COURSE: CE 599-002 Data Science in Transportation (Spring 2017)

TIME and PLACE: Tuesday and Thursday at 12:30 pm - 1:45 pm
WTYL B-25 (Basement of W.T. Young Library)

INSTRUCTOR: Dr. Greg Erhardt, 261 OHR, 323-4856, greg.erhardt@uky.edu

OFFICE HOURS: Mon 10:00-11:00 am, Wed 3:00-4:00 pm and by appointment (phone or email)
Students will be informed of any absences planned by the instructors

PREREQUISITES: Introductory course in computer programming, such as CS 115, CS 221 or EGR 102.
Introductory course in statistics, such as STA 381.

NOTE: Most of the communication for this course will be performed through email; check your email **at least once a day**!

COURSE CONTENT AND OBJECTIVES:

This course is designed around the Data Science Venn Diagram, as shown in **Figure 1**. It takes applications from the transportation realm, and introduces the practical skills needed to pursue data science both in the workplace and as a research student.

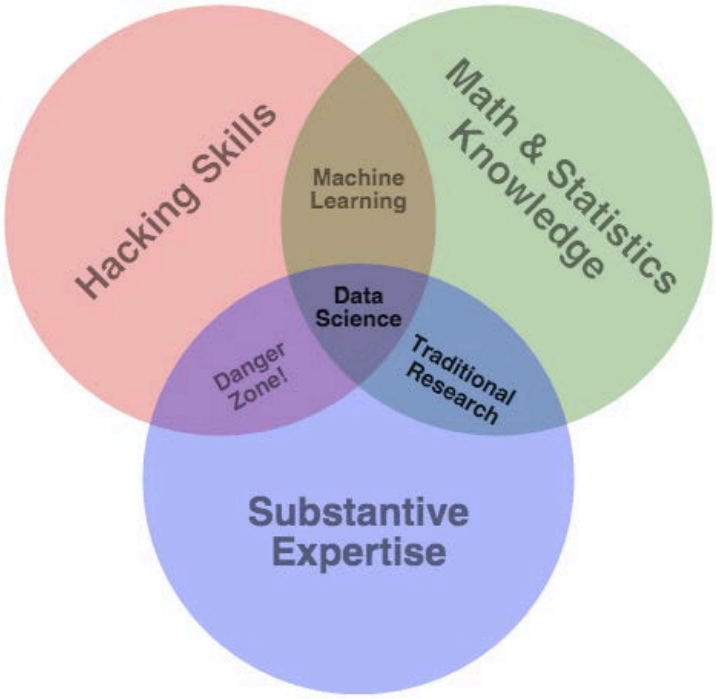

Figure 1 Data Science Venn Diagram[1]

Main topics to be covered include:

- Fundamentals of programming and data wrangling in Python
- Data visualization
- Applied statistical modelling and interpretation

1 Drew Conway, "The Data Science Venn Diagram," *Drew Conway*, accessed October 26, 2016, http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram.

- Written and oral "data storytelling"

COURSE COMPONENTS, REQUIREMENTS, AND GRADING:

Canvas

Canvas will be used for communication and posting course content outside of lectures. Please be sure you are on Canvas, in the course, and checking emails and messages sent through Canvas.

Readings

Two textbooks are required for this class:

Downey, Allen B. *Think Python: How to Think Like a Computer Scientist*. 2 edition. Sebastopol, CA: O'Reilly Media, 2015. ISBN: 978-1-4919-3936-9
Note: This is available under a Creative Commons license, and can be downloaded as a PDF for free from here: http://greenteapress.com/wp/think-python-2e/

McKinney, Wes. *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. 1 edition. Beijing: O'Reilly Media, 2012. ISBN: 978-1-4493-1979-3
Note: There is a PDF version and e-book that are available, and are suitable.

In addition, the following book is required reading, and will serve as the a foundation for the mid-term report:

Silver, Nate. *The Signal and the Noise: Why So Many Predictions Fail--but Some Don't*. 1 edition. New York: Penguin Books, 2015 (paperback). ISBN: 978-0-14-312508-2

Other sources will be identified in class and will be used as references as needed.

Class Participation and Exercises

Class attendance and participation is required, except in the cases described in the attendance policy below. Classes will require the active completion of specific programming exercises in Python. In some cases, these exercises will be completed in class, and in other cases they will be assigned as homework. In all cases, students will be required to successfully complete the exercise prior to the next class.

Questions and comments from students are welcome at any time during class. We are here to help and want to see you succeed.

