STATISTICAL LEARNING - INTRODUCTION

统计学习

课程介绍

王涛 neowangtao@sjtu.edu.cn

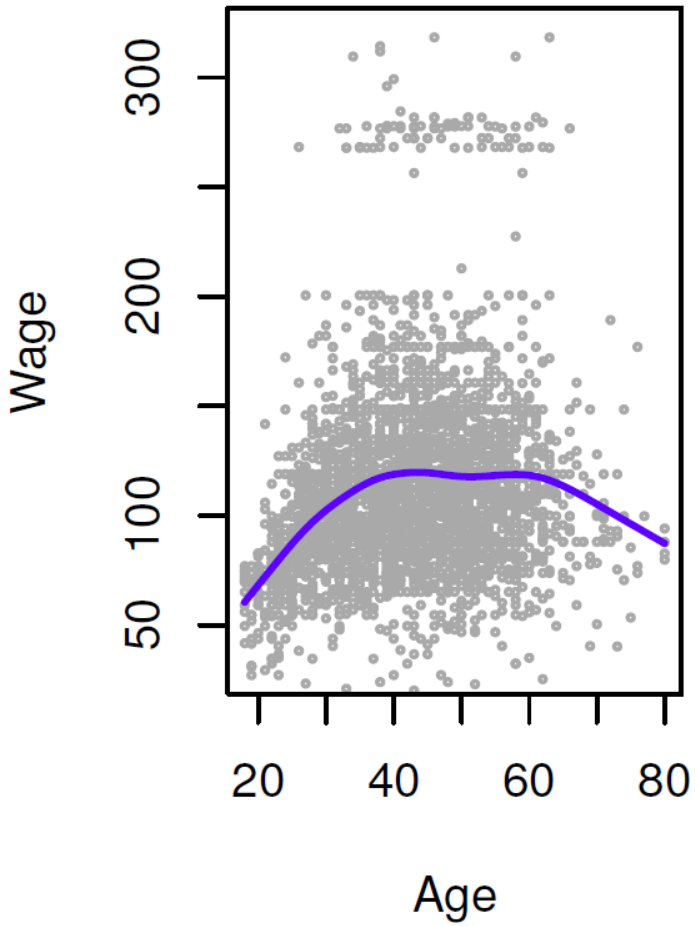
Outline

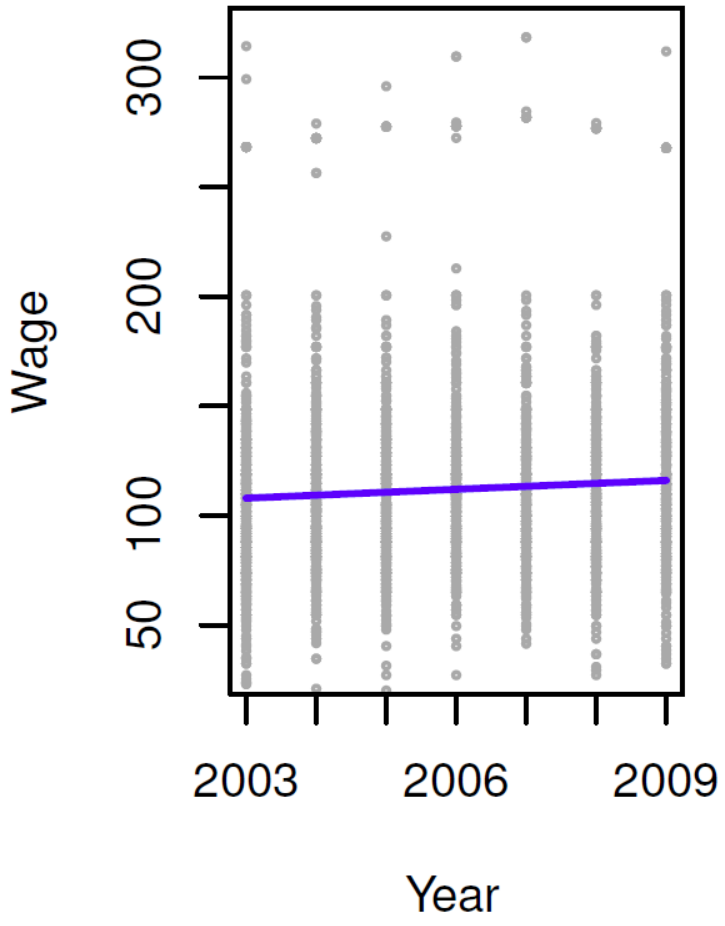
- An overview of statistical learning
- Schedules (教学安排)
- Grading policy (考核方法)
- Textbooks & references
- A brief history of statistical learning

