Intended effective date
Spring 2021

Course number
BIOL 806 / CHEM 806

Abbreviated title (30 characters)
Data science

Expanded title (100 characters)
Exploratory data science for scientists

How will this course advance time to degree?
Major elective (In fact, grad program elective, but that language is not available on form)

Prereqs
CSC 306 or equivalent, graduate standing

Do you want to restrict enrollment by student level if noted as a prerequisite?
No

Do you want to restrict enrollment by major/minor/or other program if noted in the prerequisite?
No

Do you want to enforce any of the prerequisites noted above?
No (note, saying yes would mean students who haven't completed the prereq in SFSU's system could not enroll. Since many grad students will have completed an equivalent course at another university, we don't want to add an extra layer of bureaucracy)

Text

Course description
Introduction to data science for scientists; learn the fundamentals of data science through applications in biology and chemistry research; engage in data preparation, analysis, and reporting using real scientific datasets.

Grading basis
(+/-) Letter grade, CR/NC (OPT)

Course typically offered
Spring semester only

Components
Lecture: lecture and discussion (C2) 1 WTU and 1 contact hour per unit; 3 lecture units; no final exam/assignment

Activity: Arts and Sciences activity (C7) 1.3 WTU and 2 contact hour per unit; 1 activity unit; no final exam/assignment

Repeat for additional credit
No

GOLD PLOs

1) Students will develop computational skills needed to create, debug, and run a computer programs to perform data analyses.
2) Students will be able to obtain, store, manage and share data in a distributed environment through practical, hands-on experience with programming languages and big data tools.
3) Students will be able to evaluate data, as well as to apply key technologies data science analysis including statistical analysis, machine learning, and data visualizations.
4) Students will develop data science as an aspect of their professional identity, effectively communicate their data analyses and results, and will connect with professionals outside of the University to further their data science careers.
5) Students will apply their data science skills to discipline-specific data and questions to solve real-world problems of high complexity.

Biology Graduate PLOs

1) Critically read and evaluate the significance and validity of peer-reviewed publications to develop a comprehensive knowledge of research in their field of expertise and to be able to clearly articulate such knowledge.
2) Conduct original research in a biological sub-discipline, including the design of experiments, development and testing of hypotheses, application of quantitative analyses to visualize and interpret data and derive conclusions.
3) Develop effective writing skills for both informal and formal professional communications that include a written thesis, scientific proposal, or scientific manuscript.
4) Develop skills to orally present scientific material to a broad range of audiences, including in courses and an oral thesis defense.
5) Practice the responsible and ethical conduct of research and professional integrity in carrying out scientific investigation.

Chemistry Graduate PLOs

1) Demonstrate in-depth knowledge in a sub-discipline of chemistry
2) Organize and communicate scientific information clearly and concisely, both verbally and in writing
3) Demonstrate independence in analyzing data and interpreting results
4) Demonstrate an ability to engage in collaborative scientific activities in coursework
5) Use the scientific literature to develop and implement a research project.
6) Keep accurate records of experiments and data

SLOs

1) Students will create, manipulate, and use key computational data structures including arrays, multi-dimensional arrays, and dictionaries. test, in class projects, data frame - not dictionaries
    - GOLD PLO 1, Chem PLO 3, Bio PLO 2
2) Students will be able to map out and reflectively work within the data science cycle, identifying steps in the processes of data preparation, data analysis, and data reporting.
    - GOLD PLO 4, 5, Chem PLO 2, 3, Bio PLO 2 in class projects
3) Students will learn robust habits/practices for organizing the elements of a data analysis project from the file-structure perspective. in class projects
    - GOLD PLO 2, 5, Chem PLO 3, Bio PLO 2, 5
4) Students will be able to prepare data sets, including acquiring data across computational systems and converting between data formats, for different types of data including tabular data, relational data, and text-based data. test, in class projects
    - GOLD PLO 2, 5, Chem PLO 3, Bio PLO 2
5) Students will be able to collaboratively perform computational data analyses including statistical tests, visualizations of data and results, simulations, principal components analysis, and machine learning. in class projects, self reflection, test
    - GOLD PLO 3, 5, Chem PLO 3, 4, Bio PLO 2
6) Students will be able to collaboratively report on data analyses, including accurate descriptions of data sets and analyses, publication-quality plots, and interactive data analysis visualizers. in class projects, self reflection
    - GOLD PLO 4, Chem PLO 2, 4, Bio PLO 3

Course topics

Data structures

- Arrays
- Multi-dimensional arrays
- Dictionaries
- Graphs (or another relational data structure)

Data science project management

- Data science cycle framework to think of data analyses throughout class
- Command line interface
- Files system navigation
- File system management
- Version control techniques

Data preparation

- Acquiring data remotely
- Data formats
- Wrangling, reshaping, and tidying data
- Different types of data, for example
    - Tabular data
    - Relational data
    - Text-based data
    - Image data
    - Qualitative data

Data analysis

- Summary statistics (eg: mean, variance)
- Descriptive plots (eg: scatterplot, histogram), emphasis on making choices about how to visualize data
- Statisical tests (eg: t-test, linear modeling/regression, correlation, ANOVA)
- Data clustering (eg: PCA, hierarchical models)
- Machine learning
- Bayesian statistics

Data reporting

- Data description
- Results description
- Publication-quality plots
- Interactive shiny apps
- Writing a paper
- Communicating via social media

How will final grades for the course be determined?

Grades will be based on a series of weekly quizzes to assess technical mastery (25%), several team-based applied data science projects and the weekly deliverables associated with those projects (60%), and participation (15%).

List of textbooks/reading assignments

As adherence with the NSF grant funded to support the development of this course, we will develop a set of freely available online course materials for use in this class.