DS 3010

Spring 2026

Wenting Xu

IOWA STATE UNIVERSITY

Copyright& Attribution

Some lecture slides and instructional materials in this course are adapted from the following sources:

•*An Introduction to Statistical Learning: With Applications in R (Second Edition)*
Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani
Springer, 2021

•Online course materials developed by Trevor Hastie, Robert Tibshirani, and collaborators.

DS 3010: Introduction

Wenting Xu

Iowa State University

Today's Agenda

- Course logistics & syllabus.
- Course overview.
- Your To-Do list.

Your Instructor

Wenting Xu (Prounced shÜ)

- Joined ISU in 2024
- Assistant teaching professor, Department of Statistics and Department of Computer Science
- Previously taught STAT 2260, STAT 5260, COMS 1130, COMS 1270, COMS 2070, COMS 2270, DS 3010

Teaching Philosophy

- Aim to build a strong foundation in both theory and application **without heavy reliance on math while keeping the learning process engaging and enjoyable.**
- We will focus on the important fundamentals and core ideas that are the building blocks for the majority of statistical learning techniques.

Syllabus Highlights

- **Please read the syllabus.** It is our contract.
- Course material can be found on Canvas. Check Canvas regularly for announcements, assignments, due dates, etc.
- Grading Scheme:
  - Homework (30%).
  - Two midterm exams (each 15%).
  - Final exam (30%).
  - Participation (10%).

Homework/Exams

- Weekly homeworks are generally due on **Mondays at 11:59 PM**. No late homework accepted. Submit via Canvas.
- Code and solutions should be submitted as separate documents.
- No make up midterm exams. If you miss a midterm, your final exam score will count in place of your midterm score.
- Exams are **open note/book and closed internet**. You are not allowed to communicate/discuss with others.

Participation

- Collaboration and discussion are crucial components of the learning process.
- You can earn participation points (up to 10) in this class by contributing to in-class activities. These activities may include:
  - Short conceptual questions
  - Hands-on modeling exercises (with or without code)

The **two lowest scores** will be dropped to allow for flexibility.

Lectures

- Lecture slides and R scripts will be posted **before** class starts.
- You are expected to take your own notes. The posted slides are not sufficient for studying and will require your annotations.
- Please bring your laptops to lecture!

Course Materials

Textbook

- Textbook (*An Introduction to Statistical Learning in* `R`) is optional but recommended. It is freely available here.

Software

- Coding will be done using `R` and `RStudio`.
- If your `R` is rusty, please take a look at the document 'Basics in `R`' posted on Canvas. Work through some of the simple examples to get a feel for the basics.
- Problems troubleshooting/debugging your code? Come to student hours or post in Piazza.

Communication Guidelines

- **Piazza** will be used as the class forum.
- Questions related to homework/exam material should be posted on Piazza or asked during my student hours (see Syllabus for link and hours).
- Emails should be reserved for questions related to grades or administrative issues.

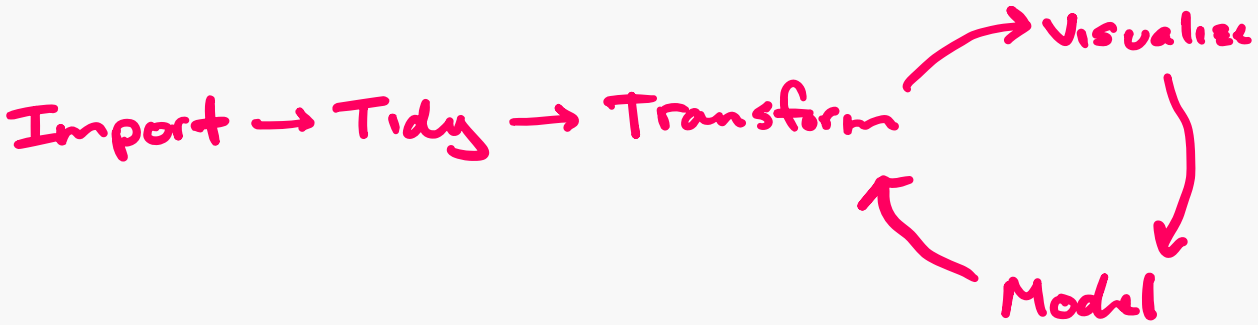
Applied Data Modeling and Predictive Analysis

This is an introductory course in **statistical learning concepts** and applications intended for students with some R experience and some introductory statistics background.