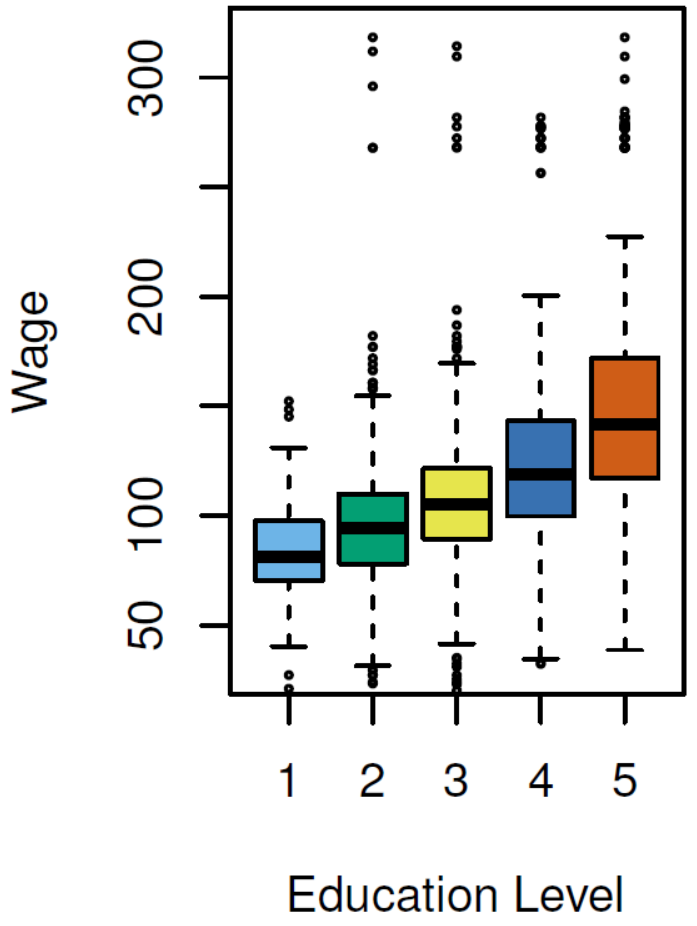
An overview of statistical learning

Wage data

- Income survey data for males in central Atlantic region of USA (Wage)
- Understand the association between an employee's wage and a number of factors, such as age, education, and the calendar year



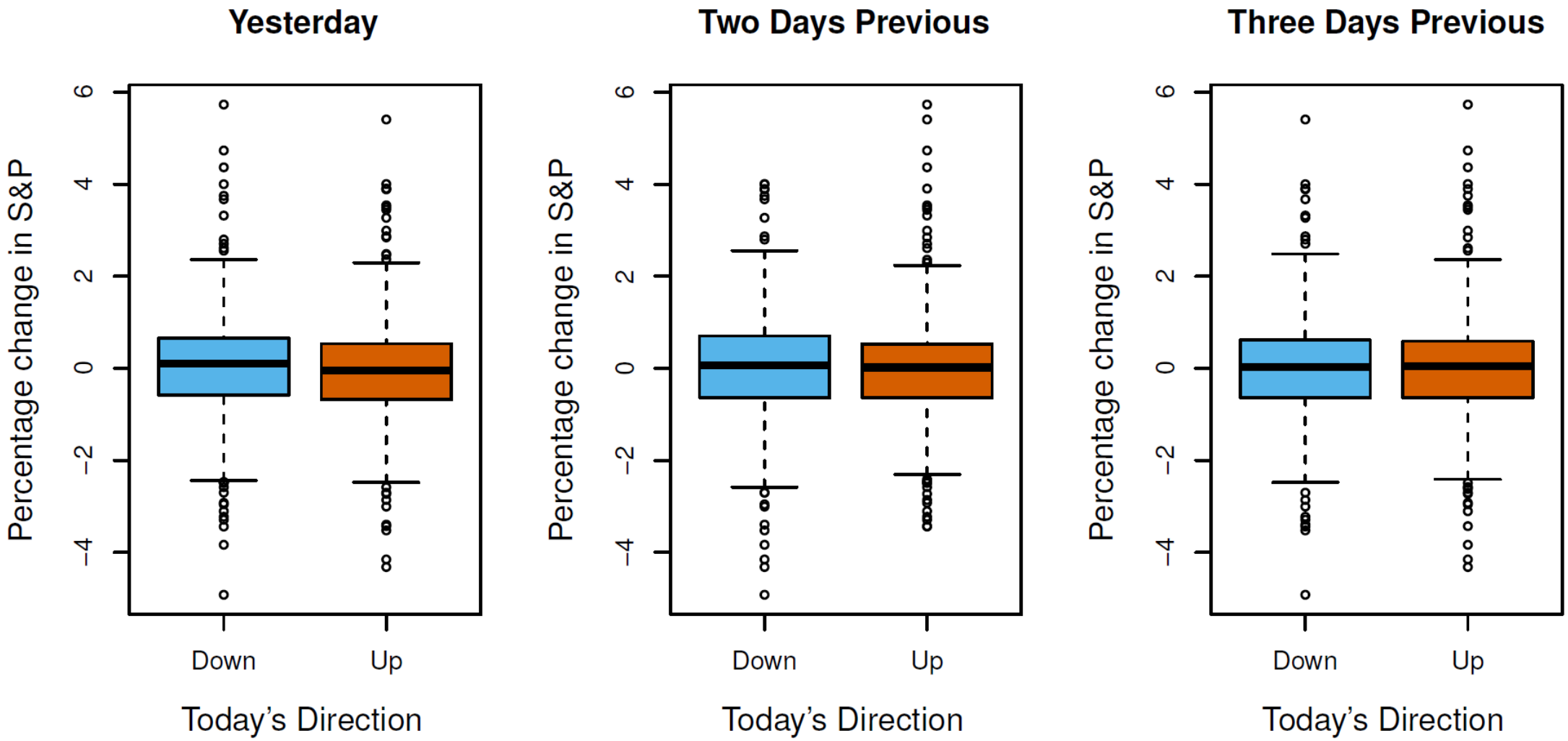




Some of the figures and tables in this presentation are taken from "*An Introduction to Statistical Learning, with Applications in R*" (Springer) with permission from the authors: G. James, D. Witten, T. Hastie, and R. Tibshirani

