

Introduction to IS 382

Predictive Modeling

Jason Bryer, Ph.D.

Fall 2026

Agenda

- Syllabus
- Class meetups
- Course Schedule
- Assignments (how you will be graded)
 - Participation
 - Labs
 - Data Project
 - Exams
- Introduction to predictive modeling

Syllabus and course materials are here: <https://fall2026.is382.net>

The site is built using [Quarto](#) and hosted on [Github](#). Each page of the site has a "Edit this page" link at the bottom right, use that to start a pull request on Github.

We will use Brightspace primary for submitting assignments only. Please submit a PDF.

PDFs are preferred for the homework as there is some LaTeX formatting in the R markdown files. The `tinytex` R package helps with install LaTeX, but you can also install LaTeX using [MiKTeX](#) (for Windows) and [BasicTeX](#) (for Mac) See this page for more information:

<https://fall2026.is382.net/course-overview/software/>

Meetups

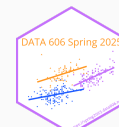
We will have meetups on Monday evenings from 7:00pm to 8:00pm.

Meetups will be recorded and made available the next day on the [course website](#).

Though attending live is not strictly required, **I expect everyone to watch the lectures during the week.** I use the class meetups to convey important information and announcements. Very often I will cover some topics not in the textbook. Students who attend the meetups tend to do well on the assignments.

One Minute Papers - Complete the one minute paper after each Meetup (whether you watch live or watch the recordings). It should take approximately one to two minutes to complete. This allows me to 1) verify you have attended/watch the meetup and 2) get feedback about what you learned and what you may still be unclear.

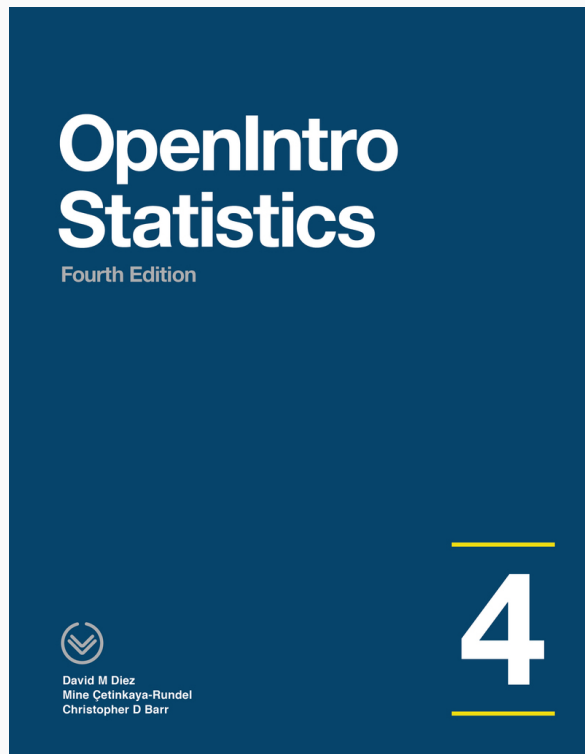
Please note: *Students who participate in this class with their camera on or use a profile image are agreeing to have their video or image recorded solely for the purpose of creating a record for students enrolled in the class to refer to, including those enrolled students who are unable to attend live. If you are unwilling to consent to have your profile or video image recorded, be sure to keep your camera off and do not use a profile image. Likewise, students who un-mute during class and participate orally are agreeing to have their voices recorded. If you are not willing to consent to have your voice recorded during class, you will need to keep your mute button activated and communicate exclusively using the "chat" feature, which allows students to type questions and comments live. [Click here for CUNY's camera use policy](#)*