- Theoretical component: conceptual understanding of methods and statistical learning theory.
- Applied component: selecting appropriate methodology, implementation, problem solving, data analysis.
- This course leans heavily on the applied component, although you cannot full appreciate the methodology without some theory.

Where does this course fit in your current progression?

Typical (over-simplified) depiction of workflow:



DS 201/202: acquired raw data, pre-processed that data, and then used exploratory data analysis tools to gain insight from the data. Repeat this as needed.

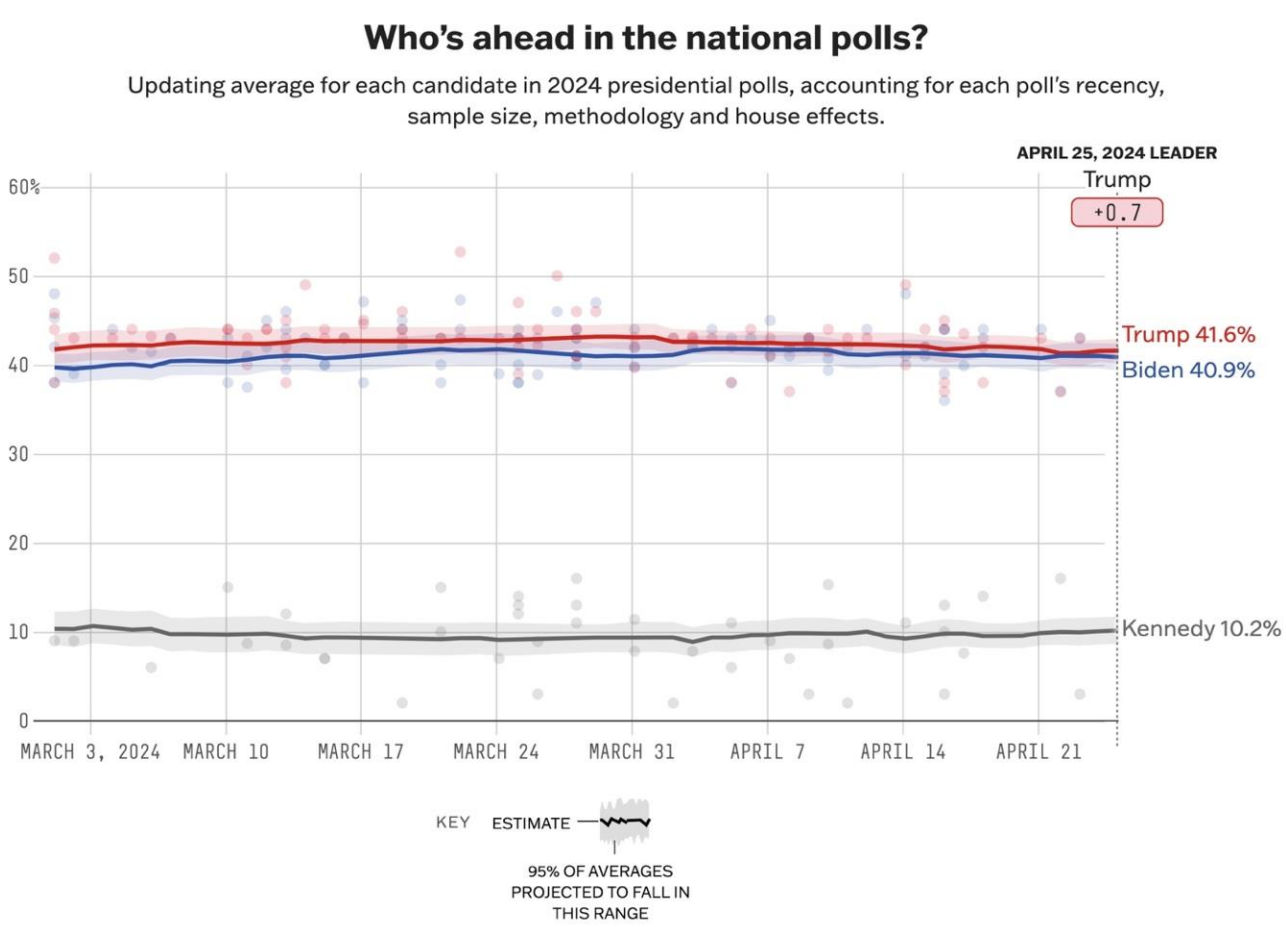
DS 301: We assume that

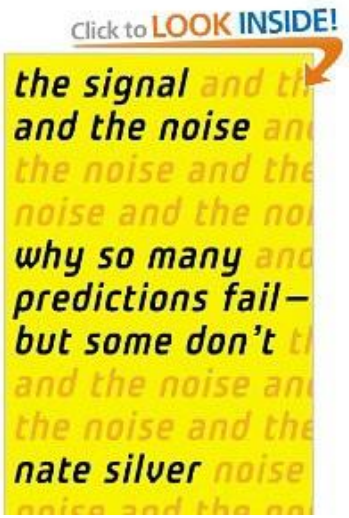
- data is more or less 'clean',
- preliminary exploratory data analysis has been done.

In reality: As a data scientist, 80% of your time will be focused on the techniques learned in DS 202, roughly 10% on the techniques from DS 301, and the remaining 10% of your time will be effectively communicating your results to others.

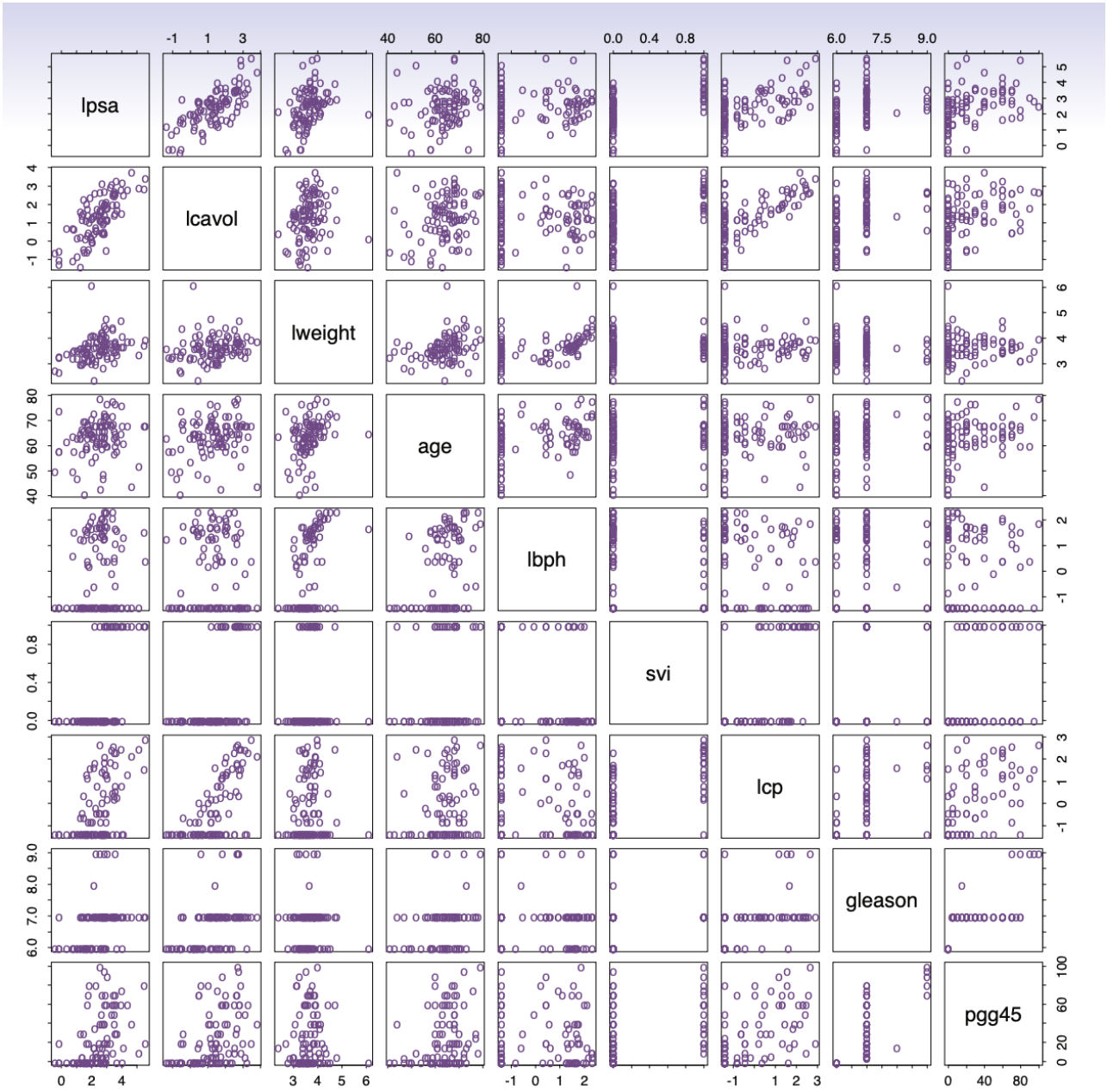
[1]Image courtesy of *R for Data Science*

FiveThirtyEight





Identify the risk factors for prostate cancer.



