COURSE INTRODUCTION

Chapter 1

An overview

- What is Statistical Learning (SL)?
  SL refers to a set of tools to understand data.

- What is Data Mining (DM)?
  DM is the analysis step of "knowledge discovery in databases (KDD)" process.

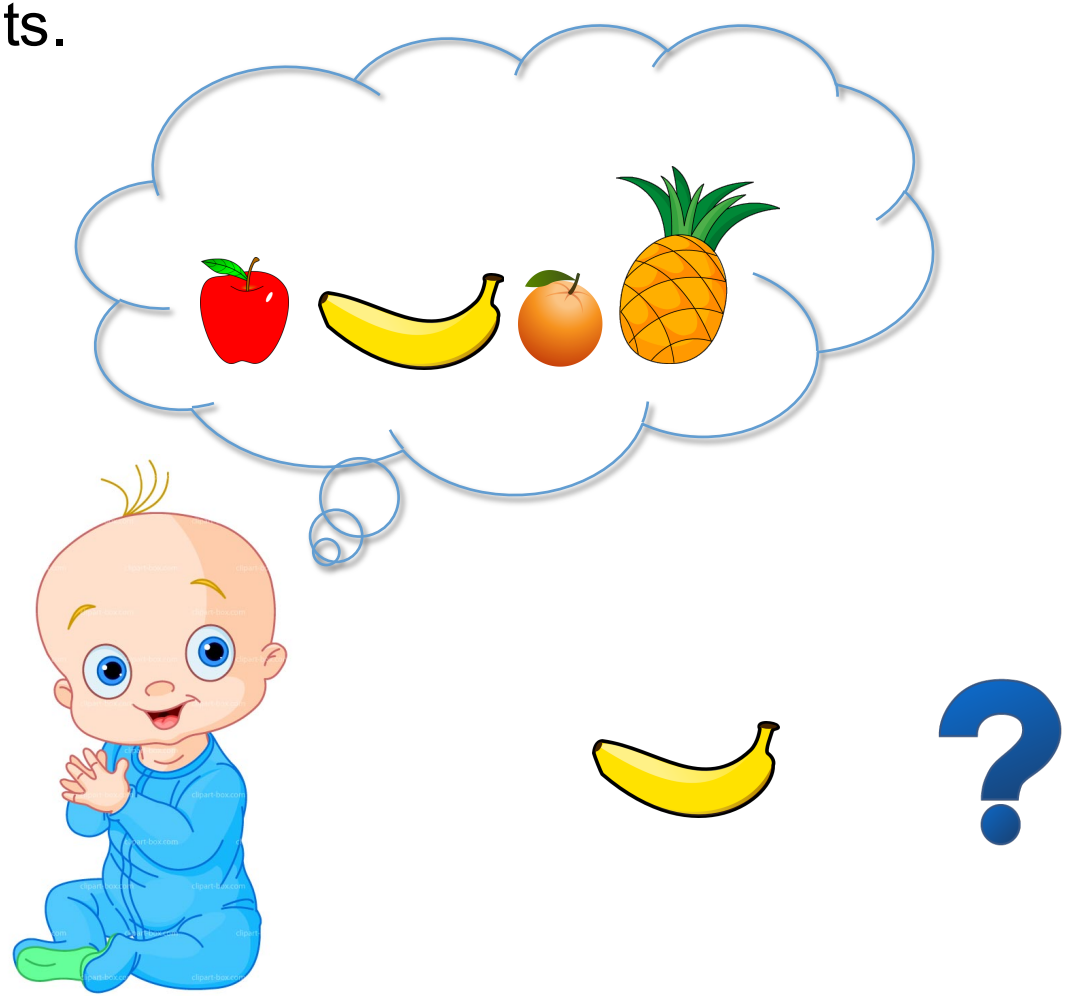
Unified objective: use data to extract useful information.

Why this course is launched?

- High demand of statistical learning skills in industry.
- Answer scientific questions across a number of fields.

