STAT 432: Basics of Statistical Learning

Introduction

Ruoqing Zhu

teazrq.github.io/stat432rpy

Fall 2026

University of Illinois at Urbana-Champaign

Course Information

Overview

1. Basic course information
2. Topics and learning objectives
3. What changes this semester
   - discussion sessions and weekly questions;
   - completion-based homework and four quizzes;
   - AI tools that may support course workflows; and
   - Week 1 setup with AI tools, skills, and GitHub.

Basic Course Information

- Course: **STAT 432, Basics of Statistical Learning**
- Semester: **Fall 2026**
- Instructor: **Ruoqing Zhu**
  - Email: rqzhu@illinois.edu; office: 137 CAB; office hours: TBA
- Teaching assistant: **Chenyu Zhang**
  - Email: chenyuz9@illinois.edu; office hours: Monday, 2:00 to 4:00 p.m. on Zoom
- Meeting days: **Tuesday and Thursday**
- Class meeting: **12:30 to 1:50 p.m.**, 2039 Campus Instructional Facility
- Course website: teazrq.github.io/stat432rpy

The course website is the source of truth for schedules, policies, materials, and announcements.

Course Materials and References

The lecture notes are the main guide. The books below provide additional explanations and examples.

SMLR Ruoqing Zhu, *Statistical Learning and Machine Learning with R*

ISL Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, *An Introduction to Statistical Learning*

- R and Python editions are available from the book website.

ESL Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning*

- This is a more advanced reference.

Prerequisites

You should be comfortable with the following ideas at an introductory level.

- Probability: random variables and common distributions
- Statistics: estimators, likelihood, and linear regression
- Mathematics: matrix operations, vectors, and derivatives
- Computing: basic programming in either R or Python

You do not need to remember every formula. You should be willing to review prerequisite material when it appears in a new method.

Topics and Objectives

What Does Statistical Learning Involve?

For every method, we will ask a common set of questions.

- What scientific, explanatory, or prediction question are we trying to answer?
- How were the data collected, and what information do the variables contain?
- What model, algorithm, or optimization problem are we using?
- How flexible is the model, and what is its computational cost?
- How should tuning choices and competing methods be evaluated?
- How should the result, uncertainty, and limitations be interpreted and communicated?

The Traditional Learning Settings

- **Supervised learning**
  - A response is observed in the training data.
  - Regression predicts a quantitative response.
  - Classification predicts a class label or class probability.
- **Unsupervised learning**
  - There is no designated response variable.
  - The goal is to find structure, groups, or useful representations.

Many modern problems combine these settings rather than fitting perfectly into one category.

Supervised-Learning Topics

- Training and testing error, model selection, and variable selection
- Linear models, ridge regression, lasso, and numerical optimization
- Logistic regression, discriminant analysis, and related classification methods
- Nearest neighbors, kernel regression, and local polynomial regression
- Support vector machines
- Classification and regression trees, random forests, and boosting

Throughout these topics, we will compare model flexibility, computation, tuning, evaluation, and interpretation.

Unsupervised Learning and Core Statistical Ideas

- Principal component analysis and low-dimensional representations
- K-means, hierarchical clustering, graph methods, and spectral clustering
- Bias-variance trade-off and prediction error
- Cross-validation, resampling, and statistical simulation
- Connections to causal inference and selected topics in artificial intelligence

The exact pace may change as we spend more time on ideas that require additional explanation or discussion.

Course Roadmap: Weeks 1–8

- **Week 1:** Course introduction, AI agents, skills, and GitHub
- **Week 2:** Training and testing error
- **Week 3:** Ridge regression, lasso, and optimization
- **Week 4:** Nearest neighbors and low-dimensional embedding
- **Week 5:** Principal component analysis and discriminant analysis
- **Week 6:** Kernel smoothing, local polynomial regression, and bias-variance trade-off
- **Week 7:** Support vector machines
- **Week 8:** Trees and random forests

Course Roadmap: Weeks 9–16

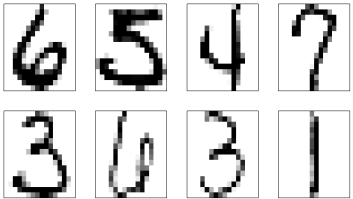
- **Week 9:** AdaBoost and gradient boosting
- **Week 10:** K-means and hierarchical clustering
- **Week 11:** Spectral clustering
- **Week 12:** Causal inference
- **Week 13:** Selected topics in artificial intelligence
- **Week 14:** Fall Break
- **Week 15:** Final Presentations I
- **Week 16:** Final Presentations II

What Should You Learn?

