# DS 7333 Quantifying the World

## Course Overview

Quantifying the World (QTW) is built in two-week chunks of varying subject matter. The first week of each chunk consists of seventy – ninety minutes of theoretical and methodological background that will be necessary to solve the case study for the second week. Topics covered include multiple imputation, branching processes, parallel processing, management of very large data sets, time series, machine learning, and predication of location for indoor positioning systems.

Before taking this class, you should know

- The Nolan and Temple-Lang text (NTL) contains example R code for 12 different case studies. We present three of them (the other case studies will be presented in-class) and the statistical/computational methods to fully understand them in this course. However, due to the ever-changing nature of R, this code doesn't always work as stated in the text. The same is true of the Python. You are expected to work out these issues as much as possible.

## Learning Objectives

Students will:

- Form a testable hypothesis from an unstructured problem
- Create an analysis plan to test afore-mentioned hypothesis
- Use code in R and Python to perform various advanced analyses
- Learn about key advanced statistical and computational methods to analyze data
- Communicate the findings of a research project in a clear, concise, and scientific manner.

## Textbooks and Materials

- Nolan, D., and Temple Lang, D. (2015), *Data Science in R: A Case Studies Approach to Computational Reasoning and Problem Solving*. Boca Raton, FL: CRC Press (NTL)
- Killick, R., and Eckley, I. (2014), "changepoint: An R Package for Changepoint Analysis." *Journal of Statistical Software*, 58(3).
- Strobl, C., Malley, J., and Tutz, G. (2009), "An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests." *Psychological Methods*
- Rosenstrom, T. (2014), *Lecture Notes: Some Core Ideas of Imputation for Nonresponse in Surveys*. University of Helsinki

- Hastie, T., Tibshirani, R., Friedman, J. *The Elements of Statistical Learning* (2017) http://web.stanford.edu/~hastie/ElemStatLearn/

## Grading

| Assignment/Assessment | Weight on Final Grade |
|---|---|
| Case Studies | 50% |
| Participation | 30% |
| Asynchronous material | 20% |

**Case Studies (50%):** Every other week in class will be spent working through a Case Study using R or Python. These case studies will be written up and turned in for a grade. Case study write ups are to be “technical report ready”, which basically means that it is suitable for public viewing.

**Participation (30 %):** Weekly postings to discussion board on various topics. During case study weeks, the discussion board will consist of hints and help with the case study. Please do not post the code that you have used to solve a problem. You may post hints and pseudo code as needed.

**Asynchronous material (20%):** Watching asynchronous material and answering questions in BLTs as required.

**Rescheduling Exams:** Life happens. Instructors schedule major due dates on the same day. Should you need to an extension on an assignment, please let me know at least 24 hours prior to the due date. The notice should be given via e-mail. We will discuss the best course of action given your circumstances. Note: Issues with technology is NOT an excuse for late work. You have been warned to start your assignments early enough so that you can resolve such issues before they affect your ability to turn in the work.

## Assignment and Assessment Information

**Submission guidelines for assignments**

- Your name must be at the top of the first page and on each successive page.
- Submit case studies as a formal written paper. The case study should have an abstract, an introduction, a literature review, a methods section, a results section, and a future work/discussion/conclusion section. Code should be included in an appendix to the document. Spelling and grammar count!
- Use an easy-to-read variable-width font (Document is in Arial 11) with a minimum of 11 point font.
- Relevant code and output must be included in-line at the appropriate point using Courier New (or other fixed width) font, in 10 point size. **Inclusion of irrelevant code or output, even in an appendix, will be penalized.** All software output must be given in the text or as a table created in Word (or the software you are using).

  ```
  An Example of a fixed width font
  ```

- Any graphics must be electronically cut and pasted in-line at the appropriate point of the write-up. You can use Word to resize the graphics appropriately**. Screen shots from R, Python, or tabular output are not allowed in the document text or in the appendix.** All tables and figures should have descriptive titles and captions. In short, the reader should be able to understand the content of the figure or table without reading the associated text.
- Any mathematical notation must be provided with appropriate use of subscripts, superscripts, and symbols. Use MS Equation or another equation editor if you submit your work in Word.
- Acceptable formats include MS Word, PDF, HTML, or Jupyter Notebook.

## Weekly Schedule

| Unit | Topic | Readings | Assignments due |
|---|---|---|---|
| 1 | R | Required Readings<br>Nolan, D., and Temple Lang, D. (2015), *Data Science in R: A* | None |

| | | | |
|---|---|---|---|
| | | *Case Studies Approach to Computational Reasoning and Problem Solving*. Boca Raton, FL: CRC Press (NTL). Chapter 1.<br>Additional Readings (Not Required)<br>Ault, A., Zhong, X., and Coyle, E. J. (2005), *K-Nearest-Neighbor Analysis of Received Signal Strength Distance Estimation Across Environments.* Technical Report. Purdue University: Center for Wireless Systems and Applications,=.<br>Madigan, D., Ju, W. H., Krishnan, P., Krishnakumar, A. S., and Zorych, I. (2006), "Location Estimation in Wireless Networks: A Bayesian Approach." *Statistica Sinica*, 16, 495-522.<br>Tarrio, P., Bernardos, A. M., and Casar, J. R. (2011), "Weighted Least Squares Techniques for Improved Received Signal Strength Based Localization." *Sensors*, 11(9), 8569-8592. | |
| 2 | Clustering | Case Study Coding and write up | None |
| 3 | Web scraping | Required Readings<br>Nolan, D., and Temple Lang, D. (2015), *Data Science in R: A Case Studies Approach to Computational Reasoning and Problem Solving*. Boca Raton, FL: CRC Press (NTL). Chapter 2.<br>Killick, R., and Eckley, I. (2014), "changepoint: An R Package for Changepoint Analysis." *Journal* | Case Study #2 |