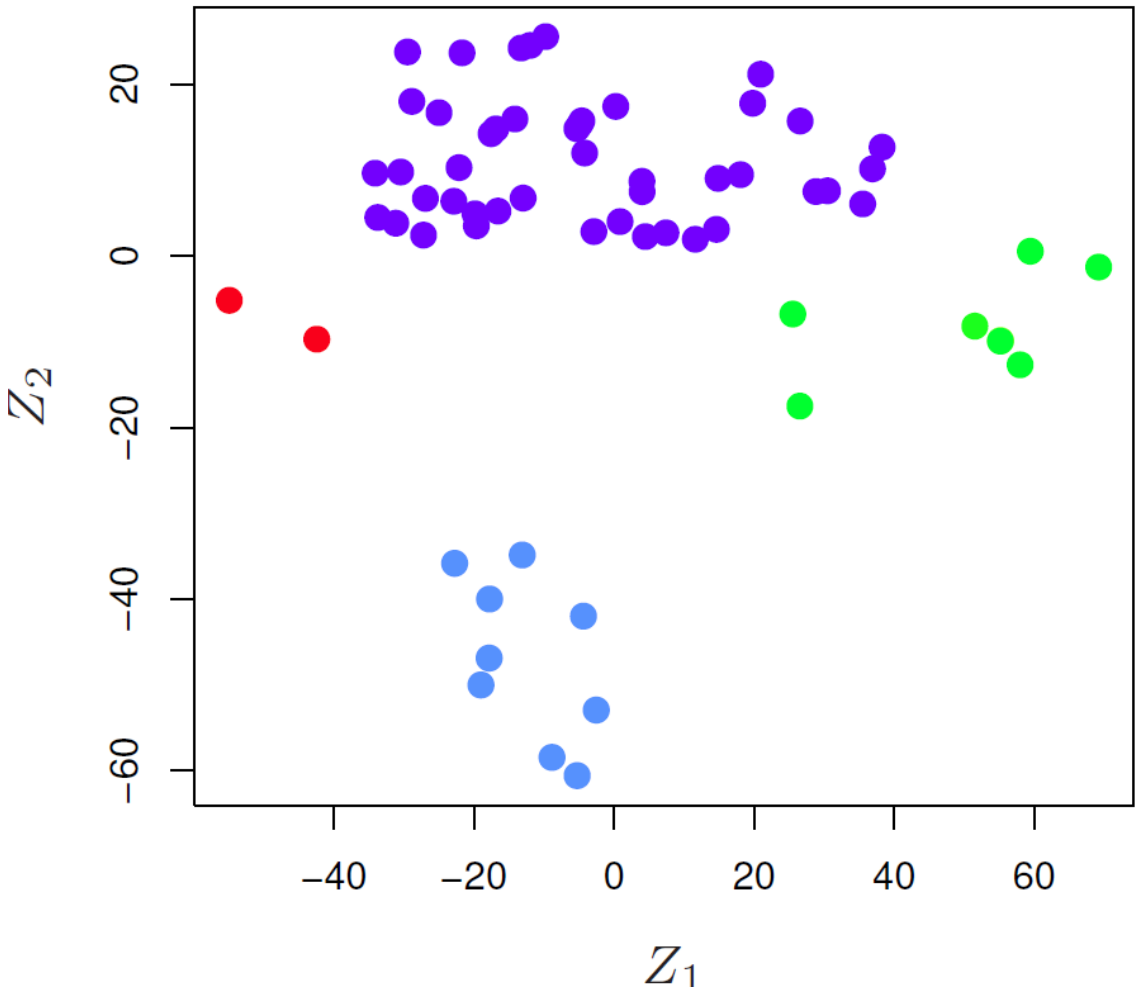
Stock market data

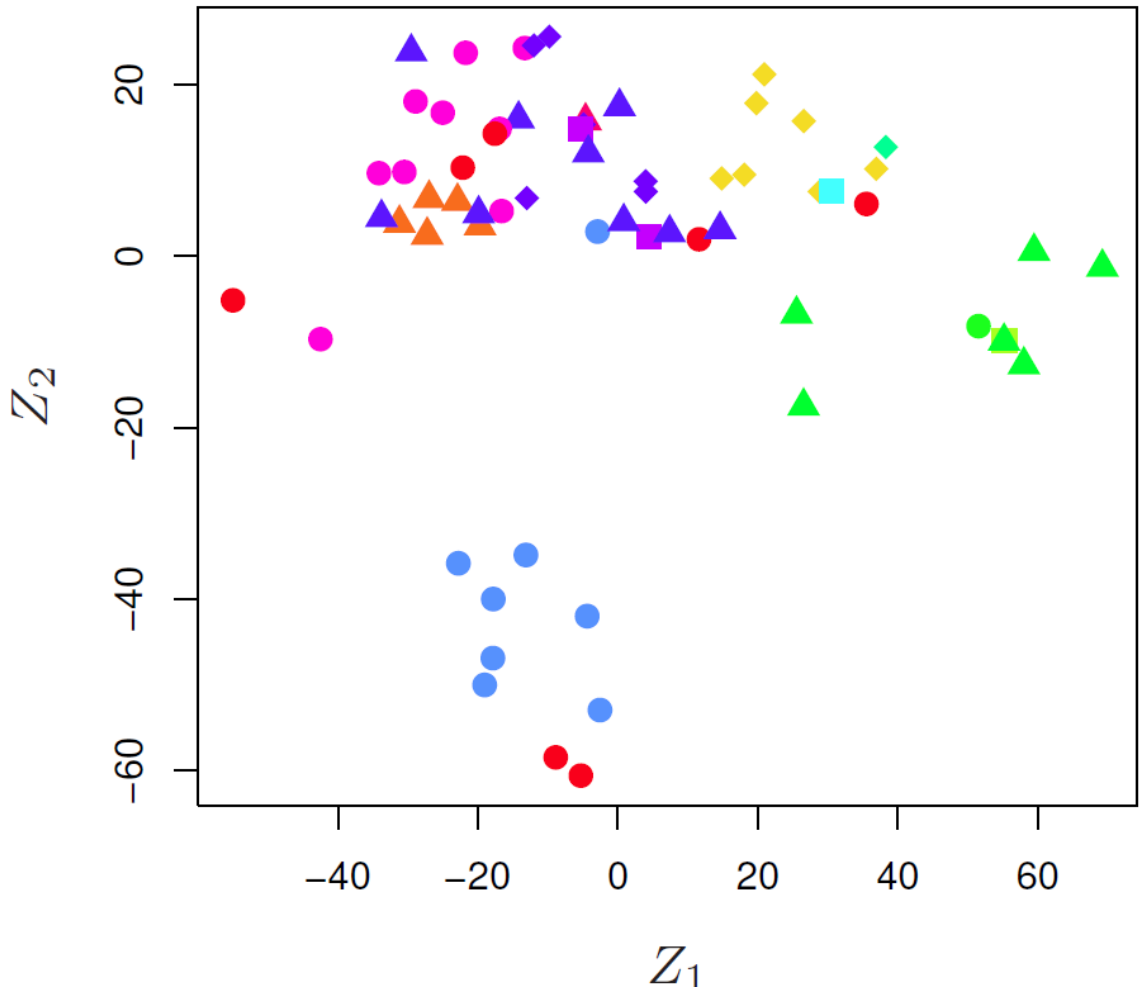
- Daily percentage returns for S&P 500 stock index over a 5-year period (Smarket)
- Predict whether the index will increase or decrease on a given day using the past 5 day's percentage changes in the index


