Data Technologies —

Stat 5080, Section 001 & Stat 6080, Section 001

Fall 2021 (2 Credits)

Instructor: Dr. Jürgen Symanzik
Office: AnSc 313
Phone: 435–797–0696
e–mail: symanzik@math.usu.edu
Web: `http://www.math.usu.edu/~symanzik/`
`http://www.math.usu.edu/~symanzik/teaching/2021_stat5080/stat5080.html`

Office Hours: Tuesday (T) & Thursday (H) 1:30pm – 2:20pm in AnSc 313 (please wear a face mask), Friday (F) TBA (via Zoom), and by appointment.

Classes & Rooms:
TH 12:00noon – 1:15pm, T 9/14 – H 11/18, 2021 (tentatively): HPER 114A (face–to–face).

Please visit the course Web page listed above for emergency announcements, e.g., when Canvas is unavailable. Otherwise, visit Canvas frequently for lecture notes, data sets, R code, etc. — in particular if you miss our face–to–face lectures for any reason. All (additional and updated) materials, announcements, discussions, recordings, etc. from Canvas are part of the course materials. Not seeing one of these in time does not serve as an excuse for not getting point deductions for the course. Deadlines may change or *Coronavirus/Covid–19* regulations and requirements may be updated. It is your responsibility to make sure to receive all announcements in time.

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	8/31 No class	9/2: No class
2	9/7: No class	9/9: No class
3	9/14: Lecture 01	9/16: Lecture 02
4	9/21: Lecture 03	9/23: Lecture 04
5	9/28: Lecture 05	9/30: Lecture 06
6	10/5: Lecture 07	10/7: Lecture 08
7	10/12: Lecture 09	10/14: Lecture 10
8	10/19: Lecture 11	10/21: Lecture 12
9	10/26: Lecture 13	10/28: Lecture 14
10	11/2: Lecture 15	11/4: Lecture 16
11	11/9: Lecture 17	11/11: Lecture 18
12	11/16: Lecture 19	11/18: Lecture 20
13	11/23: Backup	11/25: No class
14	11/30: Backup	12/2: Backup
15	12/7: Backup	12/9: Backup

Note: "No class" means guaranteed no class that day. I have marked the dates in the three last weeks of the semester as "Backup", e.g., in case we miss lectures because of Covid–19 restrictions, sickness, or any other reason. If nothing goes wrong, our tentative last lecture date will be on H 11/18/2021. Backup lectures (if any) will be held via Zoom (and not face–to–face). There is no need to return to Logan after Thanksgiving for a backup lecture in this course.

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:

STAT 5050 ("Introduction to R") with a C− or better. STAT 3000 ("Statistics for Scientists") or STAT 5100 ("Modern Regression Methods") with a C− or better. STAT 5550 ("Statistical Visualization I") is recommended. 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 that you still have "operational" knowledge from your STAT 3000 or STAT 5100 course. "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://stat.ethz.ch/R-manual/R-devel/library/utils/doc/Sweave.pdf`), knitr (`https://yihui.org/knitr/`), 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 3 HW assignments for this course, roughly one every three weeks. Each HW assignment will include a value (typically 20–100 points) that it will be scored out of. HW assignments will contribute to 100% (for Stat 5080), repectivley 70% (for Stat 6080), of your course grade. The value of each HW assignment will be roughly proportional to its importance and the amount of work involved.

You will be allowed to discuss general approaches to questions on the HW assignments with other students, but each student must write and submit their own R code and comments. Any students caught sharing R code or other parts of their homework submissions will fail the class.

Unless otherwise stated on the HW assignment sheet, all homework assignments have to be submitted electronically via Canvas. **You will have 2 or 3 weeks after the last lecture to finalize and submit the last HW assignment.**

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. We will have a few worksheets for training purposes only. Nevertheless, this will be a very challenging course that requires a lot of individual time to work on the assignments (and project for Stat 6080). 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 6080 only):

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

Every student should have access to each of these books, but it is not necessary that every student buys all of these books. The USU library holds several of these books or provides online access. 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 a recent version of R, i.e., 4.0.4 or later, on your own computer so we can exchange code. Also install RStudio (`https:`

//www.rstudio.com/) as a front end to R and a version of LaTeX that can be used in connection with RStudio. My personal recommendation is TeX Live (https://www.tug.org/texlive/).

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 Kaltura recording of the lecture available in Canvas (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 – https://www.usu.edu/drc/), preferably 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.

Covid–19 Requirements and Restrictions:

1. Although not formally required, please wear a face mask in all university buildings, in particular in all classrooms and all private offices. I will do the same.
2. Please continue to maintain a 6–foot distance to other students, instructors, and staff. Leave at least one (or more if possible) empty seat to the next student in the classroom. Fill in the seating chart in Canvas in the week of 9/20/2021.
3. If you aren't vaccinated against Covid–19 yet, please get vaccinated and upload your proof of vaccination to https://aggiehealth.usu.edu. As you may be aware, USU is finalizing plans to require students to be vaccinated for the Spring 2022 semester. According to a recent release on 9/10/2021 from the Centers for Disease Control and Prevention (CDC), fully vaccinated people have a reduced risk of infection ($5\times$), hospitalization ($> 10\times$), and death ($> 10\times$) from Covid–19 (including the Delta variant), compared to unvaccinated people (see https://www.cdc.gov/mmwr/volumes/70/wr/mm7037e1.htm). So, get vaccinated — the sooner the better (if you aren't vaccinated yet).
4. Stay home if you feel sick or suffering even mild symptoms of illness that could be related to Covid–19 and to get a free Covid–19 test. All lectures will be recorded via Kaltura so you can watch them in Canvas afterwards.
5. Make sure that you regularly check the guidelines for risk reduction and case management as they are updated on the main USU Covid–19 website (https://www.usu.

edu/covid-19), and specifically the "Communications to Student" section (https://www.usu.edu/covid-19/communications/students).

6. Check for announcements in Canvas (and/or via e–mail), ideally each time before you plan to leave for one of our face–to–face lectures.

Note:

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