

**2016-2017 Yearly Assessment Report Template
College of Arts and Sciences (CAS)**

1. Identifying Information

Name of Program: [Data Science](#)

Type of Program: [Major](#)

College of Arts and Sciences Division: [Sciences](#)

Name/Title/Email Address of Submitter:

Nathaniel Stevens
Assistant Professor of Statistics
ntstevens@usfca.edu

James Wilson
Assistant Professor of Statistics
jdwilson4@usfca.edu

Name/Email Address of Additional Individuals Who Should Receive Feedback: [Nathaniel Stevens](#), [James Wilson](#) and [Steve Devlin \(smdevlin@usfca.edu\)](#)

2. Mission Statement:

To deliver a high-quality data science program that instructs students in the theory and practice of mathematical and computation analysis of applied data driven problems, and to graduate students with appropriate experience in industry-standard data science tools.

Has this statement been revised in the last few years?

Not in the last academic year.

3. (Optional) Program Goals:

Have these goals been revised in the last few years?

4. Program Learning Outcomes (PLOs)

- L1. Think logically and analyze information critically in a mathematical setting.
- L2. Reformulate and solve problems in an abstract framework.
- L3. Express mathematical results verbally, working individually and in collaborative groups.
- L4. Apply mathematical techniques to specific problem domains
- L5. Demonstrate competence with programming concepts, including software development techniques and data structures
- L6. Apply mathematical and computational techniques to real-world problems involving large, complex data sets.
- L7. Visualize, present and communicate analytical results.

Have these PLOs been revised in the last few years?

Not in the last academic year.

5. Brief Summary of Most Recent Assessment Plan

We assessed L1, L4, L5, L6 via an end-of-degree exit exam. This was given to all graduating seniors in the Spring 2017 semester. This exam consisted of 14 multiple choice questions spanning topics from the required curriculum.

6. Academic Program Review

Date of most recent Academic Program Review's External Reviewer Visit:

Math/BSDS APR: 2017

Date of most recent Action Plan Meeting:

Spring 2016

Brief Summary of the most recent Action Plan:

We revised and modernized the BSDS curriculum which has taken effect as of Fall 2017.

7. Methods

What did you do with regard to assessment of your program/department in 2016-2017?

The assessment exam is attached as a separate document. Overall, we were pleased with the performance of our graduating seniors though we believe there is still room for improvement. As a result, we intend to highlight key points and important topics in our BSDS required classes. In the future, we plan to give this exam to every graduating student, allowing us to assess their improvement on a year-over-year basis.

What were your questions?

See attached assessment exam (titled BSDSexit.pdf)

How are these questions related to your most recent Academic Program Review and/or Action Plan?

They are unrelated, as our most recent APR happened after the development of this exam.

What PLOs are these questions related to?

L1, L4, L5, L6

What direct (most important) and/or indirect methods did you employ?

Direct: b, c, f, g, j, l, o, p;

Indirect: none

Some Possible Direct Methods (pick ≥ 1 and *briefly describe*):

- a. Published (Standardized) Test (e.g., Major Field Test)
- b. Class Tests & Quizzes with Embedded Questions
- c. Class Presentations
- d. Off-Campus Presentations (NGOs, clients, agencies, etc.)
- e. Research Projects Reports
- f. Case Studies
- g. Term Papers
- h. Portfolio
- i. Artistic Performances, Recitals & Products
- j. Capstone Projects
- k. Poster Presentations
- l. Comprehensive Exams
- m. Thesis, Dissertation
- n. Pass Rates on Certification or Licensure Exams
- o. Group Projects
- p. In/Out-of Class Presentations
- q. Competency Interviews (e.g., oral exams)
- r. Simulations
- s. Juried Presentations
- t. Other

Some Possible Indirect Methods (*briefly describe*):

- a. Student Survey
- b. Student Interview
- c. Focus Groups
- d. Reflection Sessions
- e. Reflection Essays
- f. Faculty Survey
- g. Exit (end of program) Survey
- h. Exit (end of program) Interview
- i. Alumni Survey
- j. Employer Survey
- k. Diaries or Journals
- l. Data from Institutional Surveys (e.g., NSSE, SSI, GSS)
- m. Curriculum/Syllabus Analysis
- n. Other

8. Results

What were the direct data results?