Mid-Term Report

Students will write a mid-term report in which they will recommend improvements to the practice of transportation forecasting. Reports should start from a recognition of the issues in transportation forecasting, as discussed in:

Wachs, Martin. "Ethics and Advocacy in Forecasting for Public Policy." *Business and Professional Ethics Journal* 9, no. 1 & 2 (1990): 141–57.

Hartgen, David T. "Hubris or Humility? Accuracy Issues for the next 50 Years of Travel Demand Modeling." *Transportation* 40, no. 6 (2013): 1133–57.

And consider the lessons from other fields, as described in:

Silver, Nate. *The Signal and the Noise: Why So Many Predictions Fail--but Some Don't*. 1 edition. New York: Penguin Books, 2015 (paperback). ISBN: 978-0-14-312508-2

Further details of the assignment will be provided during the course.

Final Project

A semester project is part of the course requirements. Students will be required to apply the skills they have learned to a problem of interest in transportation. The project should draw from the three areas of data science. A final report will be required. The format of the report will be that of a research paper, and will follow the Transportation Research Board (TRB) guidelines. High-quality papers will be recommended for submission to the TRB Annual Meeting. A formal project proposal and at least one interim report will be required to ensure sufficient progress.

Students will also present the results of their project orally during class time. Further details of both components will be provided during the course.

Grading

Attendance and Exercises	40%	90 - 100 = A
Mid-Term Report	20%	80 - 89 = B
Final Project	30%	70 - 79 = C
Final Presentation	10%	60 - 69 = D
	100%	<60 = E

Appeals:

Grades on problem sets and tests (with the exception of the final exam) can be appealed according to the following procedure:

*After each assignment has been returned, you will have **ONE WEEK** to review it and plead your case for a grade change. Beyond the one-week period no changes will be made.*

IMPORTANT NOTE:

Students are requested to review the policies of the university with respect to discrimination and sexual harassment, and the policies of the College of Engineering with respect to academic dishonesty, excused absences (see below) and the final examination schedule.

OTHER ITEMS:

Please note that the use of any tobacco products during the class period is not allowed.

ATTENDANCE POLICY:

Students are required to attend all classes and participate in all fieldwork associated with assignments. Excusable absences include: (1) illness of student or serious illness of an immediate family member; (2) the death of a member of the student's immediate family; (3) trips for members or student organizations sponsored by an academic unit, trips for university classes, and trips for participation in intercollegiate athletic events; and (4) major religious holidays. Students are responsible for notifying the instructor in writing (NO EMAILS) prior to an event in categories (3) and (4). The instructor has the right to request verifications for all cases. Such verification needs to be provided by the student within two weeks of the absence in question.

ACKNOWLEDGEMENTS:

Many of the exercises are based on those developed by Paul Waddell and Geoff Boeing for CP255: Urban Informatics and Visualization at the University of California at Berkeley. Those materials are available from:

https://github.com/waddell/urban-informatics-and-visualization

Course Schedule:

Date	Topic
Jan 12	Intro to Jupyter and Python
Jan 17	Python Data Types, Version Control with Git
Jan 19	Programming Logic
Jan 24	Working with Data Files, Data Cleaning
Jan 26	Pandas Basics, Part 1
Jan 31	Pandas Basics, Part 2
Feb 2	Introduction to Open Data APIs
Feb 7	Working with APIs in Python
Feb 9	Resources: Open Data, Open-Source Software and StackExchange
Feb 14	Data Wrangling with Pandas
Feb 16	Data Visualization with Pandas, Matplotlib and Plotly
Feb 21	Data Visualization with Python
Feb 23	Mapping and Spatial Analysis
Feb 28	Group, Split, Apply, Combine
Mar 2	Working with Matrix Data
Mar 7	Working with Network Data
Mar 9	**Project Proposals** **Mid-Term Report Due**
Mar 14	**Spring Break – No Class**
Mar 16	**Spring Break – No Class**
Mar 21	Regression Modeling, Part 1
Mar 23	Regression Modeling, Part 2
Mar 28	Discrete Choice Modeling
Mar 30	Time-Series Modeling
Apr 4	Correlation, Causality and Co-linearity
Apr 6	Undestanding Biases and Errors
Apr 11	Performance Metrics: Good, Bad and Ugly
Apr 13	Telling a Story: Examples from Data Journalism **Interim Project Report Due**
Apr 18	Ethics in Data Science
Apr 20	**No Class**
Apr 25	**Final Presentations**
Apr 27	**Final Presentations**
May 4	**Final Project Report Due**