

Course Syllabus

Course Information

BUAN 6340; Programming for Data Science; Spring 2023

Tuesday 4-7pm	Section s01	Cohort
Thursday 4-7pm	Section 002	Flex
Friday 4-7pm	Section 003	Flex

Professor Contact Information

Jason Parker

Office:	JSOM 3.807
Office number:	(972) 883-5141
Online Office hours:	Tues 10am-12pm Thurs 10am-12pm

To join the online office hours, simply go to Blackboard Collaborate on the elearning Platform and click “Join Course Room”. From there you can talk with me and share your screen so that I can help you to master the material.

Course Pre-requisites, Co-requisites, and/or Other Restrictions

Prerequisites: BUAN 6356 or MIS 6323 or MIS 6334 or MIS 6356 or MIS 6382.

Course Description

BUAN 6340 - Programming for Data Science (3 semester credit hours) This course covers many aspects of programming for data science and analytics, including syntax, handling data, data visualization, and implementation of statistical analysis models. The course will be taught using Python language and may use a different programming language as applicable. Prerequisites: BUAN 6356 or MIS 6323. (3-0) Y

Student Learning Objectives/Outcomes

- To apply core Python structure as well as object-oriented design in Python to build own solutions to free-form programming challenges in the problem sets.
 - To use and fully understand core database and linear algebra packages in Python for the analysis of data.
-

Required Textbooks and Materials

Access to a computer which can run the Python language interpreter. Ideally, this would be the student's own machine which they could bring into class, although that is not necessary.

For this course, we will be using the Anaconda Python v 3.9 which can be freely downloaded from: <https://www.anaconda.com/download/>

Books on programming

These books are available through O'Reilly Safari through the UTDallas library.

VanderPlas, J. (2016). Python data science handbook: essential tools for working with data. " O'Reilly Media, Inc."

This is the main book for the course. It has in-depth information about the main data science packages in Python. You must read this book to pass this course. That doesn't mean you have to memorize it. But occasionally I will give you a challenge that requires you to find information out for yourself and this book is an excellent place to start looking.

Kapil, S. (2019). Clean Python: Elegant Coding in Python.

I love this book. There are many resources to teach how to get code working in python, but not a lot of books on proper code structure and rules. This book is a must read for anyone who is going to be developing with a team. Good structure is helpful for your team mates, and that's what this book teaches. I have put this second on the list for a reason.

Martelli, A. (2006). Python in a Nutshell. " O'Reilly Media, Inc."

This is a desk reference of python features. I first learned python from this book, and it still has a prized location on my bookshelf. I look thing up in it from time to time. It's a solid reference book.

Martelli, A., Ravenscroft, A., & Ascher, D. (2005). Python cookbook. " O'Reilly Media, Inc."

Whenever you are learning a new language, it's helpful to look for a cookbook in that language. A cookbook will show you what programs are easy and hard in the language you are learning. This is very helpful when you are trying to solve new challenges.

Lutz, M. (2001). Programming python. " O'Reilly Media, Inc."

This is a book for programmers. It is an exhaustive text on the python language with all the features that I will just gloss over in my course. If you plan on doing serious software development in python, this is the book for you.

Karau, H., Konwinski, A., Wendell, P., & Zaharia, M. (2015). Learning spark: lightning-fast big data analysis. " O'Reilly Media, Inc."

Big data problems can be solved with Spark. Just getting it running on a distributed data set is a big of a challenge, but if it's already running on a system, then PySpark is an excellent resource if you know how to think about the data.

Smith, K. W. (2015). Cython: A Guide for Python Programmers. " O'Reilly Media, Inc."

If you need to speed up your python code, Cython is a solution that might help. This book will more than get you started writing Cython code. I might reference this in the course.

Sweigart, A. (2015). Automate the boring stuff with Python: practical programming for total beginners. No Starch Press.

This is the one book on this list that is not available through O'Reilly Safari. The entire text is available at:

<https://automatetheboringstuff.com/>

This book is extremely useful for any basic task. Do you need to learn how to read JSON files for your internship? This book has the solution. It's an excellent reference that should be a bookmark in the browser of every python coder.

Books on data science

These books are available through SpringerLink through the UTDallas library.

Gentle, J. E. (2007). Matrix algebra. Springer texts in statistics, Springer, New York, NY, doi, 10, 978-0.

This is the book I will be using for matrix algebra. It is a very well written and clean matrix algebra book which conveniently skips all the hard jobs that a computer will solve for us. Do not read a book on row operations for learning matrix algebra. It's a waste of your time. This book is perfect for our needs.

Gentle, J. E., Härdle, W. K., & Mori, Y. (Eds.). (2012). Handbook of computational statistics: concepts and methods. Springer Science & Business Media.

If you want to learn statistics for data science, then a book on computational statistics is the perfect blend of computer programming and statistical theory. I highly recommend any book by James Gentle. He's an amazing author.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning Vol. 1, No. 10). New York: Springer series in statistics.

This is the data science handbook. It's amazing. It has every advanced machine learning tool developed in the 20th century. My goal for this course is that you can read, understand, and implement any tool in this book.

Grading Policy

50% Problem sets (5 total, equally weighted)

50% Exams (2 total, equally weighted)

Exams will be taken online through the eLearning platform. You are required to use Honorlock for the tests. Exams are open book and open note. Any collusion on the test will be prosecuted.

Grading is on an absolute scale. A= 93.5 and above, A-=89.5 to 93.5, B+= 87.5 to 89.5, etc.

Assignments & Academic Calendar

Class	PS	Topic	Data Science	Monday
1		Core python, Jupyter	Vector algebra	16-Jan
2		Scripting, IO, Parallel	Matrix algebra	23-Jan
3		Matrices, Numpy, MPL	Matrix calculus	30-Jan
4	1	OOP, Pandas	Linear models	6-Feb
5		Optimization	Logit, MLE	13-Feb
6	2	Simulation	Endogeneity	20-Feb
7		Training, LP/QP	Lasso, RF	27-Feb
8	3	Midterm		6-Mar
9		Validation	LassoCV, GBM	20-Mar
10		Clustering, EM	Kmeans	27-Mar
11	4	Network models	MLP, RNN	3-Apr
12		Bonus: Time series		10-Apr
13	5	Bonus: Panel		17-Apr
14		Bonus: Sympy		24-Apr
15		Final		1-May

Course & Instructor Policies

Extra credit is not available for this course. Make-up exams will not be given. With clear, physical proof for an excused absence in an exam (e.g., hospitalization or death of an immediate family member), an exam may be dropped (i.e., the other exam will count double). Classroom attendance will not be taken, but attendance is highly encouraged.

Comet Creed

This creed was voted on by the UT Dallas student body in 2014. It is a standard that Comets choose to live by and encourage others to do the same:

“As a Comet, I pledge honesty, integrity, and service in all that I do.”

UT Dallas Syllabus Policies and Procedures

The information contained in the following link constitutes the University’s policies and procedures segment of the course syllabus.

Please go to <http://go.utdallas.edu/syllabus-policies> for these policies.

The descriptions and timelines contained in this syllabus are subject to change at the discretion of the Professor.