Too early to make any definitive conclusions. More data needed to provide an accurate/satisfactory response.

What were the indirect results?

n/a

What surprised you?

n/a

What aligned with your expectations?

More data needed to provide an accurate/satisfactory response.

What do you understand these results to mean?

We're pleased with this first round of results, and we endeavor to improve the program on an ongoing basis.

What are the implications of the data?

The assessment results suggest that our students come away from the program with a solid comprehension of data science and its intricacies. That said, we hope to strengthen this level of understanding as the program continues to evolve.

9. Closing the Loop

What might you do as a result of these assessment results? What curricular or programmatic changes might you implement?

This is the first year for the revised curriculum and so it is unclear at this point. In the future we hope to incorporate student internships with Bay Area companies to provide more hands-on experience with data science.

Possible Closing(s) of the Loop(s) (pick ≥ 1 and *briefly describe*):

In line with our desire to engage students in internships, this corresponds to a change in "program modality of delivery" (see items f. and p. below). In order to develop and maintain this new modality of instruction, more faculty are required for outreach and supervision (see item o.).

- a. Revision of PLOs
- b. Changes in pedagogical practices
- c. Revision of program course sequence
- d. Revision of course(s) content
- e. Curriculum Changes (e.g., addition and/or deletion of courses)

- f. Modified program policies or procedures
- g. Designed measurement tools more aptly suited for the task
- h. Improved within and across school/college collaboration
- i. Improved within and across school/college communication
- j. Revised student learning outcomes in one or more courses
- k. Modified rubric
- l. Developed new rubric
- m. Developed more stringent measures (key assessments)
- n. Modified course offering schedules
- o. Changes to faculty and/or staff
- p. Changes in program modality of delivery
- q. Other

Have you or will you submit any course or program change proposals as a result of these results?

No.

BSDS EXIT EXAM Instructions:

Don't consult with any outside materials (e.g., notes, web sites, R documentation, etc.) or individuals as you work through this examination. Feel free to use scratch paper and a calculator.

1. Suppose $P(A) = 0.60$, $P(B) = 0.47$ and $P(A \cap B) = 0.19$. Calculate $P(A|B^C)$.

- (a) 0.36
- (b) 0.41
- (c) 0.68
- (d) 0.77
- (e) 0.87

2. Suppose a store has 100 light bulbs in stock. Assume 40 light bulbs are from Distributor A and the remainder of the light bulbs are from Distributor B. Assume 5.0% of the light bulbs from Distributor A are defective and 10.0% are defective from Distributor B.

If a consumer purchases 3 light bulbs, what is the probability that exactly 2 of the light bulbs are defective? Choose the correct expression.

- (a) $\binom{2}{3}0.08^3(1 - 0.08)^2$
- (b) $0.08^2(1 - 0.08)$
- (c) $\binom{3}{2}0.08^2(1 - 0.08)$
- (d) 0.08
- (e) $0.08(1 - 0.08)$

3. Suppose that you know that exactly 10% of all emails sent to you are spam. Moreover, you know that 80% of the emails that you have received in the past that were not spam contained the word "analytics" and that 40% of the emails that you received that were spam contained that word. Suppose that you receive a new email that contains the word "analytics." What is the probability that the email is spam?

- (a) 0.40
- (b) 0.10
- (c) 0.053
- (d) 0.056

4. Consider the following Python code:

```
def f(x):  
    ind=0  
    val=x[0]  
    for i in range(len(x)):  
        if x[i]>val:  
            val = x[i]  
            ind = i  
    return ind  
l=[1,22,13,194,5,-4,0]  
f(l)
```

The result of running the above code is:

- (a) 3
- (b) 4
- (c) 194
- (d) 0

5. Consider the following R code:

```
mat <- matrix(rnorm(200000,50,7),10000,20)  
rms <- rowMeans(mat)  
par(mfrow=c(2,1))  
hist(rnorm(10000,50,7), breaks=50,xlim=c(30,70))  
hist(rms, breaks=50,xlim=c(30,70))
```

The result of running the above code is best described as illustrating:

- (a) The standard normal distribution
- (b) A confidence interval for a population mean
- (c) The central limit theorem
- (d) A chi-square distribution