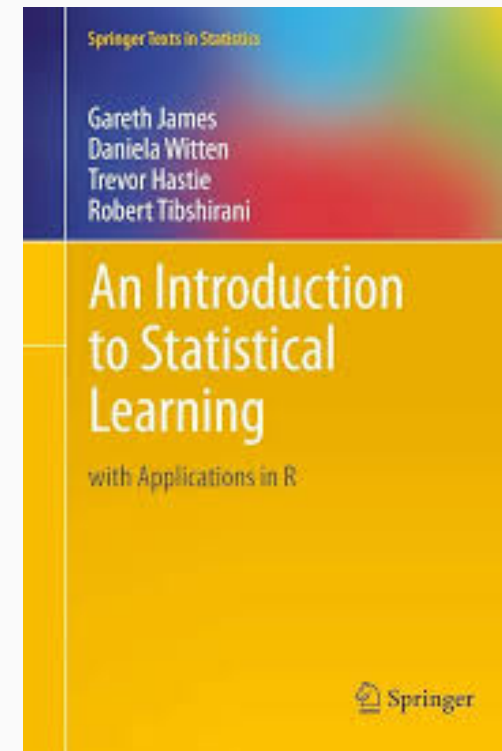
Schedule

Week	Topic
1	Introduction to the Modeling
2	Linear regression
3	Multiple regression
4	Maximum Likelihood Estimation and Logistic Regression
5	Generative Models for Classification
6	Resampling Methods
7	Tree Based Methods
8	Bagging and Random Forests
9	Support Vector Machines
10	Principal Component Analysis
11	Clustering Methods
12	Project Proposal
	No Meetup - Thanksgiving
13	Missing data
14	Bayesian Analysis
15	Wrap up / Final Presentations

Diez, D.M., Barr, C.D., & Çetinkaya-Rundel, M.
(2019). *OpenIntro Statistics* (4th Ed).



James, G., Witten, D., Hastie, T., & Tibshirani, R.
(2021). *An Introduction to Statistical Learning
with Applications in R* (2nd Ed).



Assignments

- **Weekly labs** (50%). They are available on the course schedule page:
fall2026.is382.net/schedule.html
- **Data Project** (25%). We will cover this more as the semester progresses, but is similar to what you did for IS381.
- **Final Exam** (15%). This will be completed on Brightspace during the last week of the semester.
- **Participation** (10%). Attending/watching the weekly meetups and completing the one minute papers.

Communication

- Slack Channel: <https://cuny-msds.slack.com>
 - [Click here to join the group](#)
- Email: jason.bryer@cuny.edu
- Phone/Zoom: Please email to schedule a time to meet.
- Office hours by appointment.

This is an applied statistics course so we will make extensive use of the **R statistical programming language**.

Install **R** and **RStudio** on your own computer. I encourage everyone to do this at some point by the end of the semester. I have instructions on the course website here:

<https://fall2026.is382.net/course-overview/software/>

You will also need to have **LaTeX** installed as well in order to create PDFs. The **tinytex** R package helps with this process:

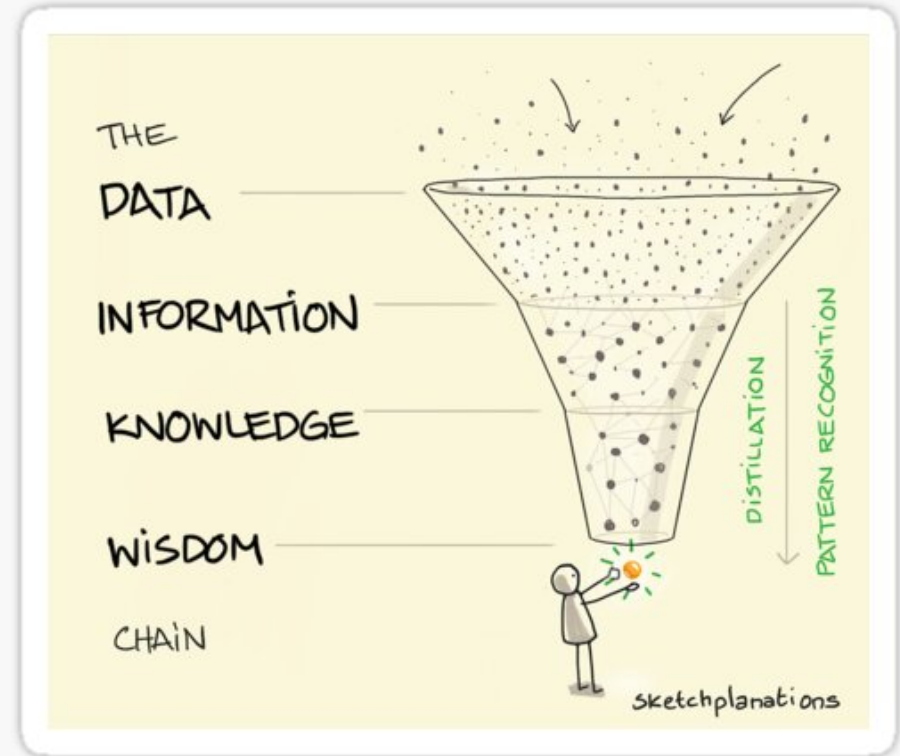
```
install.packages('tinytex')
tinytex::install_tinytex()
```

