# Data Technologies —

# Stat 5810, Section 004 & Stat 6910, Section 004

# Fall 2018

Instructor: Dr. Jürgen Symanzik
Office: AnSc 313
Phone: 797–0696
FAX: 797–1822
e–mail: symanzik@math.usu.edu
Web: `http://www.math.usu.edu/~symanzik/`
`http://www.math.usu.edu/~symanzik/teaching/2018_stat5810_004_fall/stat5810_004.html`

Office Hours: Tuesday (T) & Thursday (H) 1:30pm – 2:30pm and by appointment.

Classes & Rooms:
TH 12:00noon – 1:15pm, T 9/11 – H 11/29, 2018 (tentatively):
Eccles Business Building (EBB) 216.

**Please visit the course Web page listed above and/or Canvas frequently for lecture notes, data sets, graphical examples, R code, etc. — in particular if you miss class for any reason.**

Detailed Class Schedule:
For a 2–credit course, we need 20 lectures/lecture days (in contrast to 29 or 30 lectures/lecture days for a 3–credit course). Those days are marked as "Lecture 01" to "Lecture 20" in the overview below:

| Week | Tuesday | Thursday |
|---|---|---|
| 1 | 8/28 No class | 8/30: No class |
| 2 | 9/4: No class | 9/6: No class |
| 3 | 9/11: Lecture 01 | 9/13: Lecture 02 |
| 4 | 9/18: Lecture 03 | 9/20: Lecture 04 |
| 5 | 9/25: Lecture 05 | 9/27: Lecture 06 |
| 6 | 10/2: Lecture 07 | 10/4: Lecture 08 |
| 7 | 10/9: Lecture 09 | 10/11: Lecture 10 |
| 8 | 10/16: Lecture 11 | 10/18: Lecture 12 |
| 9 | 10/23: Lecture 13 | 10/25: Lecture 14 |
| 10 | 10/30: Lecture 15 | 11/1: Lecture 16 |
| 11 | 11/6: No class | 11/8: No class |
| 12 | 11/13: Lecture 17 | 11/15: Lecture 18 |
| 13 | 11/20: No class | 11/22: No class |
| 14 | 11/27: Lecture 19 | 11/29: Lecture 20 |
| 15 | 12/4: Backup | 12/6: Backup |

Note: "No class" means guaranteed no class that day. I have marked the last week as "Backup", e.g., in case we miss lectures because I am sick or have to travel. But hopefully, this won't happen. If nothing goes wrong, our tentative last lecture date will be on H 11/29/2018.

**Course Objectives:**

Note that Andreas Buja, the Liem Sioe Liong/First Pacific Company Professor in the Statistics Department, The Wharton School, at the University of Pennsylvania in Philadelphia, USA, already stated in a 2006 interview: *"I think in education we still have some ways to go to find a balance of things that we want to teach students. There is the traditional curriculum that gives Ph.D. students a solid foundation in theory, but then they also should acquire computational skills, they should become good applied statisticians who have good sense, good data sense. Many of these things are really hard to teach. You can teach them details, but ultimately they have to pick up the high level of thinking, the creative way of thinking, on their own or by being thrown like fish into the water, either they swim or they don't. That is hard to teach. Something that I see lacking right now is, especially if you are interested in education with industry in mind, what you need out there in industry I think is not specifics of modeling, it is good data sense, it is data skills, data literacy. I think that was a term used by one of the earlier interviewees. Data literacy, in general, is the ability to get data and start doing something sensible, and that is of utmost importance. Part of that is that we cannot assume that other people are doing data cleaning for us. We have to do that ourselves. So here I see a gap actually in our education. I don't think most statistics programs teach something like a scripting language and practice data cleaning, reshaping of data, basic tabulations, mild aggregations, getting subsamples, systematic ones and random ones, and so on. These are very important activities, and we still need to get better at teaching them."* (see Computational Statistics (2008) 23:177–184 for the full interview).

In the "Introduction to R" course, you have learned basic data skills and data literacy via R to achieve this goal. You were (likely) thrown like fish into the water (without any prior swimming course) — and you learned to swim! Now, we will extend these basic skills to do something really meaningful with a variety of data sets. Eventually, this course will be one of your most valuable courses for a future career in research (assuming you are working with any kind of data) or in industry.

**Prerequisites:**

I expect basic knowledge of R as taught in the "Introduction to R" course. Moreover, you should be familiar with a tool such as R Markdown, knitr, or sweave that allows you to combine text, R code, graphics, and numerical results in high-quality documents. LaTeX is a plus but is not formally required at the 5000 level, but it will be required at the 6000 level of this course.

Moreover, I expect basic "operational" knowledge from an introductory stats course such as Stat 2000, Stat 3000, or higher. "Operational" means that you still recall sufficient details from regression, ANOVA, hypothesis tests, etc. (it is not sufficient that you have taken such a course several years ago and have forgotten almost all details).

