syllabus: honors introduction to data science

This document is current as of 2021-08-22. An updated version may be found at *https://kevinlanning.github.io/DataScienceF2021/*](https://kevinlanning.github.io/DataScienceF2021/).

basics

This is **COP 3076 Section 1**. It is a 3-credit class offered in the **Fall 2021** term. The class meets **MW 800-920 on the Zoom platform**. There is no lab.

The professor is Kevin Lanning. My office hours are **TTh 2-4;** these will typically be held virtually, but if you would like to meet me in person we can schedule an appointment on the Jupiter campus (please send me an email, or ask me in class). You may schedule virtual appointments at https://calendly.com/kevin-lanning/student-advising. My office phone is (561) 594-1018, but it is quicker to reach me via email at lanning@fau.edu.

course prerequisites / co-requisites: STA 2023 (or equivalent, or permission of instructor) is a prerequisite.

course description: COP 3076 is an introductory seminar on data science.

course delivery mode: Online via Zoom.

note of honors distinction: The course receives honors credit by virtue of its small class size, by virtue of a dialectic approach in the classroom structure, and by the fact that students receive extensive exposure to supplementary materials and primary sources. This course differs substantially from a non-Honors course in that (a) the expectations for participation in class discussions will be greater than in a typical undergraduate course with a larger number of students, (b) class projects will be undertaken in heterogeneous groups in which students will be teaching and learning from their peers as well as the instructor, (c) assignments and expectations will be, to some extent, tailored to the backgrounds and interests of the individual student (d) the data sets we will collaboratively examine will be chosen to foster disciplinary breadth.

course objectives / student learning outcomes

Hochster (in Hicks & Irizarry, 2017) describes two broad types of data scientists: Type A (**Analysis**) data scientists, whose skills are like those of an applied statistician, and Type B (**Building**) data scientists, whose skills lie in problem solving or coding, using the skills of the computer scientist. Our course is closer to a Type A than a Type B treatment, one which is closer to Statistics than to Computer Science, but it is also essentially concerned with **Content** in the concentrations in the arts, humanities, and natural and social sciences. It is thus best understood as a third (Type C) approach, one which has as its objectives progress not just in the understanding of statistics and computing, but also in skills such as collaboration and communication, in exposure to the methods and tools of reproducible science, and in fostering a heightened sensitivity to the ethical challenges of the digital age.

The course will be taught using the statistical and graphical language R. In addition to R, we'll use a range of other tools, including the Canvas Learning Management System (LMS) for communication and collaboration, and spreadsheets such as Excel or Google Sheets. These tools will be used in service of a hierarchy of goals, ranging from literacy through proficiency, then fluency, and ultimately towards 'leadership.'

The course is intended to count towards FAU's Undergraduate Research Certificate program by virtue of its emphases on

- helping students to formulate questions (students will formulate research questions, scholarly or creative problems with integration of fundamental principles and knowledge in a manner appropriate to Data Science),
- critical thinking (students will apply critical thinking skills to evaluate information, their own work, and the work of others),
- ethical conduct (students will identify significant ethical issues in research and inquiry in Data Science),
- as well as helping them to develop plans of action - in essence, programs - to address research and inquiry questions or scholarly problems, and finally,
- communication, ranging from annotating code to facilitate reproducibility to designing data visualizations which are clear, effective, and truthful.

paths beyond the introductory course

Students interested in specializing in Data Science have several possibilities. At this writing, there is enthusiasm across units of FAU and its affiliated institutes, including Max Planck and FAU's Colleges of Science and Engineering, for integrating data science into our curriculum. The WHC is at the forefront of this, with a data science minor and, in collaboration with the College of Engineering, a data analytics concentration. Students interested in concentrating in Data Science may also pursue an individual concentration (see Dr. Lanning for details). In addition, there are several integrated '4 + 1' pathways which will lead to a master's degree in the College of Engineering.

student learning outcomes

The course is intended to count towards FAU's Undergraduate Research Certificate program by virtue of its emphases on

· helping students to formulate questions (students will formulate research questions, scholarly or creative problems with integration of fundamental principles and knowledge in a manner appropriate to Computational Social Science),

· critical thinking (students will apply critical thinking skills to evaluate information, their own work, and the work of others),

· ethical conduct (students will identify significant ethical issues in research and inquiry in Computational Social Science),

· as well as helping them to develop plans of action - in essence, programs - to address research and inquiry questions or scholarly problems, and finally,

· communication, ranging from annotating code to facilitate reproducibility to designing data visualizations which are clear, effective, and truthful.