Yesterday
Two Days Previous
Three Days Previous
Percentage change in S&P
6
4
2
0
−2
−4
Down
Up
Today's Direction

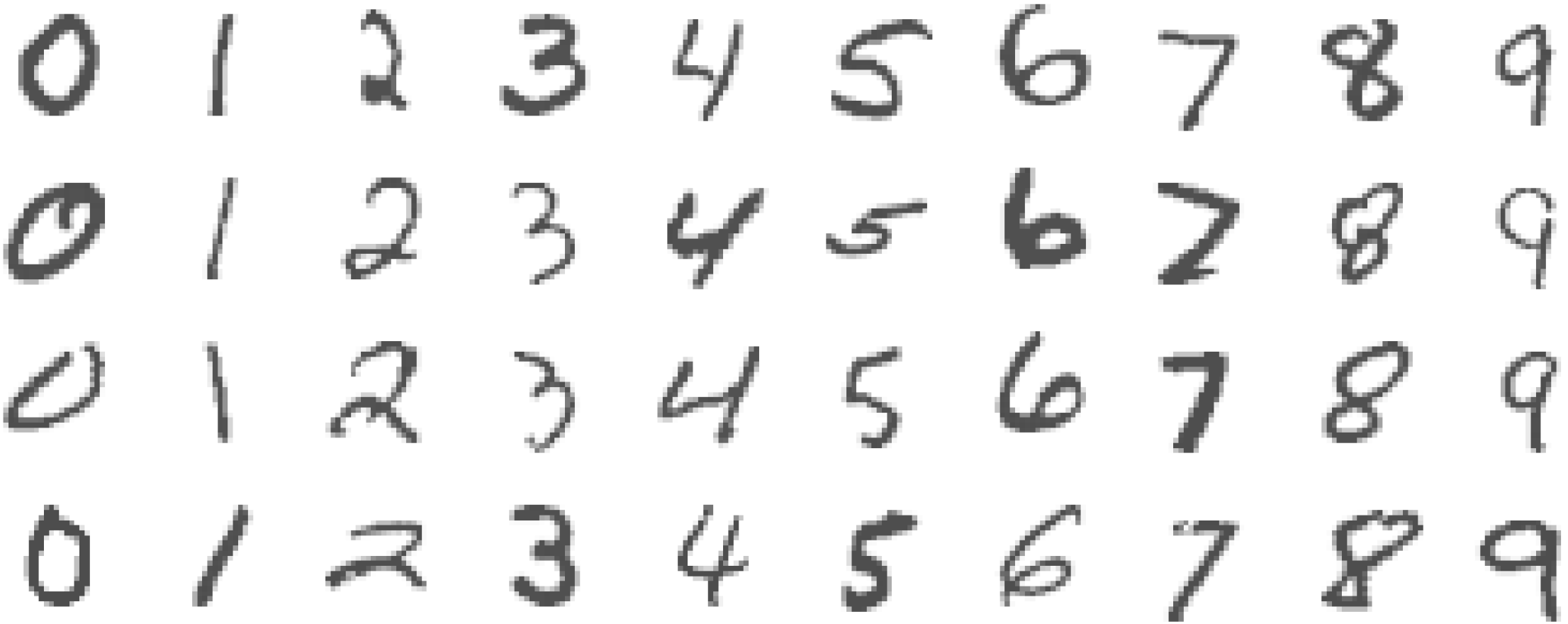
Gene expression data

- Gene expression measurements for 64 cancer cell lines (NCI60)
- Determine whether there are subgroups among 64 cancer cell lines based on 6830 gene expression measurements