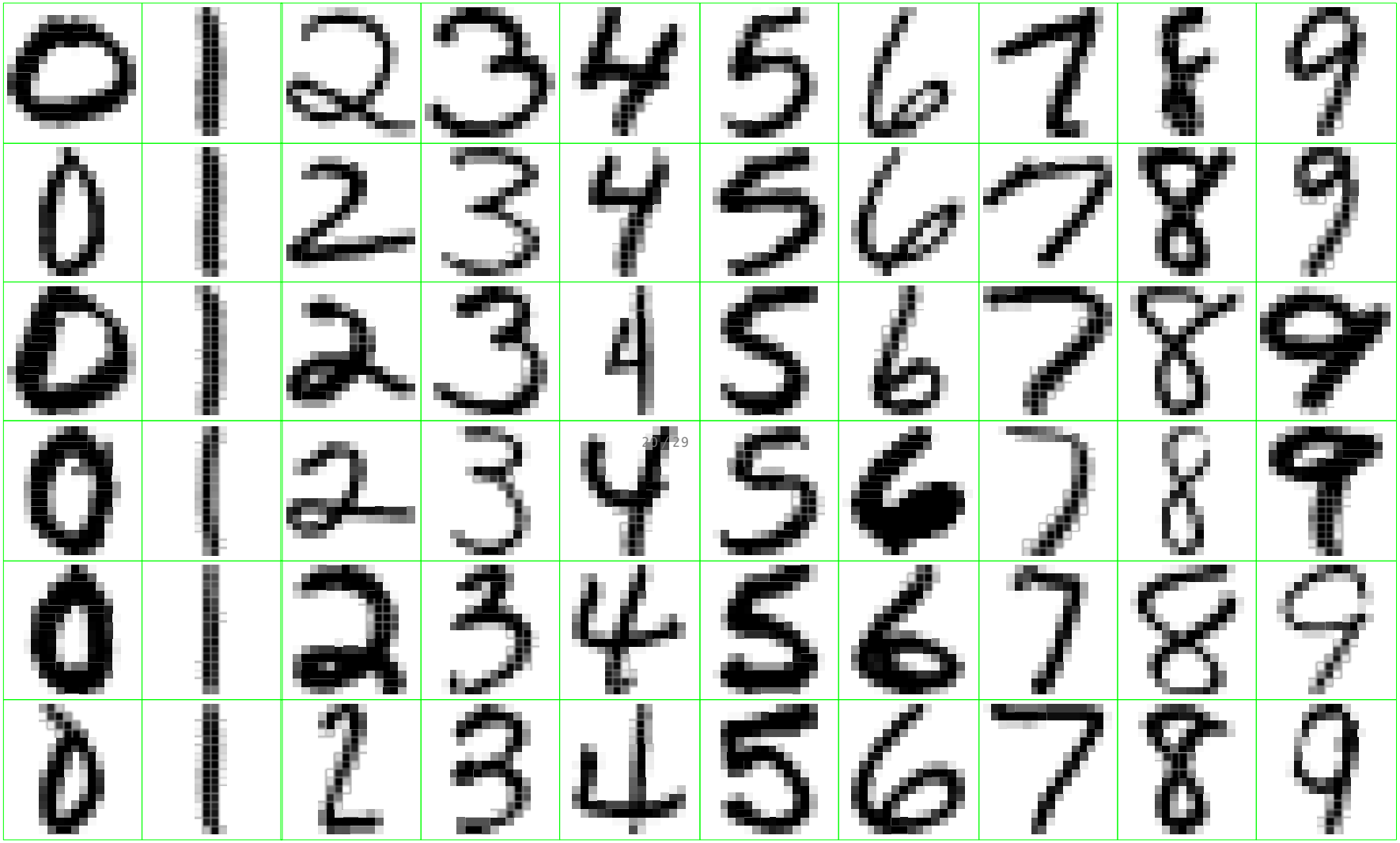
Spam Detection

- data from 4601 emails sent to an individual (named George, at HP labs, before 2000). Each is labeled as spam or email.
- goal: build a customized spam filter.
- input features: relative frequencies of 57 of the most commonly occurring words and punctuation marks in these email messages.