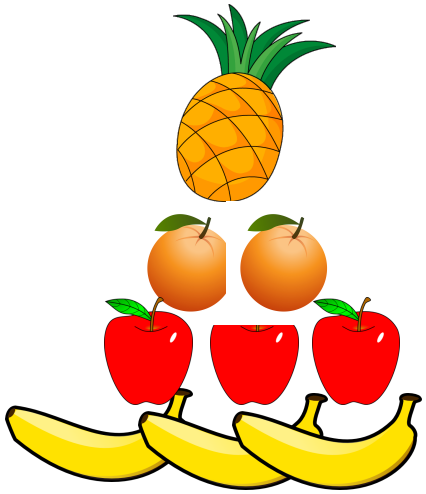
Two classes of learning problems

- Supervised statistical learning involves building a statistical model for predicting, or estimating output based on one or more inputs.



Two classes of learning problems

- Unsupervised statistical learning involves only inputs but no output, and we want to learn the structure and patterns from data.



Supervised vs. unsupervised

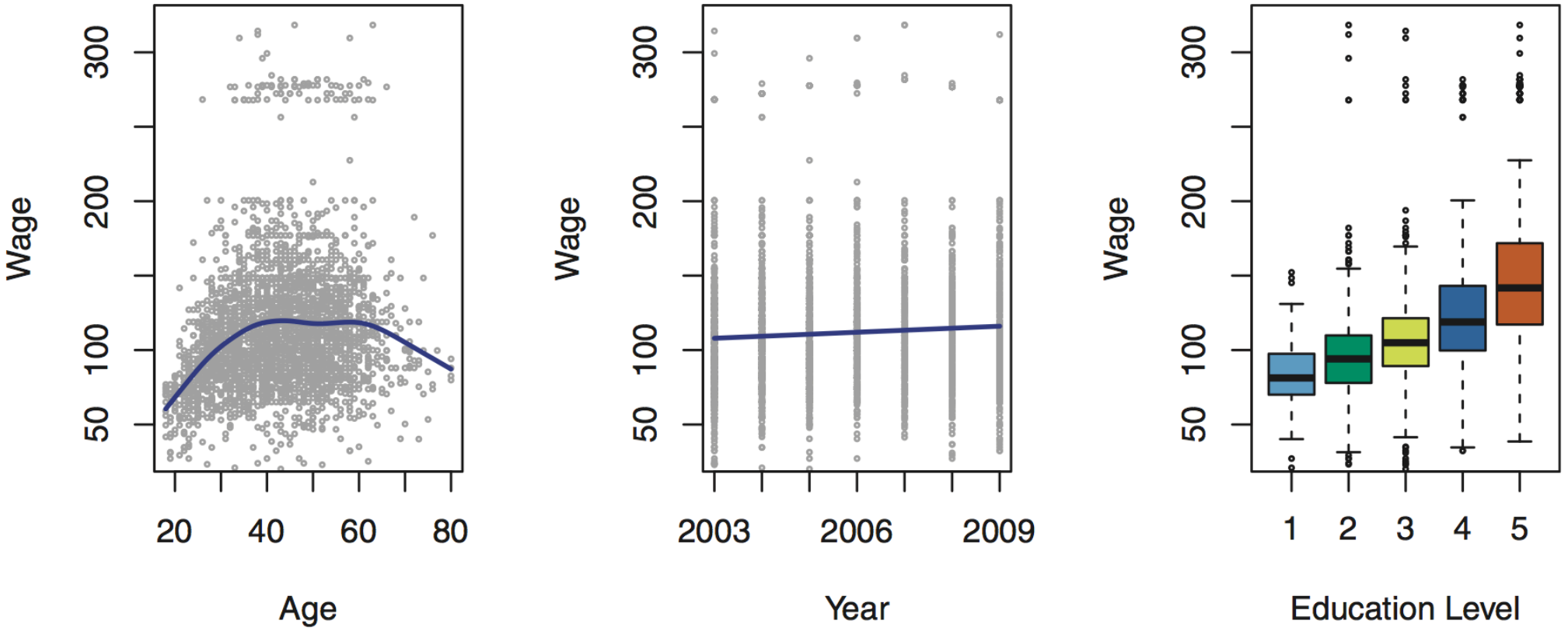
- Supervised learning has a training set of data, in which we observe the output and inputs. Unsupervised learning doesn't have training set.