| | | | |
|---|---|---|---|
| | | *of Statistical Software*, 58(3). | |
| 4 | Loess | Nolan, D., and Temple Lang, D. (2015), *Data Science in R: A Case Studies Approach to Computational Reasoning and Problem Solving*. Boca Raton, FL: CRC Press (NTL). Chapter 2, 45-103. | None |
| 5 | Naïve Bayes (Bayes Theorem, Metrics, Confusion Matrix) | Required Readings<br>Nolan, D., and Temple Lang, D. (2015), *Data Science in R: A Case Studies Approach to Computational Reasoning and Problem Solving*. Boca Raton, FL: CRC Press (NTL). Chapter 3.<br>Strobl, C., Malley, J., and Tutz, G. (2009), "An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests." *Psychological Methods* 14(4): 323-348. | Case Study #4 |
| 6 | Boosted Trees, CART, Random Forests, Boosting | None | None |
| 7 | Python & virtual environments | In class Notes | Case Study #6 |
| 8 | Time Series (+ ARIMAS) | In Class Notes | None |
| 9 | Missing Data, Monte Carlo, Markov Chains | Rosenstrom, T. (2014), *Lecture Notes: Some Core Ideas of Imputation for Nonresponse in Surveys*. University of Helsinki. | Case Study #8 |
| 10 | Missing Data (Metropolis Hastings MCMC function) | Case study coding and write-up | None |
| 11 | Neural Nets 1: Out of core Memory (Stochastic Gradient Descent), Vowpal Wabbit, Dealing with big data, Intro to Neural Nets | In Class Notes. Elements of Statistical Learning | Case Study #10 |
| 12 | Neural Nets 2: Convolution Nets for images, Recurrent Nets for language | Case study coding and write-up | None |
| 13 | Putting it together: cleaning data, staking models | In Class Notes. | Case Study #12 |
| 14 | Putting it together & multiclass | Case Study Coding and Write-Up | |

| | | | |
|---|---|---|---|
| | classifiers | | |
| 15 | Review | None | Final Case Study due |

## University Policies

**Grading Policy: Graduate Students must receive a C or better in a course in order to pass the course. If a student must retake a course, then the second grade and the first grade are averaged for the purposes of the overall GPA. Failure to maintain a GPA of 3.0 or better will result in dismissal from the program.**

**Incompletes** will be given only in the case of extraordinary circumstances that prevent you from finishing the semester. You must have completed at least 50% of the course with a passing grade to be eligible for an incomplete.

**Religious Observance:** Religiously observant students wishing to be absent on holidays that require missing class should notify the live session instructor via e-mail, and should discuss with the instructor, in advance, acceptable ways of making up any work missed because of the absence. (See University Policy No. 1.9.)

**Excused Absences for University Extracurricular Activities:** Students participating in an officially sanctioned, scheduled University extracurricular activity will be given the opportunity to make up class assignments or other graded assignments missed as a result of their participation. It is the responsibility of the student to make arrangements with the instructor prior to any missed scheduled examination or other missed assignment for making up the work (University Undergraduate Catalogue).

## Best Practices for Success

**Attendance**. Take responsibility for your commitment. Attendance means not only being there for synchronous sessions but also participating in asynchronous work.

**Citizenship.** You need to be actively engaged to succeed in this class. Talking on cell phones, texting, “facebooking,” tweeting, or leisure web browsing are prohibited in class. I consider these to be a disruption (not to mention rude).

**Integrity.** A lot of the graded work occurs outside of class, so I expect honesty and integrity in what you submit for evaluation. Evidence of academic dishonesty will minimally result in zeros for all involved parties, and perhaps University-level disciplinary action. Don’t risk your academic career.

**Humility.** Don’t get lost! Ask questions in class. If something isn't clear to you, it probably isn't clear to others either. Questions may arise because I haven’t made a connection clear or have inadvertently left out an important point. Your question gives me a chance to explain more clearly. Don't be proud or shy.

**Organization.** Don’t procrastinate! This is a technology-driven course. Count on your computer failing or your wireless connection breaking the night before a due date. Start early and give yourself a chance to succeed.

**Deadlines.** You will generally have a week to complete an assignment. Due dates and times will be clearly indicated. Late submissions will be penalized, but it is much better to turn in work late than not at all (or to turn in incomplete/sloppy work). Work turned in after solutions have been posted to the course website will receive no credit.

**Getting help.** If questions arise while doing assignments/exams, do your best to resolve these questions before the assignment is due, first by taking time to seek answers yourself, and then via e-mail to your instructor or other students. **I encourage you and expect you to seek help.**

**Collaboration.** I encourage the formation of study groups and collaboration with your fellow students in tackling the assignments. Working in groups on homework is permitted, even encouraged. **However, every student should write up and complete his or her homework independently. Students who chose to turn in exactly the same work will share the grade**

**assigned.** Talking about problems with other people does help in learning, but just copying the solutions from one another doesn't help!

**Looks do matter!** All assignments must be NEATLY executed and organized. You risk a zero on any assignment submitted in a sloppy manner. See submission guidelines for more detail.

*This syllabus is only a guideline and is not a legal contract. The professor of record for the course has final say on any policies, due dates, etc.*