DSCI 470: Research Methods in Data Science

Salisbury University · Fall 2026 · Section 001

On this page

Course information

Instructor

Randall E. Cone, Ph.D.

Email

recone@salisbury.edu

Office

Henson Science Hall 116

Office hours

2:00–3:30 p.m. Monday, Tuesday, Wednesday, Friday; and by appointment.

Class meetings

1:00–1:50 p.m. Monday, Tuesday, Wednesday, Friday

Room
Monday, Wednesday, Friday: Devilbiss Hall 213; Tuesday: Devilbiss Hall 109

Credits
4 credit hours. Expect roughly nine to twelve hours of work outside class each week.

Prerequisite
C or better in DSCI 318

Pre- or corequisite
COSC 311

Course description and purpose

Research and problem solving in data science, culminating in an original capstone research project (Project 2 below). The course covers:

- methodologies for ensuring that data used in problem solving is relevant and properly handled;
- the use of experimentation to build scientific knowledge;
- the components of sound experimentation and presentation;
- information literacy, including how to search for primary literature using library reference materials and online databases;
- writing reports and research papers;
- analyzing and presenting graphical data;
- the ethical use of information;
- presenting research to technical and non-technical audiences.

> **What this course is for:**
>
> DSCI 470 is the data science capstone. Project 1 builds the habits: Choosing a defensible question, Handling data honestly, and Defending a method under scrutiny. Project 2 is the culminating deliverable, an original research project carried from question to execution to presentation within this semester.
>
> Project 2 begins with a proposal. It must be approved before the work itself starts, the same way Project 1's dataset needs approval before you begin.

Recommended resources

- James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An Introduction to Statistical Learning with Applications in Python*. Springer.
  Free PDF available from the authors at statlearning.com. This is the same reference used in DSCI 318, so you already own it. It is our methods reference for Topics I through V.
- Müller, A. C., & Guido, S. (2016). *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly.
  Useful as a practical scikit-learn companion. Be aware that the scikit-learn API has changed substantially since publication; check the current documentation when code in the book does not run.

Assigned articles are listed in the reading list and posted in Canvas. All are available through SU Libraries or open access.

Computing environment

Class meets in Devilbiss Hall 213 on Monday, Wednesday, and Friday, and in Devilbiss Hall 109 on Tuesday. Both are standard classrooms. Bring your own laptop or tablet every session, in either room. Python and Jupyter notebooks are the primary platform.

Recommended: Install Miniforge, a free, fully open-source distribution of the conda package manager maintained by the conda-forge community. The installer is a graphical program on Windows and a shell script on macOS and Linux. Full instructions, including troubleshooting, are at the conda-forge/miniforge GitHub repository.

Once installed, use `conda` to install JupyterLab and the packages we'll use throughout the semester, with others added as needed:

- NumPy, pandas
- scikit-learn
- Matplotlib, seaborn, Plotly
- Git, for version control (introduced in Week 9)

Beginning in Week 9, every project deliverable must include an environment specification: A `requirements.txt` or `environment.yml` file. This lets a reader reconstruct the conditions under which your results were produced.

Grading

Grade weighting

Component	Weight
Homework 0 through 4	30%
Project 1 (written report)	20%
Project 2 (proposal 10%, report 20%, presentation 10%)	40%
Lab engagement and peer review	10%
Total	**100%**

Letter grades. A: At least 90. B: At least 80 and below 90. C: At least 70 and below 80. D: At least 60 and below 70. F: Below 60.

Course survey. A short survey is given in Week 1. It is not graded and does not affect your course grade. Its purpose is to show both of us what the class already knows so I can adjust the first several weeks.

Homework assignments

Unless stated otherwise, each homework solution is a Jupyter notebook containing:

- a narrative explanation of the research questions, the methods used, and the results;
- all code, commented;
- data and analysis visualizations;
- references and citations;
- a discussion of ethical considerations;
- an AI use disclosure, if you used AI tools.

Homework 0: The data science life cycle by example

Build a k-nearest-neighbor classifier using real scientific research data. Along the way you will get fluent with Jupyter, Python, pandas, scikit-learn, and Markdown. More importantly, you will work with imperfect data and get imperfect results. That is the normal condition of research, and it is what this course is actually about.