**IDEA Center Learning Objectives:**

**Objective 1)** Gaining factual knowledge (terminology, classifications, methods, trends).

**Objective 2)** Learning fundamental principles, generalizations, or theories.

**Objective 3)** Learning to apply course material (to improve thinking, problem solving, and decisions).

**Topics:** (subject to change)

1. Data.
2. Basics of simulation.
3. Representation of information.
4. Regular expressions.
5. Web scraping.
6. XML.
7. Data bases and SQL.
8. Resampling/bootstrap.
9. The *Tidyverse* (R packages for data science).
10. Others (as time permits).

We will work with real "messy" data that have not been preprocessed nor analyzed so far. These data will contain surprises — for you and for me. Do not expect that someone is going to give you the final answer or model. We jointly will have to work towards such an answer or model.

For MS and PhD students majoring in Statistics, it is important to learn LaTeX — from basic document preparation, over the inclusion of R graphics into your LaTeX documents to advanced topics such as Sweave (`https://leisch.userweb.mwn.de/Sweave/`) and the LaTeX bibliography BibTeX (`http://www.bibtex.org/`). LaTeX is essential for graduate work (at the MS and PhD level) and will be used for many theses, dissertations, and scientific publications. Therefore, LaTeX will have to be used for all homeworks, projects, presentations, etc. at the 6000 level of this course.

**Homework Assignments:**

There will be a variety of assignments throughout the semester. Each assignment will include a value (typically 20–100 points) that it will be scored out of. Your final grade will be determined by the sum of your points in all assignments. Some assignments will include combinations of computer work in R (or others) and short oral presentations. The value of each assignment will be roughly proportional to its importance and the amount of work involved.

Regular homework assignments will be done individually or in groups of 2 or 3 students. For individual assignments, you will be allowed to discuss general approaches to questions on the assignments with other students, but each student must write and submit their own code and comments. Any students caught sharing code will fail the class.

Unless otherwise stated on the assignment sheet, all homework assignments have to be submitted electronically via Canvas.

The following deductions will be applied to late homework submissions: 1 min – 24 hours late: 10% off; > 24 hours – 48 hours late: 25% off; > 48 hours – 72 hours late: 50% off. Homeworks won't be accepted later than 72 hours (i.e., 3 days) after the submission deadline.

There will be no (in–class or take–home) quizzes, midterm exams, or final exams. Nevertheless, this will be a very challenging course that requires a lot of individual time to work on the assignments (and project). Just attending classes will not be enough to pass this course! In addition, you will have to do a lot of individual reading of textbooks, online documentation, and help pages, and search for available information on the web.

**Project (Stat 6910 only):**

There will be one major project towards the end of the semester. This will include the preparation of a final project report and possibly a short presentation of your work for the other students in this course. The project will be done individually or in a small group of students. The project will account for about 30% of your final grade.

**Textbooks:**

Murrell, Paul (2009) *Introduction to Data Technologies*, Boca Raton, FL: Chapman and Hall/CRC.
Note that the entire book is available online from `http://www.stat.auckland.ac.nz/~paul/ItDT/` under a Creative Commons licence.

Nolan, Deborah, and Temple Lang, Duncan (2015) *Data Science in R — A Case Studies Approach to Computational Reasoning and Problem Solving*, Boca Raton, FL: CRC Press/Taylor & Francis.

If you plan to work in the area of data science for your MS or PhD degree, you should consider to purchase these books for an ongoing use beyond this course.

**Software:**

We will primarily be using R (`http://cran.r-project.org/`), a free software environment for statistical computing and graphics. Please install the current version of R, i.e., 3.5.1, on your own computer so we can exchange code. Also install RStudio (`https://www.rstudio.com/`) as a front end to R.

**Credits:**

This course uses some of the course materials provided by Dr. Paul Murrell (University of Auckland: `https://www.stat.auckland.ac.nz/~paul/`), Dr. Duncan Temple Lang (UC Davis: `http://www.stat.ucdavis.edu/~duncan/`) and Dr. Deborah Nolan (UC Berkeley: `http://www.stat.berkeley.edu/~nolan/`). We are likely to include parts from additional web sources that will be specified later on.

**Courtesy:**

Please turn off cell phones and similar devices before class, and please keep conversations to a minimum during lectures. Please do not read/reply to your e–mails or browse other web pages than the ones discussed during class.

I will not keep track if you come to class or not. However, I would highly recommend to attend all lectures. If you have to miss a lecture, there will be a recording of the lecture available in Canvas Panopto (if the technology doesn't fail).

**Americans with Disabilities Act:**

If a student has a disability that will likely require some accommodation by the instructor, the student must contact the instructor and document the disability through the Disability Resource Center (DRC), during the first week of the course. Any requests for special considerations relating to attendance, pedagogy, taking of examination, etc. must be discussed with and approved by the instructor. In cooperation with the Disability Resource Center, course materials can be provided in alternative formats — large print, audio, or Braille.

**Note:**

The above schedule and procedures in this course are subject to change in the event of extenuating circumstances.