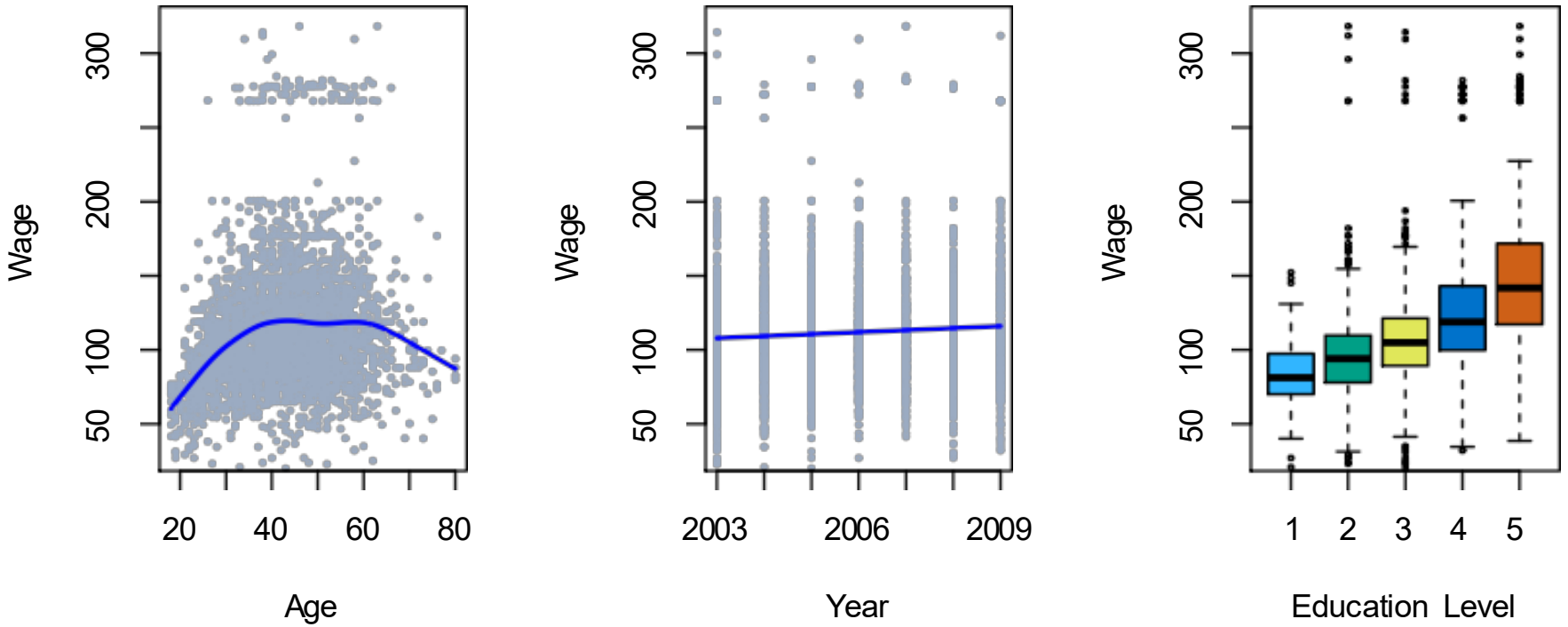
	george	you	hp	free	!	edu	remove
spam	0.00	2.26	0.02	0.52	0.51	0.01	0.28
email	1.27	1.27	0.90	0.07	0.11	0.29	0.01

Average percentage of words or characters in an email message equal to the indicated word or character. We have chosen the words and characters showing the largest difference between **spam** *and* **email**.

Identify the numbers in a handwritten zip code



Relationship between salary and demographic variables



Income survey data for males from the central Atlantic region of the USA in 2009.

Statistical Learning versus 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!*

Your To-Do List

- Fill out **start of semester survey** on Canvas.
- Review 'Basics in R' if you need to.
- HW 1 is due next Wednesday (Jan. 28).

Questions?