Homework 1: Ethics and integrity in data science

You will be given two articles on ethics and integrity in data science. Using SU Libraries' scholarly resources, locate a third article on ethics in science and read all three. Then write a paper of at least three pages, single spaced, 12-point type.

This is not a summary of what the articles say. It is your reasoned response to them, written as a formal report to an executive audience, someone who needs your judgment, not a book report. Take a position and defend it against the strongest objection you can construct.

Homework 2: Standardization and principal component analysis

Standardize a dataset's features, build a correlation coefficient matrix, and use principal component analysis to reduce the dataset's dimensionality. Compare classifier performance on the original and reduced feature sets, and discuss what that comparison shows about the relationship between variance retained and task performance.

Homework 3: Feature selection and evaluation discipline

Working with an assigned dataset containing both categorical and numerical variables, determine which features to include to improve the accuracy of a chosen baseline model. Your notebook must make clear how you separated model selection from model evaluation, and why that separation matters for the claim you end up making.

Homework 4: Developing a research question

Develop a data science question from a dataset of your choosing or from several provided. Explore the question through analysis and write up your findings. This assignment is deliberate preparation for Project 1: A question that survives this exercise is a candidate for your project.

Projects 1 and 2

The two projects share a structure and a set of standards. They differ in scope and in what you are asked to do with them:

- **Project 1** is a full research cycle on a dataset you select, submitted as a written report. It is graded on the written work alone.
- **Project 2** is a *new* research question. It may use the same dataset as Project 1 or a different one. Submit it as a written report and defend it in a 10-minute presentation. It begins with a proposal, described below, that must be approved before you proceed. The presentation is a separate graded component.

Select a dataset from a trusted source, approved by the DSCI 470 course professor, relevant to your Data Science Major track. Preprocess as necessary, then carry out visualization and statistical analysis in Python and Jupyter.

Explore at least two machine learning methods in service of your research questions; choose the methods because the questions call for them, not the reverse.

Report structure

1. **Title header.** Title, authors, collaborators.
2. **Introduction.** Narrative description of the project and its context.
3. **Data management.** Dataset description, provenance, and a data management and ethics plan.
4. **Research and methods.** Research questions and associated exploratory analysis, including an honest account of which techniques worked and which did not.
5. **Conclusion.** Summary of results and possible future directions.
6. **Bibliography.** Complete and annotated.
7. **Appendix.** AI use disclosure and environment specification.

Standard of the writing

Write for a professional audience that is intelligent but not expert in machine learning. Reports should be technical, with enough narrative exposition that a reader can follow the reasoning without already knowing the answer. For each research question, describe:

- its connection to the data;
- how completely you were able to answer it;
- where more work is needed;
- how your plans changed from your original intentions.

Ethical considerations should run through the report rather than sit in a section at the end.

Rubrics for both projects are posted in Canvas at the start of the semester. Read them before you begin, not after.

Project 2 proposal

Before starting substantial work on Project 2, submit a short proposal for approval. This works the same way Project 1's dataset approval does. The goal is to catch a doomed question or an unavailable dataset early, not after you have spent two weeks on it.

The proposal must include:

- the research question, stated so that someone could tell whether you had answered it;
- why the question matters and to whom;
- the data you intend to use and evidence that you can actually obtain it;

- a sketch of methods and why they suit the question;
- known risks and what you will do if the data or the method does not cooperate.

Submit it early. A proposal with time left to revise before the deadline is far more useful to you than one that arrives on the deadline itself.

Lab engagement and peer review

Friday is lab day all semester, in Devilbiss 213. It is homework studio through Week 7, and project studio from Week 8 on, once Project 1 gets underway. Monday, Tuesday, and Wednesday carry lecture content, with the balance shifting toward peer review and consultation as the semester moves toward the Project 2 proposal. This component makes the lab time accountable through three graded peer reviews:

1. **Peer review 1 (Week 12).** Written review of a classmate's Project 1 draft.
2. **Peer review 2 (Week 15).** Written review of a classmate's Project 2 proposal draft.
3. **Peer review 3 (final assessment period).** Structured written feedback on Project 2 presentations.

Each review is graded on the quality of the feedback, not on agreement with the author. A review that says only that the work is good receives no credit. Reviewing is a research skill, and it is the one most reliably neglected in undergraduate training.