- Supervised learning has output but unsupervised learning doesn't.

- They have different tasks. Our task for supervised learning is to predict or estimate output. The task for unsupervised learning is to describe how the data are organized or clustered.

Statistical Learning Problems

- *Establish the relationship between wage and demographic variables in population survey data.*



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

Statistical Learning Problems

- *Customized email spam detection system*

➢Data: 4601 emails sent to an individual (named George). Each is labeled as spam or email.

➢Goal: build a customized spam filter.

➢Features: relative frequencies of 57 of the most commonly occurring words and punctuation marks in these email messages.

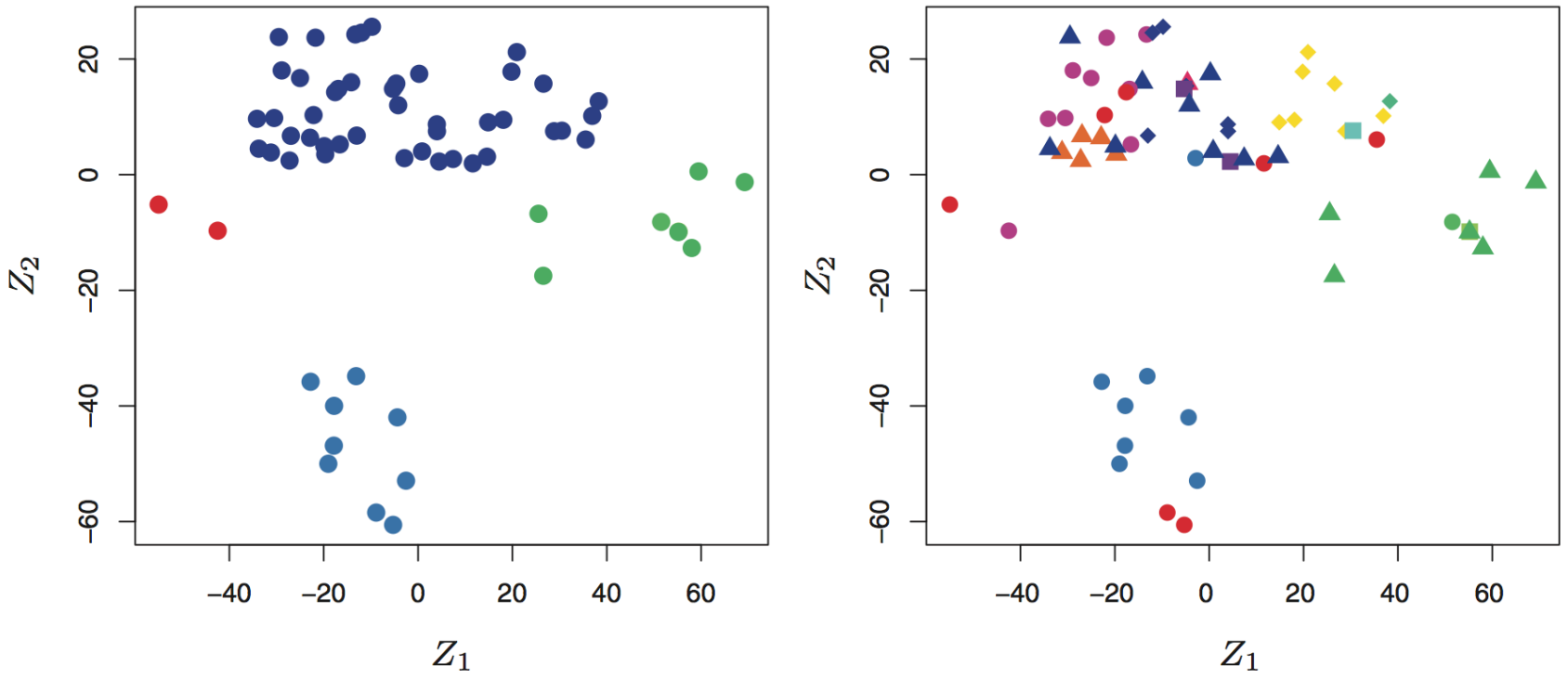
	george	you	hp	free	!	edu	remove
spam	0.00	2.26	0.02	0.52	0.51	0.01	0.28
email	1.27	1.27	0.90	0.07	0.11	0.29	0.01