required texts and materials

Wickham, H. & Grolemund, G. (2016) *R for data science.* Sebastopol, CA: O'Reilly or online at http://r4ds.had.co.nz. (our primary text)

Lanning, K. (in preparation). *Data Science for the Liberal Arts.* https://kevinlanning.github.io/DataSciLibArts/

In addition, there are a number of papers, manuals, and websites which we will access at least occasionally. These will be provided to you in Canvas (under the 'Files' tab) and/or in links in the schedule below.

minimum technology and computer requirements: You'll need a reliable laptop computer, ideally running either Windows or Mac OS, and a wi-fi connection with sufficient bandwidth so that all of us can see as well as hear each other. (If you don't have these resources, please let me know as soon as possible).

a note on teaching and learning in a pandemic: I appreciate that you are persisting in the face of real obstacles, including the stresses of being disconnected from friends and family, distractions from proximal (working from home) and distal (the news) sources, tech hiccups with your home wi-fi and/or an overburdened internet, concerns for loved ones, and possibly challenges to your own physical as well as mental health. You are to be commended for what you have already accomplished under these taxing circumstances. I will do everything I can to help you succeed in this class.

course assessments, assignments, & grading policy

Grades will be based on a 100 point scale, with points earned by participation, homework and quizzes, two term projects, a self-assessment, and a final report.

participation (10 points). Attendance is a necessary but not sufficient part of class participation. Your participation grade will be based also on the extent to which you contribute to our class by asking constructive questions and helping your classmates solve the numerous challenges which we will collectively face. That is, you can earn participation points by showing up, learning, engaging, and helping your classmates (particularly in small breakout groups).

homework/quizzes (20 points). These are also linked to attendance. Most homework projects will be submitted as in-class quizzes on the assigned date.

two term projects (40 points total). Learning is social. The term projects will be collaborative, data-based projects which you will undertake with two to four of your peers and which you will submit as fully-contained R markdown documents (that is, as reproducible documents which will include your argument, links to data, commented code, and the results of statistically appropriate analyses and/or data visualizations). The projects will be empirical, typically from data that I provide you with or we find together. The datasets that we will be working with will be small enough to analyze on your laptops in R.

In order for us to assess your individual contributions and to minimize social loafing, I ask that all group members digitally sign cover sheet describing the primary contribution and percent effort of each person. We'll work together on creating groups that will, hopefully, maximize synergies among you. Groups and paper topics will be developed in class. You'll present your projects in class as well.

a final report (30 points). Your final not-an-exam will include three parts (a) submission of a sample of your best work (code) in the class, (b) responses to a brief set of take-home questions about the class presentations as well as exercises taken from R4DS, and (c) your own self-assessment of how much you have learned this term.

extra credit (5 points maximum). You'll have the opportunity to earn additional points by solving one or more data challenges that we will develop as the class goes forward.

time commitment per credit hour: This course has three (3) credit hours. For traditionally delivered courses, not less than one (1) hour of classroom or direct faculty instruction each week is expected for fifteen (15) weeks per Fall or Spring semester, and a minimum of two (2) hours of out-of-class student work for each credit hour. Equivalent time and effort is required for Summer Semesters, which usually have a shortened timeframe. Fully online courses, hybrid, shortened, intensive format courses, and other non-traditional modes of delivery will demonstrate equivalent time and effort.

course grading scale

note that in borderline cases, students may receive the higher of two grades if there is evidence of sustained effort and/or improvement over the course of the term

grade	A	A-	B+	B	B-	C+	C	C-	D+	D	D-	F
min	93	90	87	83	80	77	73	70	67	63	60	0
max	100	92	89	86	82	79	76	72	69	66	62	59

schedule and due dates

The schedule is a dynamic document. While due dates for exams are, pending any university-wide mandates, fixed, all other dates and content are subject to change. Please monitor the announcements in Canvas for the latest updates. Our working schedule may be found at https://bit.ly/DataSciF2021Schedule. Here is the schedule as of 2021-08-22:

Again, please see https://bit.ly/DataSciF2021Schedule for the latest updates.

course policies

incomplete grade policy: University policy states that a student who is passing a course, but has not completed all work due to exceptional circumstances, may, with consent of the instructor, temporarily receive a grade of incomplete ("I"). The assignment of the "I" grade is at the discretion of the instructor, but is allowed only if the student is passing the course.