Policy on generative AI

You may use generative AI tools in this course. You must disclose that use. You are responsible for everything you submit, without exception.

Permitted

- Brainstorming approaches, questions, or study designs
- Debugging code you wrote
- Explaining concepts and methods you are learning
- Editing prose you wrote yourself
- Locating candidate literature, provided you then read and verify the sources

Not permitted

- Generating Python code for homework or project submission

- Generating mathematical solutions for homework or project submission
- Submitting model-generated prose as your own analysis or interpretation
- Including a citation you have not personally located and read
- Using a model as a substitute for reading assigned material
- Any use you have not disclosed

Your obligations

1. Verify the accuracy, validity, and appropriateness of any content or citation a model produces, and correct errors before submission.
2. Check that model output has not reproduced substantial text from a source. Language models can and do reproduce copyrighted text, and doing so unknowingly is still plagiarism.
3. Cite and reference properly, following the guidance in this syllabus.

Disclosure format

If you used AI tools, append an AI use disclosure to your submission. If you did not, no disclosure is needed. Use a table with one row per distinct use:

Example AI use disclosure

Tool and version	What you asked it to do	Where it appears in your work	What you verified or changed
[Tool name, version, date accessed]	Explain why my cross-validation loop was leaking test data	Section 3, model evaluation code	Confirmed against scikit-learn docs; rewrote the fix myself after the suggested code failed on our data

Do not paste entire transcripts. I want a record you can defend, not a volume of text neither of us will read. If a particular exchange substantially shaped your thinking, quote the relevant part.

Late work

Homework. Accepted up to three calendar days late with a 10% reduction per day. After three days, homework is not accepted.

Projects, including the Project 2 proposal. Late submissions receive a 20% reduction for each class day late.

Excused work. If illness, a family emergency, a university-sponsored activity, or a religious observance affects a deadline, contact me as early as you can. Reach out before the deadline if possible. Official documentation is required for make-up work. Talking to me early almost always produces a better outcome than explaining afterward.

Attendance and classroom conduct

Plan to attend every class meeting. Come prepared, and bring your laptop or tablet. Friday lab sessions and the studio-heavy weeks later in the semester only work if you are present.

When classmates present, pay attention, behave respectfully, and ask questions. Peer feedback during presentations is part of your grade.

Devices. Use computing devices in ways that do not distract your classmates: No social media, video, or games during class. Always put phones away during class meetings unless I direct otherwise.

If you use a phone or other device for a medical, accessibility, or caregiving reason, you do not need to explain yourself. Examples include glucose monitoring, seizure alerts, hearing aid controls, and emergencies involving a dependent. You also do not need an accommodation letter to keep the device accessible. Let me know so I do not misread the situation.

Academic integrity and collaboration

Homework. You may collaborate with one other student on homework. You must write your own solution notebook, and you must name any collaborator in your title header.

Projects 1 and 2. Individual work. Do not collaborate. This applies to the Project 2 proposal as well.

Peer review. Reading and commenting on a classmate's draft is required, not collaboration. Writing part of their draft is.

The Salisbury University Academic Misconduct Policy is available from the Office of the Provost.

Course materials. Course notes and materials are for the use of students enrolled in this section. Ask me before sharing them outside the course.

Course schedule

The schedule is tentative and subject to change. Changes are announced in class and in Canvas as early as possible. Calendar dates below (semester start, holidays, and the final assessment) are confirmed against the department's Fall 2026 course schedule.

Class meets four days a week: Monday, Wednesday, and Friday in Devilbiss 213, and Tuesday in Devilbiss 109. Monday, Tuesday, and Wednesday carry lecture content; Friday is homework studio through Week 7 and project studio from Week 8 on.

Key dates

- Semester begins: Monday, August 31, 2026
- Add/drop deadline: Friday, September 4, 2026
- Labor Day, no class: Monday, September 7, 2026
- Fall Break, no class: Monday–Tuesday, October 19–20, 2026
- Last day to withdraw with a grade of W: Friday, November 6, 2026
- Thanksgiving recess, no class: Wednesday–Friday, November 25–27, 2026
- Last day of classes: Wednesday, December 9, 2026
- DSCI 470 final assessment: Wednesday, December 16, 2026, 4:15–6:45 p.m.