6. A researcher has a dataset consisting of, among other things, the annual income for a large sample of individuals in the US. If the researcher were to make a histogram of this income data, it is likely that the distribution would be:

- (a) Exponential
- (b) Symmetric
- (c) Skewed to the left
- (d) Skewed to the right

7. A researcher studying a sample of U.S. cities plots, for each city, the number of churches per thousand people on the x -axis and the number of violent crimes per thousand people on the y -axis. The researcher notices that the points in the scatterplot could be reasonably described as moving up and to the right in a roughly linear fashion. Which of the following statements is most correct based on this information?

- (a) Increasing the number of churches in a city will increase the crime rate.
- (b) $r > 0$ where r is the correlation coefficient.
- (c) The number of churches and the crime rate are independent.
- (d) $\beta_1 = 0$ where β_1 is the slope of the population regression line.

8. A clinician runs a double blind experiment to assess the effectiveness of a potential new drug for migraine headaches. She records the mean improvement in migraine severity in both a treatment group and an independent placebo group, and uses them to run a t -test. If she obtains a p -value of 0.03, and assuming that this test is appropriate, which conclusion is correct?

- (a) The experiment shows that the drug reduces migraine severity by 3%.
- (b) 97% of people in the treatment group improved.
- (c) There is a 3% chance that the drug is ineffective.
- (d) If the drug were ineffective, the researcher would expect data like her's 3% of the time.

9. Which technique could plausibly be used by a marketer to determine the probability that a customer will make a purchase based on several pieces of demographic information?

- (a) Analysis of variance
- (b) Simple linear regression
- (c) Logistic regression
- (d) Chi-square test

10. Suppose $\mathbf{X} = [X_1, X_2, \dots, X_n]^T$ is a vector where each X_i is a random variable with mean $E[X_i]$ and finite variance σ_i^2 . Define Σ to be the matrix $E[(\mathbf{X} - E[\mathbf{X}])(\mathbf{X} - E[\mathbf{X}])^T]$. The (i, j) entry of the matrix Σ is:

- (a) The joint density of X_i and X_j .
- (b) The conditional expectation of X_i given X_j .
- (c) 0
- (d) The covariance between X_i and X_j .

11. Which of the following is an eigenvector for the matrix

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & -2 & 4 \\ 5 & 0 & -2 \end{bmatrix}$$

- (a) 3

- (b) -2
- (c) 2
- (d) 0

12. Suppose X is an $n \times k$ matrix with linearly independent columns, and \mathbf{b} is a vector in \mathbb{R}^n that is not in the column space of X . The best approximation to \mathbf{b} in the column space of X is given by:

- (a) $X^{-1}\mathbf{b}$
- (b) $X(X^T X)^{-1}X^T\mathbf{b}$
- (c) Does not necessarily exist
- (d) $(X^T X)^{-1}X^T\mathbf{b}$

13. Let Y_1, Y_2, \dots, Y_n be a random sample of observations from a uniform distribution with probability density function given, for each $i = 1, 2, \dots, n$, by $f(y_i) = \frac{1}{\theta}$, $0 \leq y_i \leq \theta$. The maximum-likelihood estimator of θ is:

- (a) $\frac{1}{n} \sum_{i=1}^n Y_i$
- (b) $\min\{Y_1, Y_2, \dots, Y_n\}$
- (c) $\max\{Y_1, Y_2, \dots, Y_n\}$
- (d) 1

14. Consider a classification problem involving only two classes $\{0, 1\}$ using a p dimensional predictor X . The Bayes classifier sets:

$$P(Y = 1|X = x) = \frac{P(Y = 1)P(X = x|Y = 1)}{P(Y = 0)P(X = x|Y = 0) + P(Y = 1)P(X = x|Y = 1)}.$$

Suppose we take three new observations x_a, x_b , and x_c , and find that $P(Y = 1|X = x_a) = 0.55$, $P(Y = 1|X = x_b) = 0.87$, $P(Y = 1|X = x_c) = 0.32$. How should these three observations (x_a, x_b, x_c) be classified?