attendance policy: Students are expected to attend all of their scheduled University classes and to satisfy all academic objectives as outlined by the instructor. The effect of absences upon grades is determined by the instructor, and the University reserves the right to deal at any time with individual cases of non-attendance. Students are responsible for arranging to make up work missed because of legitimate class absence, such as illness, family emergencies, military obligation, court-imposed legal obligations or participation in University-approved activities. Examples of University-approved reasons for absences include participating on an athletic or scholastic team, musical and theatrical performances and debate activities. It is the student's responsibility to give the instructor notice prior to any anticipated absences and within a reasonable amount of time after an unanticipated absence, ordinarily by the next scheduled class meeting. Instructors must allow each student who is absent for a University-approved reason the opportunity to make up work missed without any reduction in the student's final course grade as a direct result of such absence.

special course requirements: None.

additional selected university & college policies

classroom etiquette/disruptive behavior policy statement: Disruptive behavior is defined in the FAU Student Code of Conduct as "... activities which interfere with the educational mission within classroom." Students who disrupt the educational experiences of other students and/or the instructor's course objectives in a face-to-face or online course are subject to disciplinary action. Such behavior impedes students' ability to learn or an instructor's ability to teach. Disruptive behavior may include, but is not limited to non-approved use of electronic devices (including cellular telephones); cursing or shouting at others in such a way as to be disruptive; or, other violations of an instructor's expectations for classroom conduct. For more information, please see the FAU Office of Student Conduct.

code of academic integrity policy statement: Students at Florida Atlantic University should endeavor to maintain the highest ethical standards. Academic dishonesty is a serious breach of these ethical standards, because it interferes with the University mission to provide a high quality education in which no student enjoys an unfair advantage over any other. Academic dishonesty is also destructive to the university community, which is grounded in a system of mutual trust and places high value on personal integrity and individual responsibility. Harsh penalties are associated with academic dishonesty. For more information, see University Regulation 4.001 and the WHC code at http://www.fau.edu/honors/academics/honor-code.php.

Plagiarism is the deliberate use and appropriation of another's work without identifying the source and trying to pass off such work as one's own. Any student who fails to give full credit for ideas or materials taken

from another has plagiarized. This includes all discussion board posts, journal entries, wikis, and other written and oral presentation assignments. Plagiarism is unacceptable in the University community. Academic work must be an original work of your own thought, research, or self-expression. When students borrow ideas, wording, or organization from another source, they must acknowledge that fact in an appropriate manner. If in doubt, cite your source.

disability (accessibility) policy statement: In compliance with the Americans with Disabilities Act Amendments Act (ADAAA), students who require reasonable accommodations due to a disability to properly execute coursework must register with Student Accessibility Services (SAS) and follow all SAS procedures. SAS has offices across three of FAU's campuses – Boca Raton, Davie and Jupiter – however disability services are available for students on all campuses. For more information, please visit the SAS website at www.fau.edu/sas/.

grade appeal process: You may request a review of the final course grade when you believe that one of the following conditions apply: There was a computational or recording error in the grading, the grading process used non-academic criteria, there was a gross violation of the instructor's own grading system. Chapter 4 of the University Regulations contains information on the grade appeals process.

religious accommodation policy statement: In accordance with rules of the Florida Board of Education and Florida law, students have the right to reasonable accommodations from the University in order to observe religious practices and beliefs with regard to admissions, registration, class attendance, and the scheduling of examinations and work assignments. For further information, please see Academic Policies and Regulations.

university approved absence policy statement: In accordance with rules of the Florida Atlantic University, students have the right to reasonable accommodations to participate in University approved activities, including athletic or scholastics teams, musical and theatrical performances and debate activities. It is your responsibility to notify the instructor at least one week prior to missing any course assignment.

drops/withdrawals: You are responsible for completing the process of dropping or withdrawing from a course. Please click on the following link for more information on dropping and/or withdrawing from a course. Please consult the FAU Registrar Office for more information.

counseling and psychological services (CAPS) center: Life as a university student can be challenging physically, mentally and emotionally. Students who find stress negatively affecting their ability to achieve academic or personal goals may wish to consider utilizing FAU's Counseling and Psychological Services (CAPS) Center. CAPS provides FAU students a range of services – individual counseling, support meetings, and psychiatric services, to name a few – offered to help improve and maintain emotional well-being. For more information, go to *_http://www.fau.edu/counseling/*.

COVID-19 statement: All students in face-to-face classes are ~~required~~ (expected, recommended) to wear masks during class and at other times while indoors. Taking these measures supports the safety and protection of the FAU community. Students experiencing flu-like symptoms (fever, cough, shortness of breath), or students who have come in contact with an infected person should immediately contact FAU Student Health Services (561-297-3512)..