By the end of the course, you should be able to:

- explain why a statistical-learning method is constructed in a particular way;
- connect mathematical results to the behavior of a fitted model;
- fit, tune, evaluate, and interpret models in R or Python;
- recognize how data collection and domain knowledge affect an analysis;
- use computation to investigate a statistical idea; and
- communicate your reasoning in writing, code, and discussion.

Use Case 1: Handwritten Digit Recognition



Each image is recorded as a grid of grayscale pixel values.

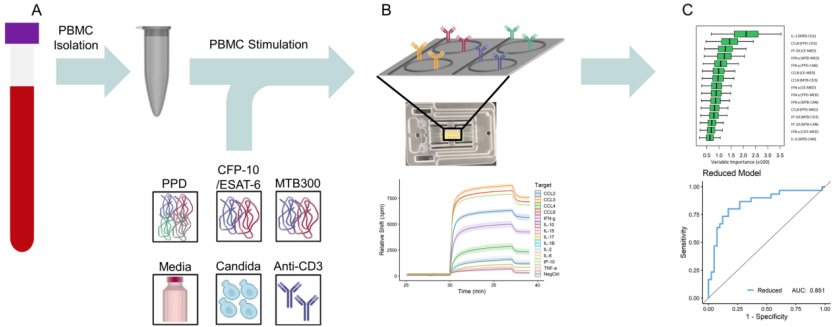
Given the pixels, which digit is shown?

- **Data:** many correlated pixel measurements
- **Task:** learn one of ten labels from example images
- **Challenge:** handwriting varies and the class boundaries are nonlinear

Statistical lesson: a useful model must work on new images, not only the images used to fit it.

Image adapted from Hastie, Tibshirani, and Friedman, *The Elements of Statistical Learning*.

Use Case 2: Biomarker Selection for Latent Tuberculosis

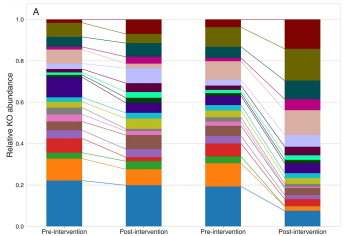


DATA 65 participants, 143 candidate features ⟶ **MODEL** random-forest feature selection ⟶ **CHECK** ROC curves

Statistical lesson: measuring more variables does not automatically produce a better model.

Robison et al. (2021), *Scientific Reports* 11, 20544. doi:10.1038/s41598-021-99754-3.

Use Case 3: Predicting Food Intake from the Microbiome



Functional microbial features before and after an almond intervention.

Can a biological profile predict food intake?

- **Data:** metagenomic features from five controlled feeding studies
- **Model:** random forests classify treatment and control groups
- **Evaluation:** 80% accuracy for almond, 87% for broccoli, and 86% for walnut

Statistical lesson: observations not used for fitting provide evidence about predictive performance.

Shinn et al. (2024), *The Journal of Nutrition* 154, 271-283. doi:10.1016/j.tjnut.2023.11.001.

What Changes This Semester

What Changes This Semester?

- Tuesday classes are organized around student questions and group discussion.
- Weekly homework is required, but evaluated only as Complete or Incomplete.
- Each student submits one weekly discussion question through GitHub.
- Four short quizzes assess completed discussion blocks. The first covers Weeks 1 and 2; each later block covers three weeks.
- Students must have an AI coding agent available. Its use is optional, but several course workflows are easier with one.

The goal is to spend less class time watching completed work and more time explaining, questioning, and checking statistical reasoning.