Introduction to Predictive Modeling

Learning...

Where is the Life we have lost in living?
Where is the wisdom we have lost in
knowledge?
Where is the knowledge we have lost in
information?

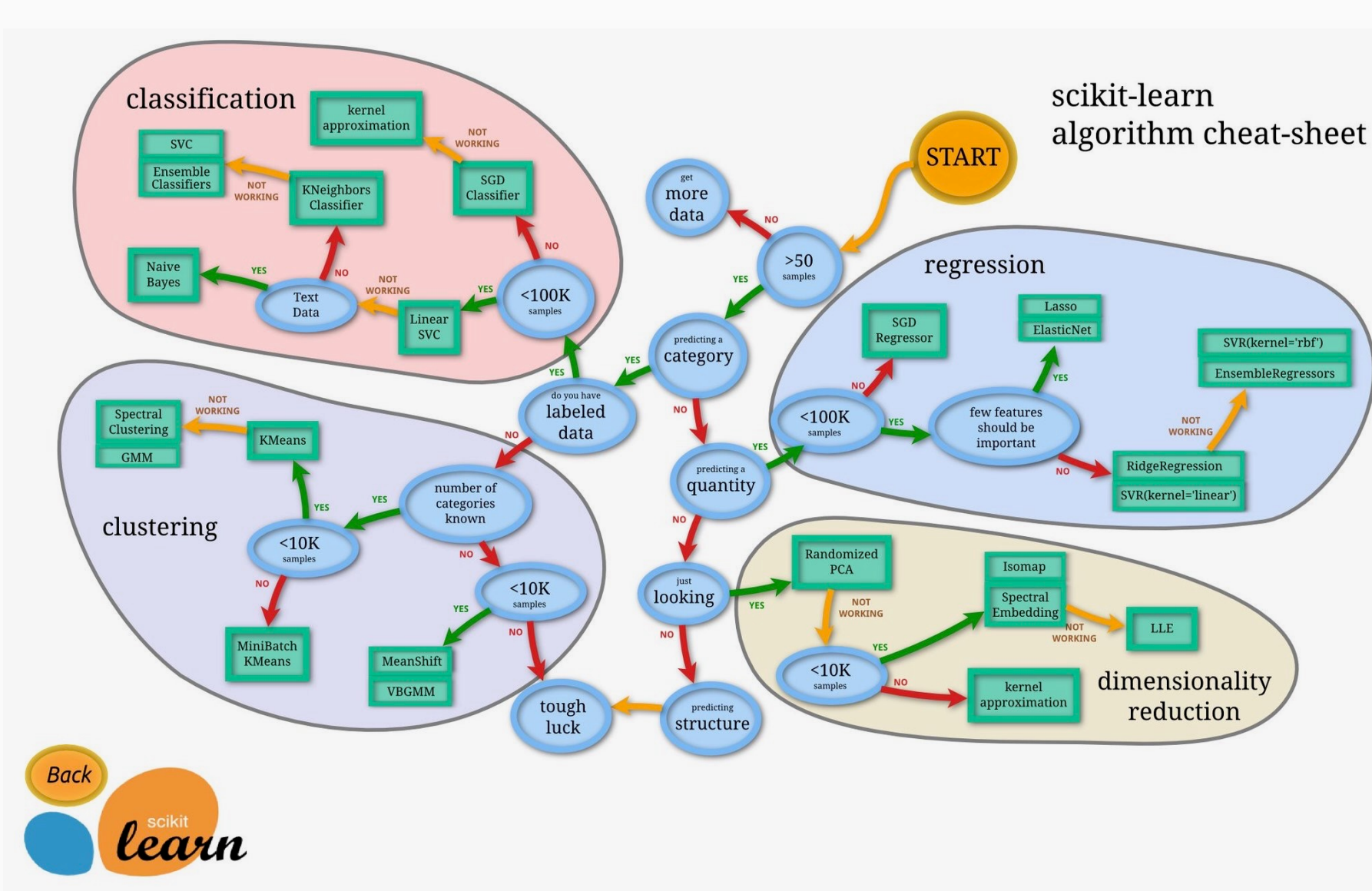
Excerpt from *The Rock* by T.S. Eliot



Data Science Landscape



Types of machine learning algorithms



Supervised vs. Unsupervised Learning

Supervised Learning

- Starting point:
 - Outcome measurement Y (also called dependent variable, response, target).
 - Vector of p predictor measurements X (also called inputs, regressors, covariates, features, independent variables).
 - We have training data we can use to estimate the model.
