

Term Information

Effective Term Spring 2018

General Information

Course Bulletin Listing/Subject Area Political Science
Fiscal Unit/Academic Org Political Science - D0755
College/Academic Group Arts and Sciences
Level/Career Undergraduate
Course Number/Catalog 3785
Course Title Data Science
Transcript Abbreviation Data Science
Course Description Introduction to new issues in data science, including big data, machine learning, network interdependencies, and automated extraction of meaning from widely varied sources of data.
Semester Credit Hours/Units Fixed: 3

Offering Information

Length Of Course 14 Week, 12 Week, 8 Week, 7 Week, 6 Week, 4 Week
Flexibly Scheduled Course Never
Does any section of this course have a distance education component? No
Grading Basis Letter Grade
Repeatable No
Course Components Lecture
Grade Roster Component Lecture
Credit Available by Exam No
Admission Condition Course No
Off Campus Never
Campus of Offering Columbus

Prerequisites and Exclusions

Prerequisites/Corequisites
Exclusions
Electronically Enforced No

Cross-Listings

Cross-Listings

Subject/CIP Code

Subject/CIP Code 45.1001
Subsidy Level Baccalaureate Course
Intended Rank Sophomore, Junior, Senior

Requirement/Elective Designation

The course is an elective (for this or other units) or is a service course for other units

Course Details

Course goals or learning objectives/outcomes

- Students become familiar with field of Data Science
- Students understand analytic tools used in Data Science
- Students learn to program in the R language

Content Topic List

- Difference Between Data and Big Data
- Scientific Method, Data, and Measurement
- Populations, Samples, and Sources of Bias
- Automated Scraping of Data
- Data Visualization
- To Explain or Predict?
- Statistical Inference
- Basics of Machine Language
- Machine Language Classification and K-means
- Recommender Systems
- Structure in Social Networks
- Visualizing Networks
- Basics of Automated Text Analysis
- Extracting Meaning from Text
- Genetics and Human Behavior
- Ethical Issues
- The Future of Data Science

Sought Concurrence

Yes

Attachments

- PolitSc_3785_Syllabus.pdf: PS 3785 Syllabus
(Syllabus. Owner: Smith, Charles William)
- List of Concurrence Requests.docx: list of concurrence requests
(Concurrence. Owner: Smith, Charles William)
- Curriculum Maps.xlsx: curriculum maps
(Other Supporting Documentation. Owner: Smith, Charles William)

Comments

- add maps *(by Herrmann, Richard Karl on 10/19/2017 02:37 PM)*

COURSE REQUEST
3785 - Status: PENDING

Last Updated: Haddad,Deborah Moore
11/28/2017

Workflow Information

Status	User(s)	Date/Time	Step
Submitted	Smith,Charles William	10/19/2017 02:26 PM	Submitted for Approval
Revision Requested	Herrmann,Richard Karl	10/19/2017 02:37 PM	Unit Approval
Submitted	Smith,Charles William	10/20/2017 01:11 PM	Submitted for Approval
Approved	Herrmann,Richard Karl	11/28/2017 05:49 PM	Unit Approval
Approved	Haddad,Deborah Moore	11/28/2017 06:20 PM	College Approval
Pending Approval	Nolen,Dawn Vankeerbergen,Bernadette Chantal Oldroyd,Shelby Quinn Hanlin,Deborah Kay Jenkins,Mary Ellen Bigler	11/28/2017 06:20 PM	ASCCAO Approval

POLITICAL SCIENCE 3785: DATA SCIENCE

Professor Skyler Cranmer 2032 Derby Hall 614-292-4529 cranmer.12@osu.edu	Class: 1245-1405 Tu, Th TBD	Office Hours: 0900-1100 W or by appointment
--	---	--

COURSE DESCRIPTION

This course is an introduction to what many consider the *second* quantitative revolution in the social sciences: that focused on big data, machine learning, network interdependencies, and automated extraction of meaning from widely varied sources of data. The primary focus of the course is on these new issues in data science, rather than traditional statistics. We will introduce and discuss some statistical concepts, but a background in statistics is not necessary to take this course. The goal of the course is to provide you with a broad survey of the field of Data Science and introduce you to the analytical tools used in this field. This is a good first course for those considering specialization in quantitative social science (e.g. economics, political science, sociology, communication, psychology) or data analytics.