The Weekly Learning Cycle

1. **Lecture:** study mathematical foundations, selected derivations, examples, and practical issues.
2. **After the lecture:** read the remaining notes, complete the weekly homework, and submit one discussion question.
3. **Monday:** the instructor selects a set of submitted questions to form the question pool for Tuesday.
4. **Tuesday:** four to eight questions are posted. Students move from **Group Discussion** to **Present and Challenge**, where teams present, peers challenge answers, and the class rates presenting teams and challengers.

On four selected Tuesdays, the class begins with a 30-minute quiz. A short lecture may follow the discussion if time permits.

Discussion Sessions

Tuesday Discussion Sessions

- The 25% classroom discussion component is based entirely on participation.
- Discussion sessions meet weekly through Tuesday, November 17, 2026.
- Each session covers the previous Thursday's topic and usually lasts 50 to 60 minutes.
- You may miss up to one discussion session without losing participation credit.
- Teams may be assigned randomly or formed voluntarily, depending on the activity.
- Each session moves from **Group Discussion** to **Present and Challenge**.

Using Popping in Tuesday Class

GitHub is used to submit your question before class. Popping is used for the in-class discussion.

- Course login: `popping.onrender.com/login/432fall2026`
- Login name: your **NetID**
- Password: check the **Popping Password** column in the Canvas gradebook.
- If an in-class network or technical problem prevents a required action, tell the instructor. You will receive full participation credit for that session.

Participation Actions and Voting Privacy

To receive full participation credit for a discussion session:

- During **Group Discussion**, give a thumbs-up to at least one teammate.
- During **Present and Challenge**, rate every presenting team and every challenger.
- Your individual voting record will remain private and will never be released.
- Aggregate results, including the top-rated individuals or teams, may be shared with the class.

Present and Challenge

- **Presenters** explain the group's conclusion, the main reasoning, and any assumptions or limitations.
- **Challengers** raise an important or interesting question to the presenting team.
  - A useful challenge may identify a missing assumption, question an interpretation, suggest a counterexample, or connect the problem to another method.
- The purpose of a challenge is to deepen the discussion, not simply to make the presentation harder.
- The instructor closes the discussion by connecting the main ideas and correcting unresolved errors.

Assessment

Assessment and Grading

Component	Weight
Weekly homework, Complete or Incomplete	15%
Weekly GitHub discussion questions	10%
Classroom discussion participation	25%
Four quizzes	25%
Final project	25%
Total	**100%**

Letter grades are based on the final course percentage. Regular letter-grade thresholds and the A+ threshold are to be announced.

Weekly Discussion Bonus Points

Each week, bonus points will be awarded to:

- every member of the highest-rated presenting team; and
- the highest-rated challenger.

The point values, tie rules, and other details will be announced later.

Homework

Weekly Homework Is Completion-Based

- Homework contributes 15% of the course grade.
- A homework is recorded as **Complete** or **Incomplete**; individual answers are not scored for correctness.
- Most questions are adapted from STAT 432 homework from previous years.
- The questions review mathematical ideas, implementation, interpretation, and common statistical difficulties.
- Solutions are provided so that you can check your reasoning and correct your work.

The six Week 1 foundation questions are **Homework 01** and follow the same completion and submission rules as later homework.

Homework Submission

- Submit `homework-NN.pdf` or `homework-NN.html` in `homework/week-NN/` through your `stat432-fall2026` GitHub repository.
- Use PDF or self-contained HTML. The report may not exceed 10 pages when printed.
- Submit by **Sunday at 11:59 p.m. Central Time** before the Tuesday discussion.
- The file version and timestamp shown on GitHub determine the submission.
- A Complete report opens correctly, meets the format and page limit, and makes a substantive attempt at every required question. An Incomplete report earns 0.
- For up to three homework assignments, you may use a two-day extension without penalty. The extended cutoff is **Tuesday at 11:59 p.m. Central Time**.

How Should You Use the Homework?

1. Read the question and identify what statistical idea it is checking.
2. Think about how you would begin before asking for help or opening the solution.
3. Work on the question yourself, use an AI agent, or combine both approaches.
4. Review the completed answer and decide whether you can explain the reasoning and interpretation.

Using AI to complete the homework is allowed. You must judge whether the way you use it is helping you learn the material needed for discussion, quizzes, and the final project.