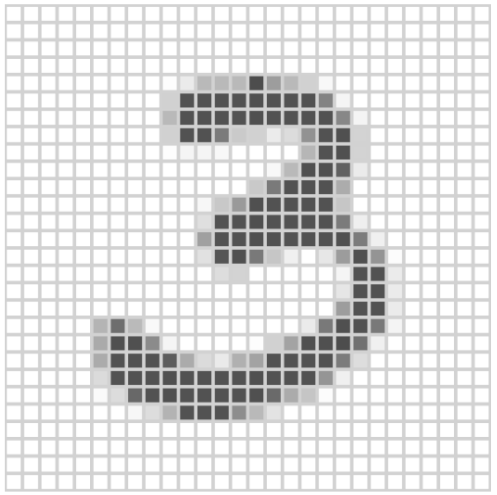

Z_2
20
0
−20
−40
−60
−40
−20
0
20
40
60
Z_1

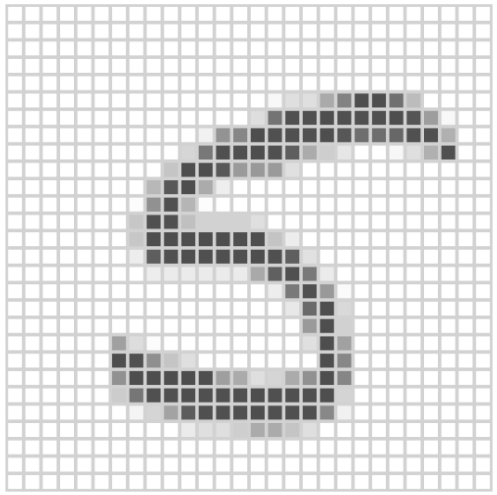

Z_2
20
0
−20
−40
−60
−40
−20
0
20
40
60
Z_1

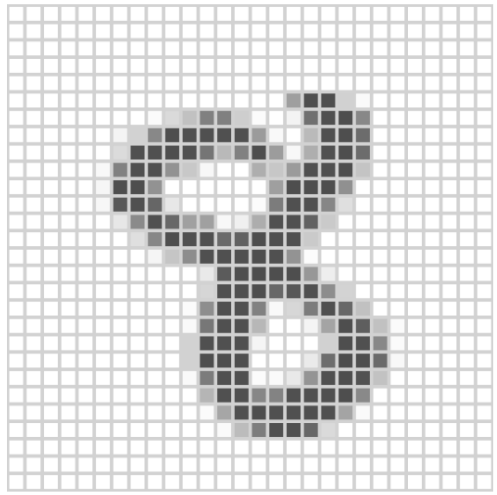
Handwritten digits from the MNIST corpus



- Modified National Institute of Standards and Technology
- 28×28 grayscale images
- Features: the 784 pixel grayscale values (0-255)
- Labels: the digit classes (0-9)
- 60K training and 10K test images

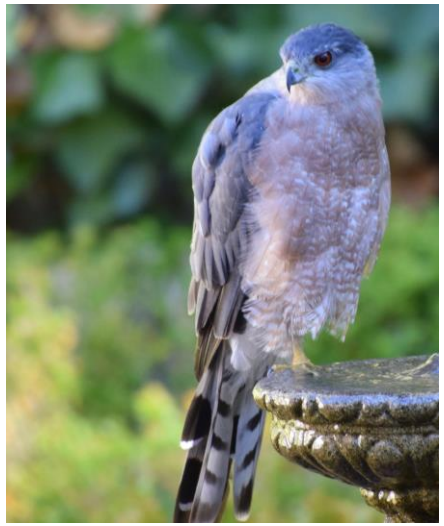


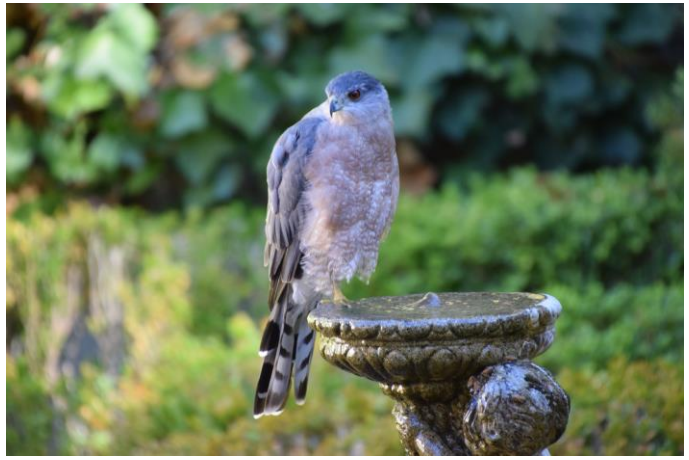




Images from the CIFAR100 database

- Canadian Institute for Advanced Research
- 32×32 color natural images, with 100 classes
- 50K training and 10K test images
- Each image is a $32 \times 32 \times 3$ array of 8-bit numbers (0-255)
- The last dimension represents the color channel (R, G, B)

