**STAT 41800**

**Computational Methods in Data Science**

## Contact Information

**Instructor:** **Dr. Timothy Reese** **email:** reese18@purdue.edu
**Office: MATH 210** **Phone:** 765-494-4129

**Lectures:** Tuesday/Thursday 1:30–2:45 PM · UNIV 127
**Office Hours:** Wednesday/Friday 1:00–3:00 PM · MATH 210
**Course Website:** https://treese41528.github.io/ComputationalDataScience/Website/index.html

## Course Description

**(Credit Hours: 3.00)** This course introduces essential computational methods in modern data science, focusing on simulation, resampling, Bayesian data analysis, and the utilization of large language models (LLMs) in data science workflows. Students will learn foundational simulation techniques, such as random variable generation through the inverse cumulative distribution function and rejection sampling. The course provides an overview of Frequentist and Bayesian inference, highlighting their theoretical foundations and practical applications. Key resampling methods, including bootstrapping and cross-validation, will be explored as tools for assessing variability, constructing confidence intervals, and validating predictive models. The course also emphasizes the practical and responsible use of LLMs in data science pipelines, including tasks such as text preprocessing, feature extraction, and leveraging pre-trained models to enhance data analysis and annotation. The course culminates in a capstone project where students will synthesize their learning by designing and implementing comprehensive solutions to real-world data science problems, demonstrating both theoretical understanding and practical implementation skills.

## Prerequisites:

Students are expected to have completed **multivariate calculus** (MA 26100), **mathematical probability** (one of MA/STAT 41600 or 5 MA/STAT 51600), **Python programming** (CS 38003), and have foundational knowledge of **statistical inference** (STAT 35000, STAT 35500, or STAT 51100). You should be comfortable with multiple integrals, partial derivatives, and vector calculus, as well as probability topics such as random variables, distributions, Bayes' theorem, and the Central Limit Theorem. Additionally, students should have experience in Python, including the use of core libraries (e.g., NumPy, Pandas) and basic data structures and algorithms (e.g., CS 18000). Understanding of statistical inference concepts such as hypothesis testing and confidence intervals is required. These prerequisites will provide the foundation necessary for the computational methods and data science applications explored in the course.

***This SYLLABUS is subject to change.***

**STAT 41800**

**Computational Methods in Data Science**

# Course Outline:

| Week | Tuesday | Thursday |
|---|---|---|
| **1** | Course introduction; Foundations of probability; Interpretations of probability; Inference paradigms | Python ecosystem; Distribution review; Distribution relationships |
| **2** | Monte Carlo fundamentals; Uniform random variates | Inverse CDF method; Transformation methods; Box–Muller |
| **3** | Rejection sampling | Variance reduction methods |
| **4** | Exponential families; Maximum likelihood theory | Statistical estimators; Sampling variability; Delta method |
| **5** | Least squares estimation; Assumptions and diagnostics | Least squares continued; Start generalized linear models (GLMs) |
| **6** | Generalized linear models (GLMs) continued | Sampling distribution problem; ECDF and plug-in principle |
| **7** | Nonparametric bootstrap | Nonparametric bootstrap continued; Start parametric bootstrap |
| **8** | **Midterm I (in class)** | Parametric bootstrap |
| **9** | Jackknife fundamentals (leave-one-out resampling) | Bootstrap confidence intervals; Bias correction |
| **10** | Spring Break | Spring Break |
| **11** | Cross-validation (LOO, k-fold); Model selection | Bayesian foundations; Prior specification |
| **12** | Conjugate priors; Analytical posteriors | Credible intervals |
| **13** | Markov chain foundations | MCMC: Metropolis-Hastings; Gibbs sampling |
| **14** | Convergence diagnostics; Model comparison | Posterior predictive checks; Hierarchical models |
| **15** | LLMs in data science: Integration; Text preprocessing; Embeddings | **Midterm II (Chapters 4–5)** |
| **16** | Pre-trained models; RAG systems | Responsible AI; Privacy; Prompt engineering |
| **17** | Capstone Presentations | |

***This SYLLABUS is subject to change.***

## Learning Resources

Lecture notes, code examples, assignments, and supplementary materials will be posted on the course website.

## Recommended Textbooks:

*There is no single textbook that covers all course topics in depth. Students seeking one comprehensive resource should start with:*

Efron, B., & Hastie, T. (2016). *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Cambridge University Press. https://doi.org/10.1017/CBO9781316576533

This text bridges classical frequentist methods, bootstrap and resampling, and Bayesian approaches, which partially mirrors the arc of the course.