Daily schedule, Fall 2026

Date	Room (Devilbiss Hall)	Session	Due
Week 1: Aug 31 – Sep 4			
Mon 8/31	213	**Course launch.** Syllabus, expectations, and the data science life cycle by example.	N/A
Tue 9/1	109	ML Topic I: K-nearest neighbors, part 1. Intuition and distance metrics.	N/A
Wed 9/2	213	ML Topic I: K-nearest neighbors, part 2. Implementation in scikit-learn.	N/A
Fri 9/4	213	**Lab.** Environment setup: Miniforge, Jupyter, Git. Course survey.	Homework 0 assigned
Week 2: Sep 7 – Sep 11			
Mon 9/7	213	**No class: Labor Day.**	N/A

Date	Room (Devilbiss Hall)	Session	Due
Tue 9/8	109	ML Topic I: Evaluating a classifier. Accuracy, confusion matrices, and their limits.	N/A
Wed 9/9	213	ML Topic I wrap-up. Open questions before the ethics unit.	N/A
Fri 9/11	213	**Lab.** Homework 0 studio.	N/A
Week 3: Sep 14 – Sep 18			
Mon 9/14	213	Lecture: Research ethics and data ethics.	Homework 0 due
Tue 9/15	109	Human subjects research; when IRB review applies to data science work.	N/A
Wed 9/16	213	Case discussion.	N/A
Fri 9/18	213	**Lab: Library instruction.** Research ethics and information use.	Homework 1 assigned
Week 4: Sep 21 – Sep 25			
Mon 9/21	213	ML Topic II: Data standardization, part 1. Why scale matters; the Z-score formula.	N/A
Tue 9/22	109	ML Topic II, part 2. Correlation and the Pearson correlation coefficient.	N/A
Wed 9/23	213	ML Topic II: Principal component analysis. Eigenvectors, eigenvalues, and dimension reduction.	Homework 1 due; Homework 2 assigned
Fri 9/25	213	**Lab.** Homework 2 studio.	N/A
Week 5: Sep 28 – Oct 2			
Mon 9/28	213	ML Topic III: Optimization and gradient descent, part 1. Geometric intuition.	N/A
Tue 9/29	109	ML Topic III, part 2. Implementation and convergence.	N/A

Date	Room (Devilbiss Hall)	Session	Due
Wed 9/30	213	ML Topic III: Connecting gradient descent to least squares; worked examples.	Homework 2 due; Homework 3 assigned
Fri 10/2	213	**Lab.** Homework 3 studio.	N/A
Week 6: Oct 5 – Oct 9			
Mon 10/5	213	ML Topic IV: Model evaluation and resampling, part 1. Train/validation/test discipline.	N/A
Tue 10/6	109	ML Topic IV, part 2. Cross-validation; what a reported accuracy actually claims.	N/A
Wed 10/7	213	ML Topic IV: Worked examples and common mistakes.	Homework 3 due; Homework 4 assigned
Fri 10/9	213	**Lab.** Homework 4 studio.	N/A
Week 7: Oct 12 – Oct 16			
Mon 10/12	213	ML Topic V: Regularization and feature selection, part 1. The bias–variance tradeoff.	N/A
Tue 10/13	109	ML Topic V, part 2. Feature selection methods.	N/A
Wed 10/14	213	**Library instruction.** Bibliography, databases, and evaluating sources. Room: Academic Commons 113.	N/A
Fri 10/16	213	**Lab.** Homework 4 studio.	N/A
Week 8: Oct 19 – Oct 23			
Mon 10/19	213	**No class: Fall Break.**	N/A
Tue 10/20	109	**No class: Fall Break.**	N/A
Wed 10/21	213	ML Topic VI: Tree ensembles and model comparison; when machine learning is the wrong tool.	Homework 4 due