Quizzes

Four In-Class Quizzes

- Four quizzes during the semester
- Each quiz lasts **30 minutes**. Quiz 1 has **8 questions**; Quizzes 2 through 4 have **12 questions** each.
- Questions may be short answer or multiple choice.
- Discussion blocks: Weeks 1 and 2; Weeks 3 through 5; Weeks 6 through 8; and Weeks 9 through 11.
- Dates: September 15, October 6, October 27, and November 17.
- Together, the quizzes contribute 25% of the course grade.
- Every quiz is completely closed book and completed independently.
- Books, notes, websites, AI tools, and help from another person are not allowed.

Where Do Quiz Questions Come From?

- Each previous discussion contributes four questions: three appear exactly as covered in class, and one comes from other student submissions.
- Quiz 1 contains six discussion questions and two other submitted questions, for eight questions total.
- Quizzes 2 through 4 contain nine discussion questions and three other submitted questions, for 12 questions total.
- 75% of the quiz questions come exactly from questions selected for previous discussion sessions; the remaining 25% come from other student-submitted questions that were not selected for live discussion.

A good way to prepare is to reconstruct the reasoning behind the selected questions and be ready to explain what your own question is asking and why it matters.

Discussion Questions

Submit One Weekly Discussion Question

- Weekly GitHub question submission contributes 10% of the course grade.
- Submit one focused question connected to the current topic.
- Include your own current thinking, not only a question copied from a note or generated by AI.
- Store the question in your `stat432-fall2026` GitHub repository.
- Use `discussion/week-NN-question.md`, with a two-digit week number.
- Keep the prose body to at most 200 words. Fenced code and LaTeX mathematics do not count toward this limit.

Deadline and Submission Check

- Submit by **Sunday at 11:59 p.m. Central Time** before the upcoming Tuesday discussion.
- Review and proofread the complete file.
- Commit the file, push the commit, and verify the version shown on GitHub.
- The file version and timestamp shown on GitHub determine the submission.
- A version posted after the deadline receives 0 points. Homework extensions do not apply to discussion questions.

The complete question format is introduced during Week 1.

GitHub and the Optional Student Skill

In Week 1, we will complete this workflow together.

1. Create and clone a repository named exactly `stat432-fall2026`.
2. If useful, install the portable `stat432-fall2026-student` skill and use it to check the correct two-digit week file.
3. Read the question carefully and revise anything that does not represent your own thinking.
4. Approve the commit and push, then verify the file on GitHub.

The skill checks the repository structure, question format, and workflow. It does not decide what you genuinely want to ask.

Final Project

Final Project

The final project contributes 25% of the course grade.

- Work individually or in a team of at most three students.
- Select a dataset and define a meaningful question yourselves.
- Connect the question, data, statistical reasoning, computation, and interpretation.
- AI-agent use is optional. If an agent is used, document its contributions clearly.
- Check the analysis carefully and explain important limitations.
- Graduate students present their projects during Weeks 15 and 16.

The report format, milestones, deadlines, and evaluation details are to be announced.

One Possible AI-Integrated Project Design

One possible project is to build a reusable skill that instructs an AI agent to:

- inspect and validate a dataset;
- carry out a justified statistical analysis;
- perform model diagnostics and explain what constitutes a good or poor fit;
- generate appropriate tables and figures;
- check its own results; and
- produce a reproducible report.

Your work is to define the question, design and test the skill, choose the statistical checks, audit the generated report, and explain the conclusions and limitations.

This is an example, not the only acceptable project format.

Coming Up in Week 1

Coming Up in Week 1

The next Week 1 lecture is a guided setup session.

- Make at least one coding agent available, such as Codex, Claude Code, or an agent available through VS Code.
- If useful, open the agent's chat interface and let it inspect your Git and GitHub setup.
- Create and clone `stat432-fall2026`.
- Optionally install and invoke the course discussion-question skill.

We will work through these steps together. Bring your computer and keep any exact error message that appears.

Your First Required Discussion Question

- Use the course skill if helpful, or check the Week 1 question format manually.
- Proofread the complete question, then commit, push, and verify it on GitHub.
- This is the first required discussion question and counts toward the 10% weekly-question component.
- The following Tuesday is the guided first discussion session. There is no quiz in that session.

Questions