For deeper study, the following topic-specific texts are recommended. Within each category, texts are ranked by accessibility and relevance to course material (★★★ = primary recommendation).

### Statistical Foundations and Inference Theory

★★★ Abramovich, F., & Ritov, Y. (2022). *Statistical Theory: A Concise Introduction* (2nd ed.). Chapman and Hall/CRC. Concise, modern treatment of estimation, hypothesis testing, and asymptotic theory. Best for building theoretical intuition.

### Monte Carlo and Simulation Methods

★★★ Robert, C. P., & Casella, G. (2004). *Monte Carlo Statistical Methods* (2nd ed.). Springer. https://doi.org/10.1007/978-1-4757-4145-2 The definitive reference for simulation techniques. Chapters 2–4 cover foundational methods used in Weeks 2–3.

### Resampling Methods

★★★ Efron, B., & Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. Chapman and Hall/CRC. https://doi.org/10.1201/9780429246593 The foundational text by the method's creators. Exceptionally clear exposition; essential reading for Weeks 6–8.

★★ **[Advanced]** Shao, J., & Tu, D. (1995). The Jackknife and Bootstrap. Springer. More theoretical treatment with rigorous asymptotic analysis. Recommended after Efron & Tibshirani.

### Bayesian Data Analysis

★★★ McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan* (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9780429029608 Outstanding pedagogical approach that builds intuition before formalism. Primary recommendation for Weeks 8–11.

***This SYLLABUS is subject to change.***

★★ Martin, O. A. (2024). *Bayesian Analysis with Python: A Practical Guide to Probabilistic Modeling* (3rd ed.). Packt Publishing. Practical implementation focus using PyMC. Excellent for translating Bayesian concepts into working Python code.

★★ **[Advanced]** Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (3rd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b16018 Comprehensive reference ("BDA3"). More encyclopedic; best used for specific topics or deeper theoretical study.

## Learning Outcomes

At the end of this course, students will be able to apply simulation techniques, including Monte Carlo methods, transformation approaches, and rejection sampling, to analyze probabilistic behavior in data science applications. They will compare and evaluate Frequentist and Bayesian inference paradigms by examining their theoretical foundations, identify their strengths and limitations, and explain their roles in statistical modeling and decision-making. Students will design, implement, and assess resampling methods, focusing on both nonparametric and parametric forms of the bootstrap, to estimate variability, construct confidence intervals, and improve statistical estimates through bias correction techniques. They will learn the principles of cross-validation, analyze its role in model assessment, and apply it to compute model performance metrics, detect overfitting and underfitting, and select models with reliable predictive accuracy using Python libraries.

Students will construct and interpret posterior distributions and credible intervals, apply Markov Chain Monte Carlo methods to approximate posteriors, and evaluate the role of prior distributions in Bayesian inference. In addition, students will utilize large language models in creative and practical ways within data science workflows, such as contextual data augmentation, feature engineering, and integrating structured and unstructured data to enhance predictive models, addressing challenges such as privacy and reliability. The course will culminate in a capstone project where students will synthesize selected course topics to design, develop, and present robust solutions to real-world data science challenges. By integrating computational methods, statistical principles, and real-world datasets, students will demonstrate their ability to create effective, data-driven solutions that showcase both theoretical understanding and applied expertise.

## Course Work and Requirements:

Final course grades are determined by the following weights.

| Category | Percentage |
|---|---|
| **Homework** | 40% |
| **MIDTERM Exams** | 15% x 2 |
| **Capstone Project** | 30% |

- **Homework (40%):** Homework will generally be posted on a Tuesday and due two weeks later. There will be **6 or 7 assignments** throughout the semester, with the **lowest homework score dropped**. Homework accounts for 40% of your course grade and will focus on applying course concepts, including coding and data analysis tasks.
- **Midterm Exams (15% each):** There will be **two midterm exams**, each contributing 15% to your final grade. These exams will assess both theoretical understanding and problem-solving skills related to the core topics covered in the first and middle portions of the course.

***This SYLLABUS is subject to change.***

- **Capstone Project (30%):** The course culminates in a capstone project, which will require students to design and implement a comprehensive solution to a real-world data science problem. Detailed project guidelines and requirements will be available on Brightspace.

### Late Policy:

- **Homework:** Homework assignments must be submitted by the due date listed in the assignment. Late submissions are permitted within a 3-day grace period but will incur a 20% penalty. Submissions beyond this window will not be accepted unless prior arrangements are made or in cases of documented emergencies.
- **Capstone Project:** The capstone project includes multiple submissions, such as a **proposal, progress report**, and **final deliverables**. Each component has its own deadlines, and late submissions may affect both the final grade and project evaluation. Details on submission dates and policies will be provided on Brightspace.