- (a) (1, 1, 1)
- (b) (1, 1, 0)
- (c) (1, 0, 0)
- (d) (0, 1, 1)
- (e) (0, 0, 1)

BSDS Course Checklist

Data Courses

BSDS 100 Introduction to Data Science with R (4)
--

Math and Stats Courses

Math 109 Calculus and Analytic Geometry I (4)

M 110 Calculus and Analytic geometry II (4)

M 230 Elementary Linear Algebra (4)

Math 201 Discrete Mathematics or Math 235 Introduction to Formal Methods (4)
--

Math 370 Probability with Applications (4)
--

Math 371 Statistics with Applications (4)

Math 372 Linear Regression (4)

Math 373 Statistical Learning (4)

CS Courses

CS 110 Intro to Programming I (4)

CS 112 Intro to Programming II (4)

CS 245 Data Structures and Algorithms (4)

CS 333 Databases (4)

Capstone / Technical Elective(s)

Math 394 or CS 490* or Upper division Math&Stats elective or Upper division CS elective (4)
--

Math&Stats Courses: minimum 8 courses, 32 units.

Data Courses: minimum 1 course, 4 units

CS Courses: minimum 4 courses, 16 units.

Capstone/Tech Elective: 1 course, 4 units (*requires CS212).

Electives: 5 Courses, 20 units.

Core(40)+Language(8): 12 courses, 48 units.

TOTAL: 32+4+16+4+20+48 = 124.

	PLO1	PLO2	PLO3	PLO4	PLO5	PLO6	PLO7
Institutional Learning Outcomes X Program Learning Outcomes	Think logically and analyze information critically in a mathematical setting.	Reformulate and solve problems in an abstract framework.	Express mathematical results verbally, working individually and in collaborative groups.	Apply mathematical techniques to specific problem domains.	Demonstrate competence with programming concepts, including software development techniques and data structures.	Apply mathematical and computational techniques to real-world problems involving large, complex data sets.	Visualize, present and communicate analytical results.
Institutional Learning Outcomes							
1. Students reflect on and analyze their attitudes, beliefs, values, and assumptions about diverse communities and cultures and contribute to the common good.							
2. Students explain and apply disciplinary concepts, practices, and ethics of their chosen academic discipline in diverse communities.							
3. Students construct, interpret, analyze, and evaluate information and ideas derived from a multitude of sources.							
4. Students communicate effectively in written and oral forms to interact within their personal and professional communities.							
5. Students use technology to access and communicate information in their personal and professional lives.							
6. Students use multiple methods of inquiry and research processes to answer questions and solve problems.							
7. Students describe, analyze, and evaluate global interconnectedness in social, economic, environmental and political systems that shape diverse groups within the San Francisco Bay Area and the world.							

Key:
I = Introductory
D = Developing
M = Mastery

		PLO1	PLO2	PLO3	PLO4	PLO5	PLO6	PLO7
Program Learning Outcomes X Courses		Think logically and analyze information critically in a mathematical setting.	Reformulate and solve problems in an abstract framework.	Express mathematical results verbally, working individually and in collaborative groups.	Apply mathematical techniques to specific problem domains.	Demonstrate competence with programming concepts, including software development techniques and data structures.	Apply mathematical and computational techniques to real-world problems involving large, complex data sets.	Visualize, present and communicate analytical results.
Required Courses								
BSDS 100	Introduction to Data Science with R	I		D	D	D	D	D
CS 110	Introductory Computer Science I	I	I			I	I	
CS 112	Introductory Computer Science II	D	D			D	D	
Math 109	Calculus I	I	I		I		I	
Math 110	Calculus II	D	D		D		D	
Math 201	Discrete Mathematics	D	D		I			
Math 230	Linear Algebra	D	D		D			
Math 235	Formal Methods	M	M	D	I			D
CS 245	Data Structures and Algorithms	D				D	D	D
Math 370	Probability with Applications	D	D	D	D			
Math 371	Statistics with Applications	D	D	D	D	I	I	I
Math 372	Linear Regression	M	M		M	M	M	M
Math 373	Statistical Learning	M	M		M	M	M	M
CS 333	Databases					M	M	
CS 490/ Math 395	Capstone	M	M	M	M	M	M	M

Key:
I = Introductory
D = Developing
M = Mastery