- Objectives:
 - Accurately predict unseen test cases.
 - Understand which inputs affect the outcome, and how.
 - Assess the quality of our predictions and inferences.

Unsupervised Learning

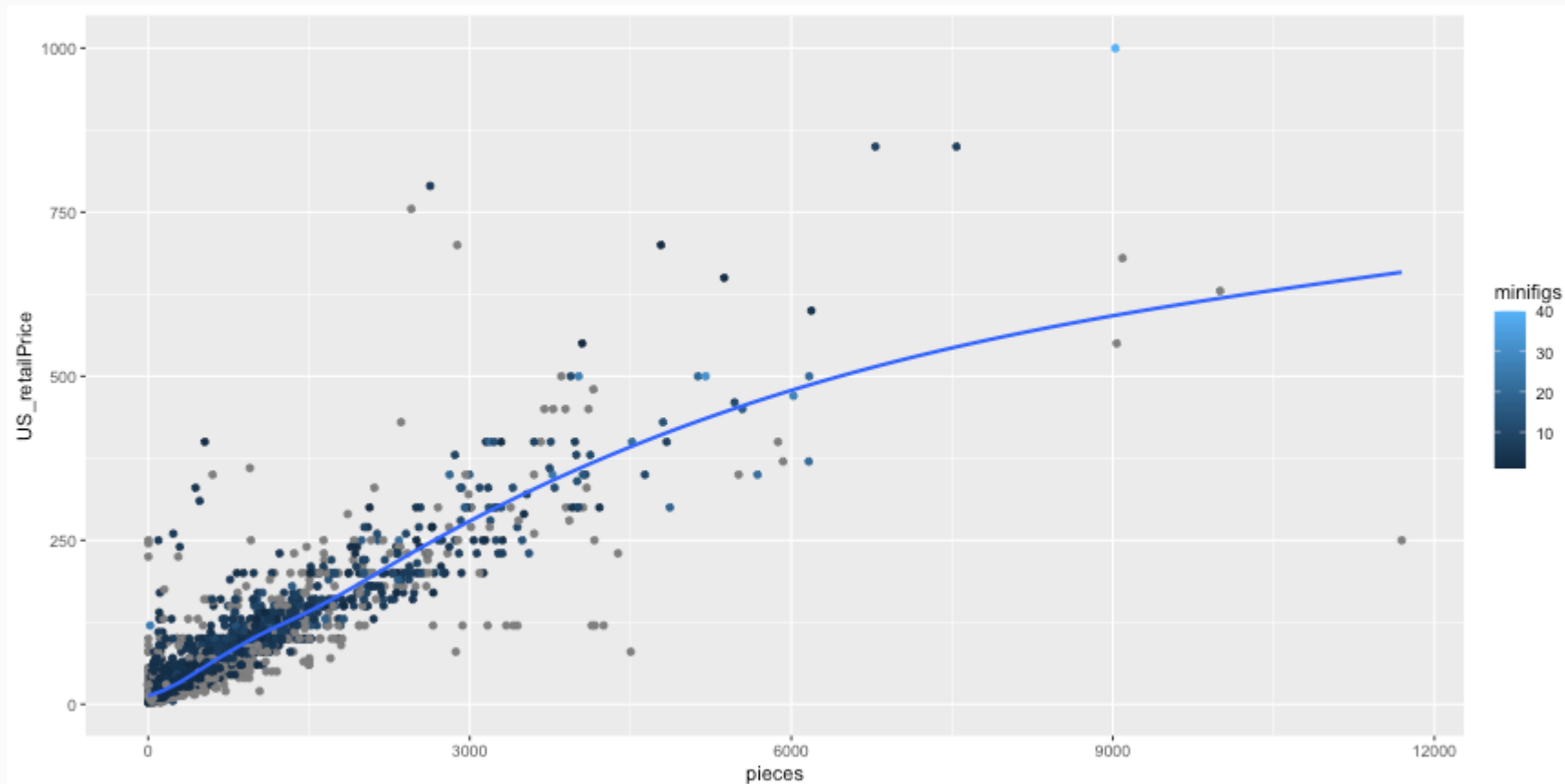
- No outcome variable, just a set of predictors (features) measured on a set of samples.
- Objective is more fuzzy — find groups of samples that behave similarly, find features that behave similarly, find linear combinations of features with the most variation.
- Difficult to know how well you are doing.
- different from supervised learning, but can be useful as a pre-processing step for supervised learning.