Additional references

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., & Chang, W. (2017). *rmarkdown: Dynamic documents for r* [Manual]. https://CRAN.R-project.org/package=rmarkdown

Alter, A. (2017). *Irresistible: The rise of addictive technology and the business of keeping us hooked.* Penguin.

Anscombe, F. (1973a). Graphs in statistical analysis. *American Statistician*, *27*(1), 17–21.

Bajak, A. (2017a). *Digital storytelling and social media, northeastern school of journalism, JRNL 3610* [Manual]. https://aleszu.github.io/digisoc/index.html

Bajak, A. (2017b). *How to convert a Google Doc to RMarkdown and publish on Github pages.* http://www.storybench.org/convert-google-doc-rmarkdown-publish-github-pages/

Baker, M. (2016). Is there a reproducibility crisis? *Nature*, *533*, 26.

Baumer, B., Cetinkaya-Rundel, M., Bray, A., Loi, L., & Horton, N. J. (2014). R Markdown: Integrating a reproducible analysis tool into introductory statistics. *ArXiv Preprint ArXiv:1402.1894*.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., ... Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. https://doi.org/10/cff2

Biology, A. S. for C. & others. (2015). *How can scientists enhance rigor in conducting basic research and reporting research results? A white paper from the American Society for Cell Biology.*

Bonomi, F., Milito, R., Zhu, J., & Addepalli, S. (2012). Fog computing and its role in the internet of things. *Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing*, 13–16.

Bravo, H. C. (2017). *Introduction to data science, u maryland CMSC 320* [Manual]. http://www.hcbravo.org/IntroDataSci/

Broman, K. W., & Woo, K. H. (2018). Data Organization in Spreadsheets. *The American Statistician*, *72*(1), 2–10. https://doi.org/10.1080/00031305.2017.1375989

Bryan, Jennifer. (2017). *Practical data science for stats—A peer j collection.*

Bryan, Jenny. (2017). *Data rectangling [2017 video].* https://www.rstudio.com/resources/videos/data-rectangling/

Bryan, Jenny. (2017). *Data wrangling, exploration, and analysis with r, UBC STAT 545A and 547M* [Manual]. https://github.com/STAT545-UBC

Bryan, Jenny, & TAs, the S. 545. (2017). *Happy git and GitHub for the useR.* http://happygitwithr.com/

Carmichael, I. (2017). *Introduction to data science, UNC STOR 390* [Manual]. https://idc9.github.io/stor390/

Clarke, R., Dorwin, D., & Nash, R. (2009). Is open source software more secure? *Homeland Security/Cyber Security.*

Cleveland, W. S., & McGill, R. (1985). Graphical Perception and Graphical Methods for Analyzing Scientific Data. *Science, New Series*, *229*(4716), 828–833.

Collaboration, O. S. & others. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Cox, J., & Lindell, M. (2013). Visualizing uncertainty in predicted hurricane tracks. *International Journal for Uncertainty Quantification*, *3*(2).

Donoho, D. (2017). 50 Years of Data Science. *Journal of Computational and Graphical Statistics*, *26*(4), 745–766. https://doi.org/10.1080/10618600.2017.1384734

FitzGerald, B., Levin, P. L., & Parziale, J. (2016). *Open source software & the department of defense.* Center for a New American Security.

Freeman, M., & Ross, J. (2017). *Technical foundations of informatics, u washington INFO 201.* https://info201.github.io

Gandrud, C. (2016). *Reproducible research with r and r studio.* CRC Press.

Gerla, M., Lee, E.-K., Pau, G., & Lee, U. (2014). Internet of vehicles: From intelligent grid to autonomous cars and vehicular clouds. *Internet of Things (WF-IoT), 2014 IEEE World Forum On*, 241–246.

Grange, J., Lakens, D., Adolfi, F., Albers, C., Anvari, F., Apps, M., Argamon, S., Baguley, T., Becker, R., Benning, S., & others. (2018). Justify your alpha. *Nature Human Behavior.*

Gutierrez, D. (2018). *How tidyverse guides r programmers through data science workflows.* https://opendatascience.com/how-tidyverse-guides-r-programmers-through-data-science-workflows/

Hathaway, J. (2017). *Data wrangling, exploration, and visualization, BYU-Idaho Math/CS 335* [Manual]. https://byuistats.github.io/M335/index.html

Healy, K. (2017). *Data visualization for social science: A practical introduction with r and ggplot2.* https://socviz.co

Hicks, S. C., & Irizarry, R. A. (2017). *Introduction to data science, BST 260* [Manual]. https://github.com/datasciencelabs/2017

Hicks, S. C., & Irizarry, R. A. (2018). A Guide to Teaching Data Science. *The American Statistician*, *72*(4), 382–391. https://doi.org/10/gfr5tf

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124.