Students are encouraged to manage their time effectively and reach out early if extenuating circumstances arise.

### Policy for Make-up Exams:

The makeup exam format may vary from the standard exam, potentially involving questions necessitating python programming implementations or an oral examination.

Valid reasons for requesting a makeup exam include Purdue University required activities, direct exam conflicts, family bereavement, and medical grounds. Any other reasons will be considered on a case-by-case basis.

For non-emergency exam absences, contact me by email **reese18@purdue.edu** **at least one week** in advance of the exam. In the email include scheduling information regarding your courses and other important school activities.

For cases that fall under excused absence regulations, you or your representative should contact your instructor or go to the Office of the Dean of Students (ODOS) website to complete appropriate forms for instructor notification. Under academic regulations, excused absences may be granted by ODOS for cases of grief/bereavement, military service, jury duty, parenting leave, or emergent medical care.

## Academic Integrity

**Purdue Honors Pledge**

You are expected to uphold The Honor Code of Purdue University. The Purdue Honor Pledge is: "*As a boilermaker pursuing academic excellence, I pledge to be honest and true in all that I do. Accountable together - we are Purdue."* There is more information at https://www.purdue.edu/odos/osrr/honor-pledge/about.html.

Academic integrity is one of the highest values that Purdue University holds. Individuals, including students, are encouraged to alert university officials to potential breeches of this value by either emailing integrity@purdue.edu or by calling 765-494-8778. While information may be submitted anonymously, the more information that is submitted provides the greatest opportunity for the university to investigate the concern. More details are available on our course Brightspace table of contents, under University Policies.

***This SYLLABUS is subject to change.***

**STAT 41800**

**Computational Methods in Data Science**

Cheating on homework assignments will result in a zero for that assignment. Repeated infractions will lead to a formal report filed with the Office of Student Rights and Responsibilities (OSRR) (https://www.purdue.edu/odos/osrr/). Cheating includes but is not restricted to copying another person's work, allowing another person to copy your work, and copying previously posted keys or keys posted on the internet. It is considered cheating if any student uploads any course material to ANY website or shares the information electronically.

Cheating on exams will result in an automatic **"F"** in the course. This includes sharing details of the exam (questions, answers, or topics) with other students who have not yet taken it or using unauthorized materials, such as notes, electronic devices, or internet resources, while taking the exam.

This course emphasizes the application of computational methods in data science, including simulation, inference, and machine learning. Students will encounter real-world challenges that require critical thinking and problem-solving. While artificial intelligence (AI) tools, such as large language models (LLMs), can provide valuable support for debugging, studying, and exploring alternative approaches, students are expected to develop their skills independently and not rely excessively on AI assistance.

## Permitted Uses of AI Tools:

Students may use AI tools in the following ways:

- **Debugging:** AI may be used to identify errors or suggest improvements in code, provided the student understands the changes.
- **Study Assistance:** AI can aid in brainstorming ideas, clarifying complex theoretical concepts, and exploring multiple problem-solving approaches. This should supplement, not replace, individual study and critical thinking.
- **Resource Discovery:** AI tools may assist in finding additional learning materials or resources related to course topics, such as tutorials, documentation, and research papers.
- **Creative Exploration:** Students may use AI to enhance the presentation of their results (e.g., formatting visualizations, generating summaries), provided the core content and analysis are their own work.

**Warning:** AI tools may produce hallucinated content, including incorrect or fabricated code, explanations, or references. Always critically evaluate AI-generated responses and verify their accuracy before using them in your assignments or projects. These issues have been mitigated from the early models but the issue are still present to some degree.

## Prohibited Uses of AI Tools:

The following uses are considered violations of academic integrity:

- **Completing Assignments:** Submitting AI-generated code, analysis, or written responses without substantial modification and understanding is prohibited. All submitted work must reflect the student's original effort and comprehension.
- **Sharing AI-Generated Solutions:** Distributing AI-generated solutions to others in the course is not allowed.
- **Automating Capstone Project Tasks:** AI may support some tasks in the project (e.g., data preprocessing), but it should not replace core project requirements or be used without acknowledgment.

***This SYLLABUS is subject to change.***

- **Disclosure Requirement:** If AI tools are used to assist with coursework, students must disclose this in their submissions. Indicate how the tool was used (e.g., debugging suggestions, code optimization) and ensure that the final work is based on your understanding of the material.

