Biomedical datasets for which the above Part 2's procedures can be applied. An introduction to all aviable datasets is presented and standard operating procedures are taught for transcriptomic methods:

- **Bulk-RNAseq**
- **Single-cell RNAseq**
- **Spatial Transcriptomics**

Which have been prioritized for two reasons. First, these have been identified as wide-spread needs across the biomedical research community at UGA. Second, UGA's new medical school will be establishing a cutting edge spatial transcriptomics core which positions UGA in a position to be a national leader in the area of in vivo transcriptomic research.

The course syllabus is a general plan for the course; deviations from this schedule will be announced to the class by the instructor via email if necessary.

Academic Integrity:

The University of Georgia has a University Honor Code and Academic Honesty Policy. All students are responsible for making themselves familiar with these codes and policies. All students are responsible for maintaining the highest standards of honesty and integrity in every phase of their academic careers. The penalties for academic dishonesty are severe and ignorance is not an acceptable defense. All academic work must meet the standards contained in "A Culture of Honesty." Students are responsible for informing themselves about those standards before performing any academic work. The link to more detailed information about academic honesty can be found

http://www.uga.edu/ovpi/honesty/acadhon.htm

Accommodations for Disabilities (e.g., If you plan to request accommodations for a disability, please register with the Disability Resource Center. They can be reached by visiting Clark Howell Hall, calling 706-542-8719 (voice) or 706-542-8778 (TTY), or by visiting http://drc.uga.edu.) See https://drc.uga.edu/content_page/sample-access-statements for additional examples.

Mental Health and Wellness Resources: • If you or someone you know needs assistance, you are encouraged to contact Student Care and Outreach in the Division of Student Affairs at 706-542-7774 or visit https://sco.uga.edu/. They will help you navigate any difficult circumstances you may be facing by connecting you with the appropriate resources or services. • UGA has several resources for a student seeking mental health services (https://www.uhs.uga.edu/bewelluga/bewelluga) or crisis support (https://www.uhs.uga.edu/info/emergencies). • If you need help managing stress anxiety, relationships, etc., please visit BeWellUGA (https://www.uhs.uga.edu/bewelluga/bewelluga) for a list of FREE workshops, classes, mentoring, and health coaching led by licensed clinicians and health educators in the University Health Center. • Additional resources can be accessed through the UGA App.

FERPA Notice (e.g., The Federal Family Educational Rights and Privacy Act (FERPA) grants students certain information privacy rights. See the registrar's explanation at reg.uga.edu/ general-information/ferpa/. FERPA allows disclosure of directory information (name, address, telephone, email, date of birth, place of birth, major, activities, degrees, awards, prior schools), unless requested in a written letter to the registrar.)

Course Outline:

LECTURE. **INTERACTIVE WORKSHOPS**

PART 0: INTRODUCTION

Week 1: **Course Overview**

- The Problem
- Course Structure and Goals
- Introduction to Datasets, Data-Procedures and Data-Visualization

- **Homework:** Choose Dataset and take screenshots UMAP-app

PART 1: STATISTICAL FOUNDATIONS IN EXCEL

Week 2a: **Understanding Probability Distributions**

- Selecting the Bin Size: noise vs signal
- Selecting the scale: log vs linear
- Probability Intuition: z-scores vs P-values

- **eLC Homework**: PowerPoint, Excel File, Perusall Activity

Week 2b: **Probability Distribution Workshop**

Five Team Presentations
Muddiest Point Discussions

Week 3a: **Understanding Statistical Tests**

- Multiple Groups within Dataset
- Inferential Statistics vs Descriptive statistics
- Combining Evidences: Z-score vs p-value

- **eLC Homework**: PowerPoint, Excel File, Perusall Activity

Week 3b: **Statistical Tests Workshop**

Five Team Presentations
Muddiest Point Discussions

Week 4a: **Correlation Review: Scatter Plots and Linear regression**

Scatter Plots Review
Linear Regression Review
Predictive Statistics vs Inferential Statistics vs Descriptive Statistics
Machine Learning Concepts

- **eLC Homework**: PowerPoint, Excel File, Perusall Activity

Week 4b: **Linear Regression Workshop**

Five Team Presentations
Muddiest Point Discussions

Week 5a: **Visualizing Entire Matrices**

Dimensionality Reduction
Heatmaps

- **eLC Homework**: PowerPoint, Excel File, Perusall Activity

Week 5b: **Heatmaps/PCA/UMAP Workshop**

Five Team Presentations
Muddiest Point Discussions

PART 2: DATA SCIENCE IN R

Week 6a: **Introduction to R:** coding & distributions

Introduction to R
Syntax, variables, operations
Data Structures
Basic Functions in R
Debugging

- **eLC Homework**: Perusall Reading

Week 6b: **Distributions Analyses Team Activity**

Automating Week #2 homework
Empirical P-values: calculate from rank

Week 7a: **Statistical Tests in R**

- Visualizing multiple groups
- Calculating Statistical Tests
- Loops & False-Discovery Rate

- **eLC Homework**: PowerPoint, modified R-code, Perusall Activity

Week 7b: **Statistical Tests Workshop**

Five Team Presentations
Muddiest Point Discussions

Week 8a: **Linear Regression in R:**

Correlation Matrices:
Linear Regression
Multiple Linear Regression

- **eLC Homework**: PowerPoint, modified R-code, Perusall Activity

Week 8b: **Linear Regression Workshop**

Five Team Presentations
Muddiest Point Discussions

Week 9a: **Matrix Visualization in R:**

Dimensionality Reduction
Heatmaps
Clustering: distance vs correlation

- **eLC Homework**: PowerPoint, modified R-code, Perusall Activity

Week 9b: **Matrix Visualization Workshop**

Five Team Presentations
Muddiest Point Discussions

PART 3: BIOMEDICAL DATA SCIENCE

Week 10a: **Overview of -Omic Datasets:**

mRNA: 90% of phenotypes are epigentic
Dealing with large matrices (reducing size)
Pathway Analysis
Figures (scatter plots and heatmaps, clustering)

- **eLC Homework**: App-based Activity + PowerPoint

Week 10b: Human Cell Atlas Analyses Workshop

Individual Work
Discussion

Week 11a: **Introduction to bulk RNAseq**

Data Download, QC, Normalization
Simple Pathway analyses
Gene-set enrichement analysis
Deconvoluting bulk RNAseq

- **eLC Homework**: Perusall Discussion Bulk

Week 11b: **FALL BREAK: no class**

Week 12a: **Introduction to single-cell RNAseq**

Introduction to Seurat
Defining cell identities based on clusters
Pseudobulk methods to increase signal
Cell communication algorithms
Cell-transcriptional network algorithms

- **eLC Homework**: Perusall Discussion single-cell

Week 12b: **POH CONFERENCE: no class**

Week 13a: **Student-Sources scRNAseq Activity**

Source data
Application of SOP

Week 13b: **Continue Student Sourced scRNAseq Activity**

Week 14a: **Spatial Transcriptomics Overview**

Technology Types Overview
Overview of Standard Workflows

- **eLC Homework**: Videos Visium & CosMx workflows + R-code

Week 14b: **Workshop on CosMx / Visium Workflow**

Week 15a: **Student Sourced Spatial Transcriptomic Analyses (Final Project)**

Technology Types Overview

Week 15b: **Final Project Workshops**