Isaacson, W. (2014). *The Innovators: How a Group of Hackers, Geniuses, and Geeks Created the Digital Revolution.* Simon and Schuster.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 103). Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

Kumar, D., Wong, A., & Taylor, G. W. (2017). Explaining the unexplained: A class-enhanced attentive response (clear) approach to understanding deep neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 36–44.

Lakatos, I. (1969). Falsification and the methodology of scientific research programmes. *Criticism and the Growth of Knowledge. Cambridge University Press: Cambridge.*

Lanning, K. (1996). Robustness is not dimensionality: On the sensitivity of component comparability coefficients to sample size. *Multivariate Behavioral Research*, *31*(1), 33–46.

Lanning, K. (2018). *Data visualizations in personality and social psychology: Challenges in representing taxonomic, community, and developmental structures.*

Leek, J. T., & Peng, R. D. (2015). Statistics: P values are just the tip of the iceberg. *Nature*, *520*(7549), 612.

Loevinger, J. (1957). Objective Tests as Instruments of Psychological Theory. *Psychological Reports*, *3*(3), 635–694. https://doi.org/10.2466/pr0.1957.3.3.635

Loukides, Mike, M., Hilary, Patil, DJ Patil. (2018). *Ethics and data science.* O'Reilly.

Majumder, M. (2014). *Introduction to data science, nebraska DSC 256* [Manual]. http://mamajumder.github.io/data-science/fall-2014/

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., Glennerster, R., Green, D. P., Humphreys, M., Imbens, G., & others. (2014). Promoting transparency in social science research. *Science*, *343*(6166), 30–31.

Peng, RD, & Matsui, E. (2016). *The art of data science: A guide for anyone who works with data.*

Peng, Roger. (2018). *Teaching r to new users—From tapply to the tidyverse.* https://simplystatistics.org/2018/07/12/use-r-keynote-2018/

Persily, N. (2017). Can democracy survive the Internet? *Journal of Democracy*, *28*(2), 63–76.

R Core Team. (2017). *R: A language and environment for statistical computing* [Manual]. https://www.R-project.org/

Roger Peng, Brian Caffo, J. L. (2017). *Data science certificate program, johns hopkins university / coursera* [Manual]. https://github.com/DataScienceSpecialization/courses

Salganik, M. J. (2017). *Bit by bit: Social research in the digital age.* Princeton University Press.

Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach.* O'Reilly Media, Inc. https://www.tidytextmining.com/

Slovic, P., Zionts, D., Woods, A. K., Goodman, R., & Jinks, D. (2013). Psychic numbing and mass atrocity. *The Behavioral Foundations of Public Policy*, 126–142.

Stanton, J. M. (2013). *Introduction to data science.* CRC Press. https://archive.org/details/DataScienceBookV3

Stephens-Davidowitz, S. (2017). *Everybody lies: Big data, new data and what the Internet can tell us about who we really are.*

Sternberg, R. J. (1999). The theory of successful intelligence. *Review of General Psychology*, *3*(4), 292–316.

Storey, J. (2017). *Introduction to data science, princeton SML 201* [Manual]. https://github.com/SML201

Szucs, D., & Ioannidis, J. (2017). When null hypothesis significance testing is unsuitable for research: A reassessment. *Frontiers in Human Neuroscience*, *11*, 390.

Thies, J., Zollhöfer, M., Nießner, M., Valgaerts, L., Stamminger, M., & Theobalt, C. (2015). Real-time expression transfer for facial reenactment. *ACM Trans. Graph.*, *34*(6), 183–1.

Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.). Graphics Press.

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, *33*(1), 1–67.

Tukey, J. W. (1977). EDA: Exploratory data analysis. *Reading, Mass.*

Wickham, H. (2017). *Data challenge lab, stanford* [Manual]. https://github.com/dcl-2017-04/curriculum

Wilkinson, L. (2006). *The grammar of graphics.* Springer Science & Business Media.

Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Chapman and Hall/CRC. http://yihui.name/knitr/

Xie, Y. (2017a). *bookdown: Authoring books and technical documents with r markdown* [Manual]. https://CRAN.R-project.org/package=bookdown

Xie, Y. (2017b). *knitr: A general-purpose package for dynamic report generation in r* [Manual]. https://CRAN.R-project.org/package=knitr

Yandell, B. (2017). *R for data sciences, u wisconsin* [Manual]. https://github.com/datascience-uwmadison/R_for_data_sciences