Statistical Learning Problems

- *Cancer tissue gene expression clustering*

➢Data: 6,830 gene expression measurements for each of 64 cancer cell lines.

➢Goal: determine whether there are groups or clusters among the cell lines.



Topics we will cover in this course

- Basic concepts behind statistical learning. (Ch2)
- Linear regression. (Ch3)
- Logistic regression and linear discriminant analysis. (Ch4)
- Model evaluation methods. (Ch5)
- Linear model selection and regularization. (Ch6)
- Moving beyond linear methods. (Ch7)
- Tree-based methods. (Ch8)
- Support vector machine. (Ch9)
- Clustering methods. (Ch10)

Syllabus overview

Instructor: **Dr. Tao He**
Office: **Thornton Hall 945**
Phone: **(415)338-1368**
Email: hetao@sfsu.edu
Office Hours: **Tu Th 2:00pm-3:00pm, or by appointment**
Lecture: **Tu Th 11:00pm-12:15pm @ Thornton Hall 211**
Course web page: ilearn.sfsu.edu

Prerequisite: MATH 340 with grade of C or better or consent of instructor.

Syllabus overview (cont'd.)

Statistical Software: We are going to use R project, see http://www.R-project.org. Rstudio is a recommended interface for the R software. It is free and it runs on Windows, Mac and Linux operating systems. https://www.rstudio.com/.

Textbook: G. James, D. Witten, T. Hastie, T. Tibshirani.(2013). *An Introduction to Statistical Learning, with Applications in R*. New York: Springer. **The textbook is required**. The book's website is http://www-bcf.usc.edu/∼gareth/ISL/index.html. You can download the book and related materials for free from the book's website.

Course Objectives: At the completion of this course, students will be able to:

- Obtain a strong conceptual understanding of statistical learning.
- Learn the statistical principles behind many of the approaches to supervised & unsupervised learning.
- Understand how to perform model selection & evaluation and to effectively communicate the results.
- Learn how to rigorously analyze data using modern statistical methods and computer software.
- Obtain hands-on experience by analyzing real data sets with the skills learned throughout the course.

Syllabus overview (cont'd.)

Term Project: Each student is required to find a topic for the term project and apply what you learned in this course to analyze data sets drawn from different application domains. You can either work on the project individually or work with **one** teammate. The complete submission includes:

- A one-page project proposal which is due by Feb 23rd.
- A 10 minutes presentation of your term project in one of the last three classes.
- A final project report which is due by the May 23rd.

Term project counts a total of 180 points. **No extension will be given**.

Homework Assignments: Each assignment will contain both conceptual and applied problems, where the former ones examine your understanding of basic concepts and latter ones require you to implement methods in real data sets. Homework must be submitted in class on the due date. There will be eight assignments, each of them counts 30 points. The lowest one will be dropped. **No late homework will be accepted**.

Syllabus overview (cont'd.)

Homework Assignments: Each assignment will contain both conceptual and applied problems, where the former ones examine your understanding of basic concepts and latter ones require you to implement methods in real data sets. Homework must be submitted in class on the due date. There will be eight assignments, each of them counts 30 points. The lowest one will be dropped. **No late homework will be accepted**.

Quick Quizzes: Quick Quizzes are designed to check your understanding of the fundamental concept. Six 5-minute Quizzes will be given, each of them counts 18 points. The lowest one will be dropped.

Midterm Exam: There will be one take-home midterm exam. It counts 90 points.

Attendance: Attendance is expected and will be taken each class. Students are responsible for all missed work, regardless of the reason for absence. It is also the absentee's responsibility to get all missing notes or materials.