Date	Room (Devilbiss Hall)	Session	Due
Fri 10/23	213	**Lab.** Project 1 studio begins, opening with a brief look at data sources and how dataset approval works.	N/A
Week 9: Oct 26 – Oct 30			
Mon 10/26	213	**Lecture: Reproducibility and project architecture, part 1.** Project layout and data provenance. Reading: Wilson et al. (2017).	N/A
Tue 10/27	109	**Lecture, part 2.** Version control with Git; environment specification. Reading: Peng (2011).	N/A
Wed 10/28	213	Project 1 studio.	Dataset approval due
Fri 10/30	213	**Lab.** Project 1 studio.	N/A
Week 10: Nov 2 – Nov 6			
Mon 11/2	213	**Lecture: Analytic flexibility and the garden of forking paths, part 1.** Multiple comparisons and researcher degrees of freedom. Reading: Gelman & Loken (2014).	N/A
Tue 11/3	109	**Lecture, part 2.** Analysis plans as a remedy.	N/A
Wed 11/4	213	Project 1 studio.	N/A
Fri 11/6	213	**Lab.** Project 1 studio.	N/A
Week 11: Nov 9 – Nov 13			
Mon 11/9	213	**Lecture: Honest graphics.** Visual rhetoric, axis manipulation, uncertainty display, and the ethics of the persuasive chart.	N/A
Tue 11/10	109	**Lecture: Structuring a technical report for a non-expert reader.** Reading: Gopen & Swan (1990).	N/A
Wed 11/11	213	Project 1 studio.	N/A
Fri 11/13	213	**Lab.** Project 1 studio.	N/A

Date	Room (Devilbiss Hall)	Session	Due
Week 12: Nov 16 – Nov 20			
Mon 11/16	213	**Structured peer review workshop** on Project 1 drafts.	Peer review 1 due
Tue 11/17	109	Project 1 studio, final revisions.	N/A
Wed 11/18	213	Project 2 launched.	Project 1 report due
Fri 11/20	213	**Lab.** Project 2 planning: Dataset and question selection.	N/A
Week 13: Nov 23 – Nov 27			
Mon 11/23	213	**Lecture: Presenting technical work to a non-expert audience.** Reading: Fischhoff (2013).	N/A
Tue 11/24	109	Project 2 studio.	N/A
Wed 11/25	213	**No class: Thanksgiving recess.**	N/A
Fri 11/27	213	**No class: Thanksgiving recess.**	N/A
Week 14: Nov 30 – Dec 4			
Mon 11/30	213	Project 2 studio. Proposal drafting.	N/A
Tue 12/1	109	**Lecture: Scoping a capstone project, part 1.** Sizing the work to the time remaining in the semester.	N/A
Wed 12/2	213	Individual consultations on Project 2 scope.	Proposal draft for peer review due
Fri 12/4	213	**Lab.** Project 2 studio.	N/A
Week 15: Dec 7 – Dec 11			
Mon 12/7	213	**Lecture, part 2.** Recognizing when a project needs to change.	Peer review 2 due
Tue 12/8	109	Proposal workshop.	N/A
Wed 12/9	213	Project 2 studio, final polish.	Project 2 proposal due

Date	Room (Devilbiss Hall)	Session	Due
Fri 12/11	213	**Lab.** Wrap-up. Presentation logistics and Q&A.	N/A
Final assessment period			
Wed 12/16	213	Project 2 presentations, 4:15–6:45 p.m.	Project 2 report due Peer review 3 due

Accessibility and accommodations

Any student registered with the Disability Resource Center (DRC) who wants to use approved accommodations should contact me as soon as possible. We will arrange a meeting to coordinate them.

Students with disabilities can request reasonable accommodations, auxiliary aids and services, and modifications to University policies through the DRC by following the DRC's procedures for Requesting Accommodations, Modifications, Aids, and Services.

For more information, see the Disability Resource Center, email disabilitysupport@salisbury.edu, or visit the office at Guerrieri Student Union 229.

If any course material is inaccessible to you, tell me and I will fix it. Examples include a notebook that does not work with your screen reader, a figure without a usable description, or a video without captions. You do not need to be registered with the DRC to raise this.

Student support resources

Communication. Canvas and email are used extensively for announcements and course material. Check your SU email at least once a day.

Office hours. Office hours exist for your benefit. If you have questions about the course, about your capstone project, or about anything else, come by. You do not need a specific question prepared.

Writing support. The University Writing Center works with students at any stage of a draft, including technical writing. Bringing a project report there is a good use of your time.