This is a quantitatively oriented course and one geared very heavily towards programming in the R language. While prerequisites in mathematics, statistics, and programming are not required, we will make extensive use of all three of those skills, introducing the material we need as we go along. A background in math, statistics, and/or programming can only be helpful to the student and students without much background in these disciplines should expect to put in substantial work on developing these skills over the course of the semester. The course is designed to be reading light and coding heavy: you will learn by doing and you will be good at coding in R by the time you finish the semester.

This course will also take advantage of the web and modern computing as an instructional and research technology. Some of the readings are accessible only via the web, and students are required to use information technology tools in their assignments. For example, you will have DataCamp exercises in advance of every lecture. Also, if you are not familiar with web tools such as JSTOR, you need to become familiar with this technology as soon as possible.

COURSE MATERIALS REQUIRED FOR PURCHASE

1. Membership to DataCamp: \$29 per month. <https://www.datacamp.com/pricing>
2. King, Keohane, and Verba, 1994 *Designing Social Inquiry*, Princeton Press (note: someone put a PDF online for free... google it)

COURSE ASSESSMENT

Homework	35%
Hackathon Project	20%
Hackathon Presentation	5%
Exam 1:	20%
Exam 2:	20%

Failure to turn in an assignment on its due date without prior contact with the instructor will result in a zero for that portion of the grade. However, you are still required to turn in and proceed with the rest of the assignment.

You will be graded not only for the depth of your ideas, reasoning, and evidence, but also for your presentation (this includes the basics: visual clarity, spelling, punctuation, organization, grammar, but it also includes the effective use of technology to enhance the report and improve on your ability to communicate your ideas).

Laptops, phones, and other such electronics are not allowed except as directly related to class activities.

Office hours: Office hours are your time for one-on-one dialogue about the issues we are discussing in class. I encourage you, in the strongest possible terms, to make full and regular use of them.

HOMEWORK

Students are required to complete two code-based homework assignments per week using DataCamp. This is to develop your R coding skills, and thus your ability to implement the techniques and methods we discuss theoretically in the lectures. The purpose of these homework assignments is to give you practical, hands-on, experience with the concepts and tools we are discussing. Due dates and homework topics are given in the calendar below. To turn in an assignment, simply screenshot the “achievement” screen you get when you complete a chapter, and submit that image to the appropriate Carmen dropbox.

HACKATHON PROJECT

In teams of 2-4 people, you will prepare an info-graphic style report on a subject of interest to you. You have very wide leeway in your choice of topic, but **you must clear the topic with the instructional staff (Professor or TA) by the date indicated on the calendar below** so that we know what you are working on and can offer advice. You will also be required to present your final informational product to the class. This will be worth 5% of your course grade (while the product itself will be worth 20%).

To complete this project you will need to use and integrate many of the techniques and skillsets we develop over the course of the semester. You will also need to conduct library research. Library sources include relevant books, professional journal articles, policy magazines, and original government documents. See the Reference desk in the Library for assistance. A bibliography must be included with the infographic.

It is also highly likely that you will want/need to improve certain aspects of your coding skills by doing some of the DataCamp tutorials that are not assigned in the calendar below. Plan your time accordingly as you work on your projects.

Specific instructions will be given in class and on Carmen as the semester progresses.

Note: it is expected that this project represents original work not previously or simultaneously handed in for credit in another course, unless this is done with the prior approval of all professors involved.

EXAMS

Both the exams will be in-class, closed-book, closed-notes, exams taken on your laptop or tablet via Carmen's Browser LockDown feature. The exams may include multiple choice, true/false, short answer, and short essay questions, or any mix thereof. We will discuss the format of the exams and how to study for them in class prior to the examination period. **Missing either exam will result in a score of zero.** Makeup exams will not be provided without proper documentation, and if at all possible the student must notify the instructor of the conflict *before* the scheduled exam takes place.

Note: We will employ the *Browser LockDown* feature in Carmen both exams. This means that you will be taking your exams on your laptops. If your laptop battery does not hold a charge for 75 minutes, buy a new battery or bring an extension cord/surge protector to class on exam day. If you do not have a laptop that runs a currently supported operating system, you may use blue books instead, but please let the instructional staff know of your intention to do so before exam day.