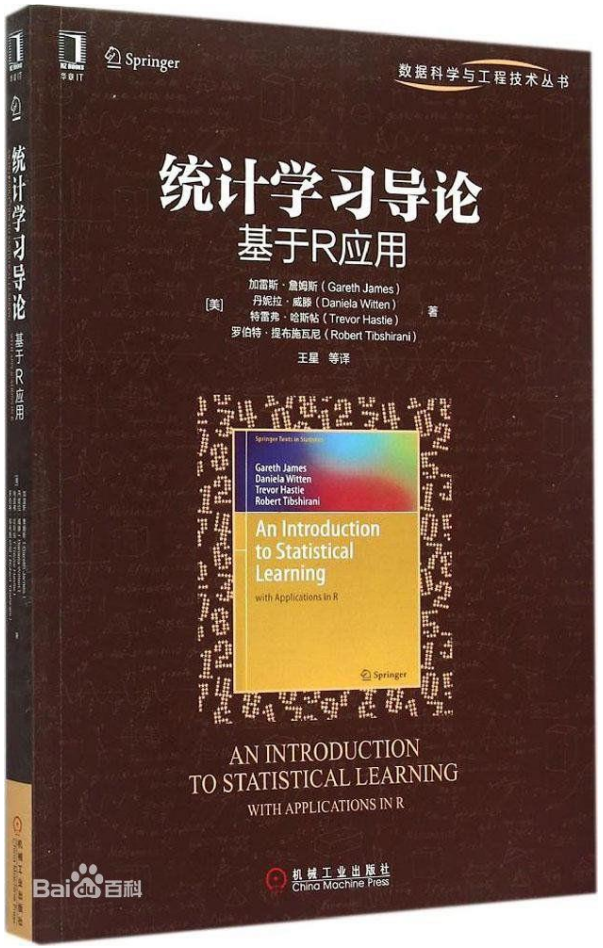



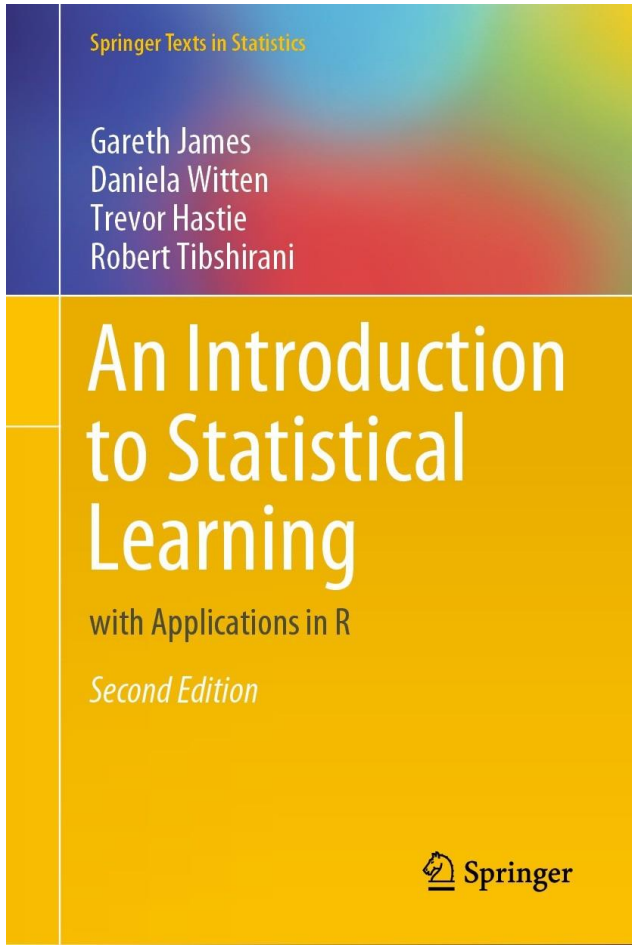


Schedules (教学安排)

- Basic terminology and concepts
- Linear regression and classification
- Resampling methods
- Model selection and regularization
- Nonlinear methods
- Tree-based methods
- Support vector machines
- Neural networks and deep learning
- Unsupervised learning
- Oral presentation and quizzes

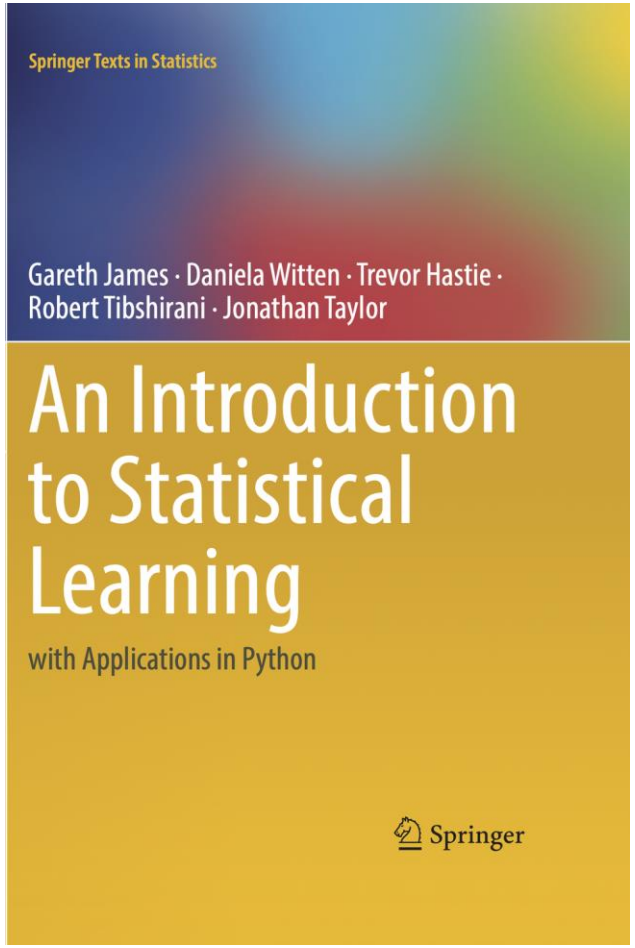
Textbooks & references

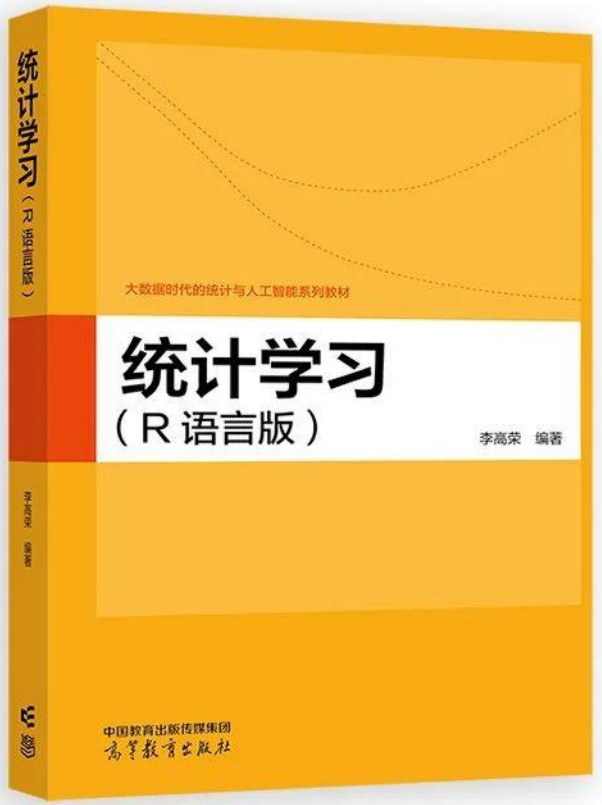

Springer
数据科学与工程技术丛书
统计学习导论
基于R应用
加雷斯·詹姆斯（Gareth James）
丹妮拉·威滕（Daniela Witten）
特雷弗·哈斯帖（Trevor Hastie）
罗伯特·提布施瓦尼（Robert Tibshirani）
[美] 著
王星 等译
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R
AN INTRODUCTION
TO STATISTICAL LEARNING
WITH APPLICATIONS IN R
机械工业出版社
China Machine Press
Baidu百科