Counseling. The SU Counseling Center provides confidential support to enrolled students at no cost. A senior year with a capstone in it is a demanding one; using this resource is ordinary, not remarkable.

Basic needs. If you have difficulty affording food or lack a stable place to live, SU has resources that can help. Contact the Dean of Students, or talk to me and I will help you find them.

Writing across the curriculum

The Department of Mathematics and Computer Science supports the University position that graduates should communicate clearly and correctly in all written work. All written work in this course, including papers, reports, notebook narratives, and peer reviews, is evaluated in part on spelling, punctuation, grammar, content, and organization.

Reading list

Assigned articles, listed in the order they appear in the schedule. All are available through SU Libraries or open access; PDFs and links are posted in Canvas.

- Peng, R. D. (2011). Reproducible research in computational science. *Science, 334*(6060), 1226–1227. https://doi.org/10.1126/science.1213847
- Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L., & Teal, T. K. (2017). Good enough practices in scientific computing. *PLOS Computational Biology, 13*(6), e1005510. https://doi.org/10.1371/journal.pcbi.1005510
- Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102*(6), 460–465. https://doi.org/10.1511/2014.111.460
- Gopen, G. D., & Swan, J. A. (1990). The science of scientific writing. *American Scientist, 78*(6), 550–558.
- Fischhoff, B. (2013). The sciences of science communication. *Proceedings of the National Academy of Sciences, 110*(Suppl 3), 14033–14039. https://doi.org/10.1073/pnas.1213273110

Two additional articles on ethics and integrity in data science are assigned for Homework 1 and posted in Canvas. You will locate a third yourself.

Learning outcomes

Course Learning Outcomes (CLO)

By the end of this course you will be able to:

1. Develop and complete an original data science capstone project, from proposal through execution and presentation.
2. Use library and scholarly information sources, and evaluate their reliability.
3. Select statistical and modeling methods appropriate to a given data science problem.
4. Apply appropriate problem-solving techniques to challenging data science problems.
5. Locate sources of data and process and manage them appropriately.
6. Present results ethically and responsibly, in a way that is not misleading.
7. Summarize and present results in a way that is reproducible.
8. Analyze, summarize, and present findings in multiple formats, with writing that supports the analysis.
9. Communicate with different audiences and guide readers through methods and results.

Experiential Learning Outcomes (ELO)

1. **Critical Thinking.** Students will be able to comprehensively analyze evidence before they create, critique, or accept an opinion or conclusion, or determine a need for further investigation.
2. **Oral Communication.** Students will be able to prepare, deliver, and reflect upon purposeful oral communication appropriate to the audience, purpose, and context.
3. **Written Communication.** Students will be able to develop and clearly express ideas through writing, in appropriate styles, incorporating evidence when warranted.
4. **Information Literacy.** Students will be able to determine the extent of information needed; access information effectively and efficiently; evaluate information and its sources critically; use information effectively to accomplish a specific purpose; and use information ethically.
5. **Intellectual Curiosity.** Students will explore a range of topics, be open minded to new ideas and ways of thinking, and be able to ask relevant questions or develop original thoughts.
6. **Ethical Reasoning.** Students will be able to reason about right and wrong human conduct; assess their own ethical values and the social context of problems; and recognize ethical issues in a variety of settings.

Outcome coverage by assessment

Which assessments address which outcomes

Assessment	Course Learning Outcomes	Experiential Learning Outcomes
Homework 0	4, 5, 7, 8, 9	1, 3, 5
Homework 1 (ethics paper)	2, 6, 8, 9	1, 3, 4, 6

Assessment	Course Learning Outcomes	Experiential Learning Outcomes
Homework 2 (PCA)	3, 4, 7	1, 3, 5
Homework 3 (feature selection)	3, 4, 5, 7	1, 3, 5
Homework 4 (question development)	3, 4, 5, 8, 9	1, 3, 5
Project 1	2, 3, 4, 5, 6, 7, 8, 9	1, 3, 4, 5, 6
Project 2 proposal	1, 3, 5, 9	1, 3, 4
Project 2 (report and presentation)	2, 3, 4, 5, 6, 7, 8, 9	1, 2, 3, 4, 5, 6
Lab engagement and peer review	4, 6, 9	1, 2, 5