Syllabus overview (cont'd.)

Letter Grade Distribution: Points are earned through homework assignments, quizzes, midterm exam, term project and attendance. The total number of points is 600. Your grade will be converted into a percentage and be determined by the following grading scale:

>= 93.00	A	73.00 - 76.99	C
90.00 - 92.99	A-	70.00 - 72.99	C-
87.00 - 89.99	B+	67.00 - 69.99	D+
83.00 - 86.99	B	63.00 - 66.99	D
80.00 - 82.99	B-	60.00 - 62.99	D-
77.00 - 79.99	C+	<= 59.99	F

Assignments	35%	$30 \times 7 = 210$ points
Quick Quizzes	15%	$18 \times 5 = 90$ points
Midterm Exam	15%	90 points
Term Project	30%	180 points
Attendance	5%	30 points

Syllabus overview (cont'd.)

Course Policies:

- Please staple your HW in order before submission.
- Quizzes are closed book, closed notes.
- **No make-up quizzes will be given under all the circumstances.** In case of illness you can have make-up midterm exam. But please notify me via email before the exam. If you inform me after the midterm exam, no credit will be given.
- Students are responsible for CR/NC, withdrawal, etc.
- Students are expected to work independently. **Offering** and **accepting** solutions from others is an act of **plagiarism**, which is a serious offense and **all involved parties will be penalized according to the Academic Honesty Policy**.
- Please use the solution manual you found wisely. Copying directly will receive 0.

Syllabus overview (cont'd.)

Students with Disabilities: Students with disabilities who need reasonable accommodations are encouraged to contact the instructor early in the semester. The Disability Program and Resource Center is available to facilitate the reasonable accommodations process. The DPRC, located in the SSB 110, can be reached by telephone at 338-2472 (voice/TTY) or by email at dprc@sfsu.edu.

Religious Holidays: Reasonable accommodations will be made for you to observe religious holidays when such observances require you to be absent from class activities. It is your responsibility to inform the instructor during the first two weeks of class, in writing, about such holidays.

Student disclosures of sexual violence: SF State fosters a campus free of sexual violence including sexual harassment, domestic violence, dating violence, stalking, and/or any form of sex or gender discrimination. If you disclose a personal experience as an SF State student, the course instructor is required to notify the [Dean of Students]. To disclose any such violence confidentially, contact:

- The SAFE Place - (415) 338-2208; http://www.sfsu.edu/~safe_plc/
- Counseling and Psychological Services Center - (415) 338-2208; http://psyservs.sfsu.edu/

For more information on your rights and available resources: titleix.sfsu.edu

The term project

Pick a real data set for which you think there are interesting question(s) to answer. It could be from any area. You need to try all the applicable statistical learning methods that are covered in this course and find the best approach to answer your question(s).

More about the project

Three deliverables:

1. A one-page *proposal* that includes the following:
   - ✓ members' names
   - ✓ description of the problem
   - ✓ description of the data (dimensions, names of the variables, sample size, sources, etc.)
   - ✓ supervised or unsupervised
   - ✓ comments or concerns (optional)

More about the project

Three deliverables:

2. A *presentation* that includes:
   - ✓ Description of the data and the question(s) that you are interested in answering.
   - ✓ Review of some of the methods that you tried.
   - ✓ Summary of the final approach you used and the reasons of your choice.
   - ✓ Summary of the results.
   - ✓ Conclusions.

Note: if there are two students in one team, then both of you need to present and answer questions from audience.

More about the project

Three deliverables:

3. A *project report* that includes:
   - ✓ A detailed version of the materials covered in the presentation (maximum 10 pages). The first page should be an extract (executive summary).
   - ✓ All R scripts used in project. (appendix)
   - ✓ The presentation slides. (4-in-1 page)

Grading of the project

- Proposal counts 20% of the project score.
- Presentation counts 40% of the project score.
- Final report counts 40% of the project score.