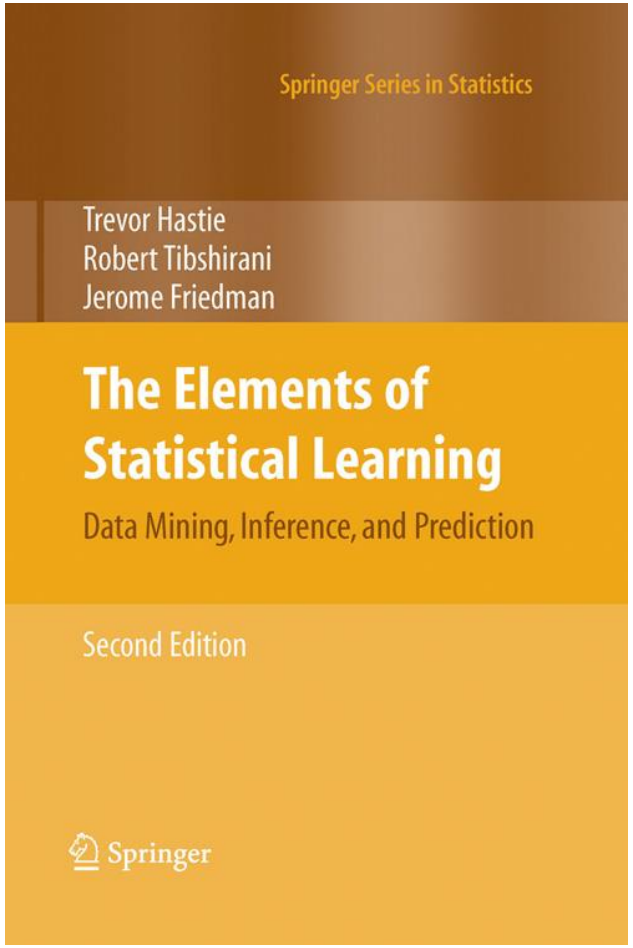

Springer Texts in Statistics
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R
Second Edition
Springer

An Introduction of Statistical Learning

- by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani
- Book website: https://www.statlearning.com/
  - Data sets, .R/Rmarkdown files
  - Slides and video lectures
  - Book PDF and errata


Springer Texts in Statistics
Gareth James · Daniela Witten · Trevor Hastie ·
Robert Tibshirani · Jonathan Taylor
An Introduction to Statistical Learning
with Applications in Python
Springer


统计学习（R语言版）
大数据时代的统计与人工智能系列教材
统计学习
（R语言版）
李高荣 编著
中国教育出版传媒集团
高等教育出版社


Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Second Edition
Springer

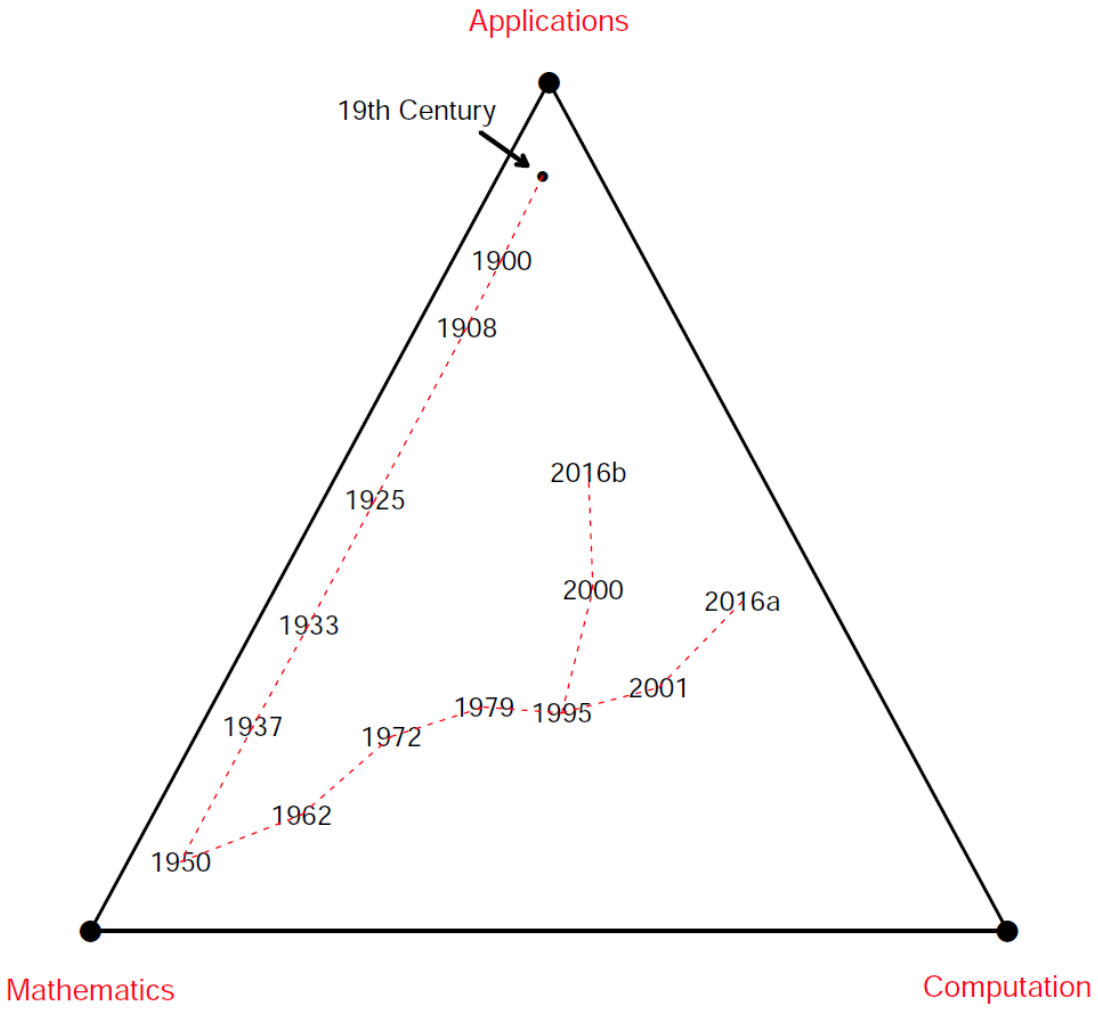
Grading policy (考核方法)