**Misusing AI tools** in a way that compromises the learning process will be treated as academic dishonesty.

These policies are designed to ensure fairness and academic integrity while fostering the development of critical skills essential for future success. Mastering computational methods, statistical reasoning, and problem-solving requires genuine engagement with course material. Reliance on dishonest practices, including unauthorized use of AI or collaboration, undermines your learning and growth. If you have any doubts regarding acceptable collaboration or resource usage, consult the instructor for clarification. All violations will be handled in accordance with university procedures.

## Accessibility

Purdue University strives to make learning experiences as accessible as possible. If you anticipate or experience physical or academic barriers based on disability, you are welcome to let me know so that we can discuss options. You are also encouraged to contact the Disability Resource Center at: drc@purdue.edu or by phone: 765-494-1247.
If the Disability Resource Center (DRC) has determined reasonable accommodations that you would like to utilize in my class, you must release your Course Accommodation Letter to me. Instructions on sharing your Course Accommodation Letter can be found by visiting: https://www.purdue.edu/drc/students/course-accommodation-letter.php. Additionally, you are strongly encouraged to contact me as soon as possible to discuss implementation of your accommodations.

## Basic Needs Statement

If you are facing challenges securing basic needs such as food, housing, transportation, health services, or access to technology or childcare resources and believe this may affect your performance in the course, please contact the Office of the Dean of Students (ODOS) to help coordinate with community resources. These services vary by location. In **West Lafayette**, see the Basic Needs Program website, or email basicneeds@purdue.edu. To connect with a Student Support Generalist on the **Indianapolis** campus, contact them by phone at 765-495-7797 or email studentlifeindy@purdue.edu.

## Nondiscrimination Statement

Purdue University is committed to maintaining a community which recognizes and values the inherent worth and dignity of every person; fosters tolerance, sensitivity, understanding, and mutual respect among its members; and encourages each individual to strive to reach his or her potential. In pursuit of its goal of academic excellence, the University seeks to develop and nurture diversity. The University believes that diversity among its many members strengthens the institution, stimulates creativity, promotes the exchange of ideas, and enriches campus life. A hyperlink to Purdue's full Nondiscrimination Policy Statement is included in our course Brightspace under University Policies.

## Violent Behavior Policy

Any student who has substantial reason to believe that another person is threatening the safety of others is encouraged to report the behavior to and discuss the next steps with their instructor. Students also have the

option of reporting the behavior to the Office of the Student Rights and Responsibilities. See also Purdue University Bill of Student Rights and the Violent Behavior Policy under University Resources in Brightspace.

## Emergency Preparedness

In the event of a major campus emergency, course requirements, deadlines and grading percentages are subject to changes that may be necessitated by a revised semester calendar or other circumstances beyond the instructor's control. Relevant changes to this course will be posted on Brightspace or can be obtained by contacting your instructor via email at reese18@purdue.edu. You are expected to read your @purdue.edu email on a frequent basis.

There is additional information concerning emergency preparedness in the syllabus module on Brightspace.

## Mental Health/Wellness Statement

**If you're struggling and need mental health services: Purdue University is committed to advancing the mental health and well-being of its students.** If you or someone you know is feeling overwhelmed, depressed, and/or in need of mental health support, services are available. For help, such individuals should contact Counseling and Psychological Services (CAPS) at 765-494-6995 during and after hours, on weekends and holidays, or by going to the CAPS office on the second floor of the Purdue University Student Health Center (PUSH) during business hours.

**If you find yourself beginning to feel some stress, anxiety and/or feeling slightly overwhelmed,** try Therapy Assistance Online (TAO), a web and app-based mental health resource available courtesy of CAPS**.** TAO is available to you at any time by creating an account on the TAO Connect website, or downloading the app from the App Store or Google Play. It offers free, confidential well-being resources through a self-guided program informed by psychotherapy research and strategies that may aid in overcoming anxiety, depression, and other concerns. It provides accessible and effective resources including short videos, brief exercises, and self-reflection tools.

**If you need support and information about options and resources**, please contact or see the Office of the Dean of Students. Call 765-494-1747. Hours of operation are M-F, 8 a.m.- 5 p.m.

**If you find yourself struggling to find a healthy balance between academics, social life, stress**, etc., sign up for free one-on-one virtual or in-person sessions with a Purdue Wellness Coach at RecWell. Student coaches can help you navigate through barriers and challenges toward your goals throughout the semester. Sign up is free and can be done on BoilerConnect.

## Emergency Preparedness

In the event of a major campus emergency, course requirements, deadlines and grading percentages are subject to changes that may be necessitated by a revised semester calendar or other circumstances beyond the instructor's control. Relevant changes to this course will be posted on Brightspace or can be obtained by contacting your instructor via email at reese18@purdue.edu. You are expected to read your @purdue.edu email on a frequent basis.