The Honor Code is in effect in this class and all others at the University. I am committed to treating Honor Code violations seriously and urge all students to become familiar with its terms set out at . If you have questions, it is your responsibility to ask me about the Code's application. **All exams, written work, and other projects must be submitted with a statement that you have complied with the requirements of the Honor Code in all aspects of the submitted work.**

GRADING

In general, I subscribe to the official grading rubric provided by the university:

A - Excellent performance	D - Low but acceptable performance
B - Above average performance	E - Course objectives not satisfied
C - Average performance	

Thus, an assignment or exam that is essentially correct but is neither strong nor extraordinary will receive a grade in the range of a C. For B grades I require assignments and exams to be comprehensively satisfactory, not just essentially correct, and I reserve A grades for *truly outstanding* work that goes beyond the requirements.¹ Substantive challenges to grades on projects and the exams may be submitted to me in writing, with the provision that I reserve the right to regrade an entire assignment or exam. The grade received on any regraded assignments or exams will be final regardless of whether that grade is higher or lower than the original grade.

¹Note that a grade ranging between 90%-93% is an A-, with a full A reserved for 94%-100%. 80-83%=B-, 84-87%=B, 87-89%=B+; 70-74%=C-, 74-77%=C, 77-79%=C+; 67%-69%=D+; 64%-66%=D; 60%-63%=D-; 0%-59%=E

THIS SHOULDN'T HAVE TO BE IN A SYLLABUS, BUT...

You are an adult. This is not high school. It is your responsibility to communicate your performance to your parents. Due to FERPA regulations, I cannot discuss your performance with your parents. Accordingly, I do not communicate with your parents, in any way or under any circumstances. If your parent contacts me, I will not reply to them but will advise you of this so that you can communicate my non-communication policy to them.

ACADEMIC MISCONDUCT

It is the responsibility of the Committee on Academic Misconduct to investigate or establish procedures for the investigation of all reported cases of student academic misconduct. The term “academic misconduct” includes all forms of student academic misconduct wherever committed; illustrated by, but not limited to, cases of plagiarism and dishonest practices in connection with examinations. Instructors shall report all instances of alleged academic misconduct to the committee (Faculty Rule 3335-5-487). For additional information, see the Code of Student Conduct <http://studentlife.osu.edu/csc/>.

STUDENTS WITH DISABILITIES

Students with disabilities (including mental health, chronic or temporary medical conditions) that have been certified by the Office of Student Life Disability Services will be appropriately accommodated and should inform the instructor as soon as possible of their needs. The Office of Student Life Disability Services is located in 098 Baker Hall, 113 W. 12th Avenue; telephone 614-292-3307, slds@osu.edu; slds.osu.edu

COURSE CALENDAR

- 1/09 *Introduction to the Course* Read before class: This Syllabus
- 1/11 *So much data... digital footprints and the difference between data and big data*
Read before class: D. Lazer and J. Radford. “Data ex Machina: Introduction to Big Data” Annual Review of Sociology Vol. 43:- (Volume publication date August 2017) <https://doi.org/10.1146/annurev-soc-060116-053457>
Homework: Install RStudio (free) and do chapter 1 and 2 of Datacamp’s free R tutorial: <https://www.datacamp.com/courses/free-introduction-to-r>
- 1/16 *Fundamentals of design part 1: Scientific method / Data / measurement*
Read before class: KKV Ch 3
Homework: Datacamp chapter 3 and 4
- 1/18 *Fundamentals of design part 2: Populations, Samples, and Sources of Bias*
Read before class: KKV Ch 4-5
Homework: Datacamp chapter 4, 5, and 6
- 1/23 *Sources of data part 1: databases and repositories*
Read before class: Instructor provided notes on types of databases.
Homework: Importing Data in R (Part 1), chapter 1 and 2