Statistical Learning vs. Machine Learning

- Machine learning arose as a subfield of Artificial Intelligence.
- Statistical learning arose as a subfield of Statistics.
- There is much overlap — both fields focus on supervised and unsupervised problems:
- Machine learning has a greater emphasis on large scale applications and prediction accuracy.
- Statistical learning emphasizes models and their interpretability, and precision and uncertainty.
- But the distinction has become more and more blurred, and there is a great deal of “cross-fertilization”.
- Machine learning has the upper hand in Marketing!

Example of a machine learning problem

```
data(legosets, package = 'brickset')
legosets <- legosets |> dplyr::filter(year > 2010)
ggplot(legosets, aes(y = US_retailPrice, x = pieces, color = minifigs)) +
  geom_point() + geom_smooth(se = FALSE)
```



An example from my own work



The Diagnostic Assessment and Achievement of College Skills (DAACS) is a suite of assessments designed to assess college readiness including:

- A Self-Regulated Learning Assessment with approximately 60 items
 - Using exploratory factor analysis, we combine the items to create four subdomains (i.e. motivation, metacognition, strategies, self-efficacy)
- Math and reading assessment (students complete 18 to 24 items and receive a total score ranging from 0 to 1)
- Writing assessment (this is machine scored using large language models and predictive models)
- Feedback views (this is a measure of how many pages students view)

We also have basic demographic data.

Our outcome measure of interest is academic success (e.g. GPA, retention, graduation).

See Bryer, J.M., Akhmedjanova, D., Andrade, H.L., & Lui, A.M. (2021). *The use of predictive modeling for assessing college readiness*. In H. Jiao & R.W. Lissitz (Eds.) *Enhancing Effective Instruction and Learning Using Assessment Data*, Emerald Insight. <https://doi.org/10.1108/978-1-64802-628-7>



Next steps...



As soon as possible:

- [Join the Slack channel](#)

Then:

- Start the *Introduction to the Modeling* lab (due September 6th)



Good luck with the semester!

 jason.bryer@cuny.edu

 cuny-msds.slack.com

 @jbryer

 @jbryer@vis.social

 fall2026.is382.net