- Homework (30%)
  - 不按时交作业1次扣3分
- Group project (40%)
  - 大作业按贡献打分
- In-class quizz (30%)

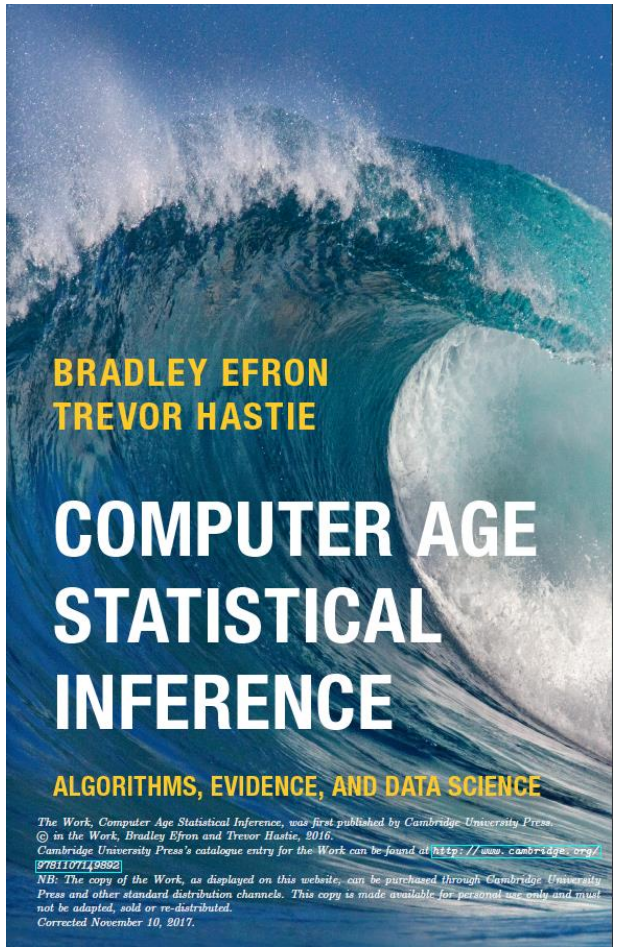
Group project

- A proposal (May 08)
  - Questions/problems, data sets, methods
- Oral presentations (June 05)
- A report (June 08)

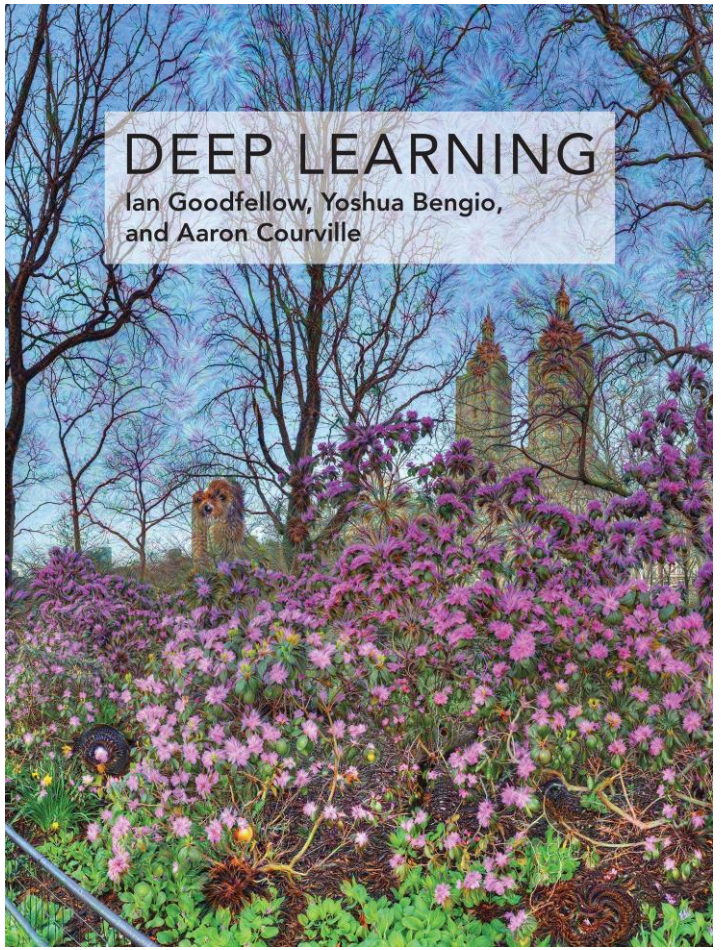
A brief history of statistical learning



Development of the statistics discipline since the end of the 19th century. This figure is taken from "*Computer Age Statistical Inference*" (Cambridge)


BRADLEY EFRON
TREVOR HASTIE
COMPUTER AGE
STATISTICAL
INFERENCE
ALGORITHMS, EVIDENCE, AND DATA SCIENCE
The Work, Computer Age Statistical Inference, was first published by Cambridge University Press.
© in the Work, Bradley Efron and Trevor Hastie, 2016.
Cambridge University Press's catalogue entry for the Work can be found at http://www.cambridge.org/9781107149892
NB: The copy of the Work, as displayed on this website, can be purchased through Cambridge University Press and other standard distribution channels. This copy is made available for personal use only and must not be adapted, sold or re-distributed.
Corrected November 10, 2017.


机器学习
周志华 著
MACHINE
LEARNING
机器学习
清华大学出版社


DEEP LEARNING
Ian Goodfellow, Yoshua Bengio,
and Aaron Courville