- 1/25 *Sources of data part 2: automated scraping of data*
 Read before class + Homework: Extracting tables (1 and 2) as well as Making a Scraper from <https://www.outwit.com/support/help/tutorials/> (Note: we won't use it in this class, but if you want to get good at automated data gathering, there is no substitute for learning Python. DataCamp has an extensive Python program to which you have access through your monthly membership.)
- 1/30 *The power of visualization*
 Read before class: Data Visualization 101: https://cdn2.hubspot.net/hub/53/file-863940581-pdf/Data_Visualization_101_How_to_Design_Charts_and_Graphs.pdf
 Homework: Data Visualization in R chapters 1 and 2 (note: if you think you will want to do extensive/serious visualizations for your hackathon project, I highly recommend the three part series DataCamp has on using ggplot... way more powerful graphics.)
- 2/1 *To Explain or Predict?*
 Read before class: S. Cranmer and B. Desmarais. "What we can Learn from Predictive Modeling" *Political Analysis* 2017 (find through Library)
 Homework: Intermediate R chapters 1 and 2
- 2/6 *Statistical Inference*
 Read before class: Instructor notes on statistical inference
 Homework: Correlation and Regression (all chapters: note, this is a pretty long homework, allocate extra time)
- 2/8 *Machine Learning part 1, basics*
 Read before class: An Introduction to Machine Learning Theory and Its Applications: <https://www.toptal.com/machine-learning/machine-learning-theory-an-introduction>
 Homework: Introduction to Machine Learning chapters 1 and 2
- 2/13 *Machine Learning part 2, classification and k-means*
 Read before class: Introduction to K-Means Clustering <https://www.datascience.com/blog/introduction-to-k-means-clustering-algorithm-learn-data-science-tutorial>
 Homework: Introduction to Machine Learning chapters 3, 4, and 5
- 2/15 *Recommender Systems*
 Read before class + Homework: Recommender Systems 101 <http://bigdata-doctor.com/recommender-systems-101-practical-example-in-r/>. Optional: <http://blog.yhat.com/posts/recommender-system-in-r.html>
- 2/20 *Structure in social networks part 1 (centrality, structure case study: Google's pagerank)*
 Read before class: Borgatti, S. P. 2005. "Centrality and network flow" *Social Networks* p55-71 <http://www.sciencedirect.com/science/article/pii/S0378873304000693>
 Homework: Instructor provided R-Lab
- 2/22 *Visualizing Networks*
 Read before class: None
 Homework: Work through Katya Ognyanova's tutorial <http://kateto.net/network-visualization>
- 2/27 *Everything (well... a lot) is contagious through networks!*
 Read before class: N. Christakis and J. Fowler "The Spread of Obesity in a Large

Social Network Over 32 Years” New England Journal of Medicine 357 (4): 370379
(26 July 2007) http://fowler.ucsd.edu/spread_of_obesity.pdf

Homework: Instructor provided R-Lab

3/1 *Exam review*

3/6 **Exam 1**

3/8 *Topic TBD*

Readings and homework TBD

3/13 *Spring Break Class Canceled*

No readings or homework

3/15 *Spring Break Class Canceled*

No readings or homework

3/20 *Text as data part 1: Basics of automated text analysis*

Read before class: Gary King, Jennifer Pan, and Margaret E. Roberts. 2014. “Reverse-engineering censorship in China: Randomized experimentation and participant observation.” *Science*, 6199, 345: 1-10

Homework: Text Mining: Bag of Words chapters 1 and 2 (3 and 4 optional)

3/22 *Text as data part 2: extracting meaning from text*

Read before class: Brice Acree, Justin H. Gross, Amber Boydston, Yanchuan Sim, & Noah A. Smith. “Etch-a-Sketching: Testing the Post-Primary Moderation Hypothesis” *American Journal of Political Science*

Homework: Instructor provided R tutorial

3/27 *Genetics and human behavior*

Read before class: JH Fowler, LA Baker, CT Dawes. “Genetic variation in political participation” *American Political Science Review* 102 (02), 233-248 http://fowler.ucsd.edu/genetic_basis_of_political_cooperation.pdf

Homework: work on your Hackathon projects

3/29 *Ethical Issues: privacy, security, and ethics*

Read before class: H Simo “Big Data: Opportunities and Privacy Challenges” <https://arxiv.org/pdf/1502.00823.pdf>

Homework: work on your Hackathon projects

4/3 *The Future of Data Science*

Read before class: “The Future of Data Science: QA with MIT Professional Educations Devavrat Shah”

Homework: work on your Hackathon projects

4/5 *Hackathon Presentations*

4/10 *Hackathon Presentations*

4/12 *Hackathon Presentations*

Hackathon Infographics Due

4/17 *Exam Review*

4/19 *Exam 2*

New Course Request

Political Science 3785 – Data Science

Concurrence Requests Sought from:

Statistics—request sent 6/28/17 to Dr. Steven MacEachern (snm@state.osu); no reply

Computer Science and Engineering—request sent 6/28/17 to Dr. Xiaodong Zhang.574; no reply