Suggestions:

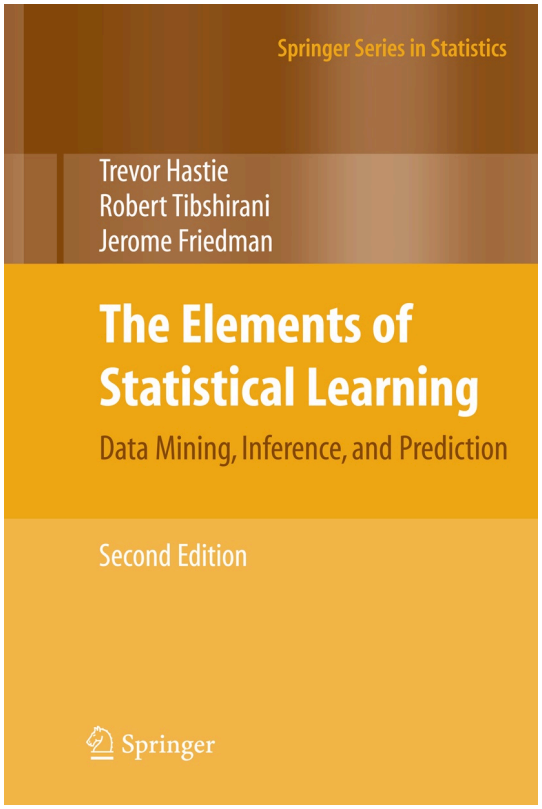
1. Start to think about the project now.
2. You can send your proposal to me before the deadline.
3. Don't wait until the last week. Start to work on the project once get my approval on your proposal.
4. The R code in the textbook and your HWs are helpful!

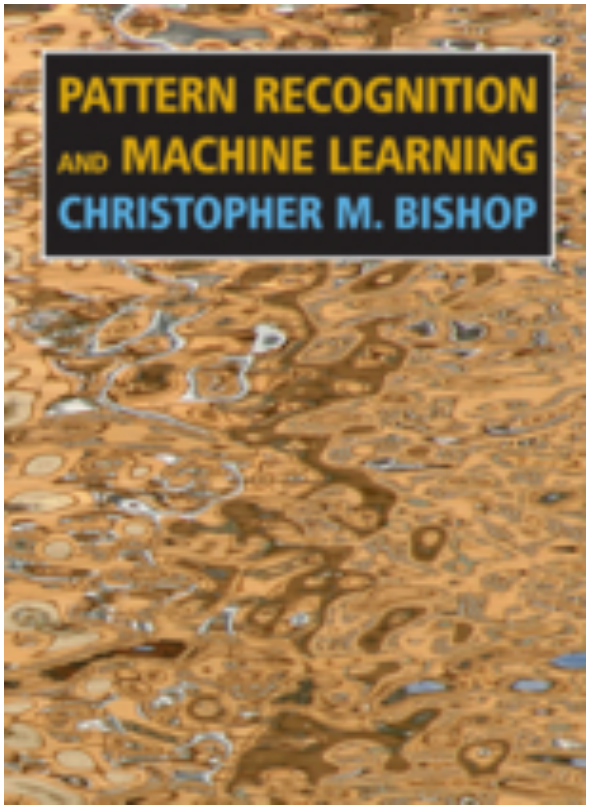
Data Repositories

1. Open Gov. Data:
   www.data.gov, www.data.gov.uk, www.data.gov.fr,
   http://opengovernmentdata.org/data/catalogues/
2. Kaggle: www.kaggle.com
3. KDD Nugets: http://www.kdnuggets.com/datasets/
4. UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/
5. StatLib: http://lib.stat.cmu.edu
6. TwitteR: http://cran.r-project.org/web/packages/twitteR/index.html
7. rfigshare: http://figshare.com,
   http://cran.r-project.org/web/packages/rfigshare/index.html

Adapted from Abbass Al Sharif, USC

Recommended reference books





Friedman, J., Hastie, T., & Tibshirani, R. (2009). *The elements of statistical learning* (2nd edition). Springer, Berlin: Springer series in statistics.
http://statweb.stanford.edu/~tibs/ElemStatLearn/
Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.