Syllabus

資料科學 Data Science
張家銘 Jia-Ming Chang
政治大學資訊科學系

Dr. Chuan Yi Tang

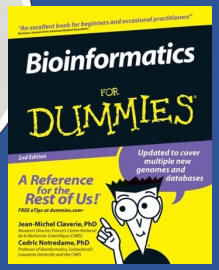
1996 ~ 2000 Bachelor (推薦甄試入學)
2002 ~2002 Master
@ Computer Science, National Tsing Hua Uni.

2002 ~ 2008 military replace service
@ Institute of Information Science
Academia Sinica

Dr. Ting-Yi Sung *Dr. Wen-Lian Hsu*

2008~2013 PhD La Caxia fellowship
@ The Centre for Genomic Regulation
Barcelona, Spain

Dr. Cedric Notredame



2014~2016 Postdoc
@ Institute of Human Genetics
Montpellier, France

Dr. Giacomo Cavalli


生物資訊
Bioinformatics for DUMMIES

https://goo.gl/photos/AT6QkCgfVH8JsmH69
https://goo.gl/photos/dBiRxWYxbWSsbCHA8

The Institute of Human Genetics
The French National Centre for Scientific Research

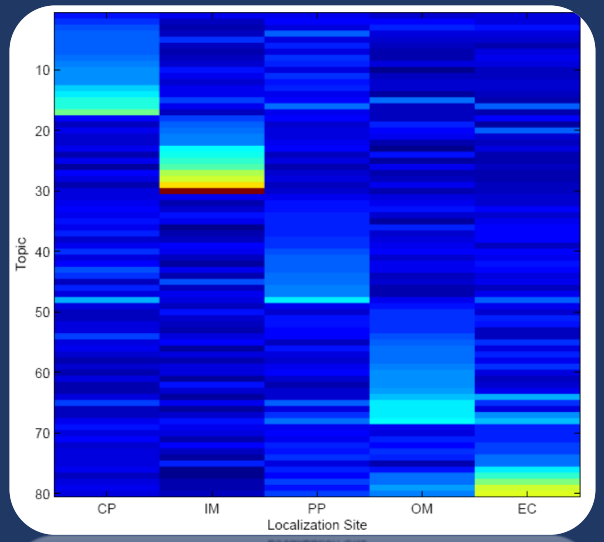
302 publications between 2008 and 2012
20 % of these research papers with an *IF* >10
2 *Nature*, 9 in other *Nature* series, 6 *Cell*, 2 *Science*, 7 *Genes & Dev*, 6 *Mol Cell*, 7 *EMBO J*, 5 *PNAS*
the mean of IF is 6.4

Genome Dynamics, Giacomo Cavalli's Lab
25> papers in high-impact journals
Cell, *Science*, *Nature Genetics*, *PLoS Biology*.
Awards:
- the silver medal of the CNRS
- EMBO fellow

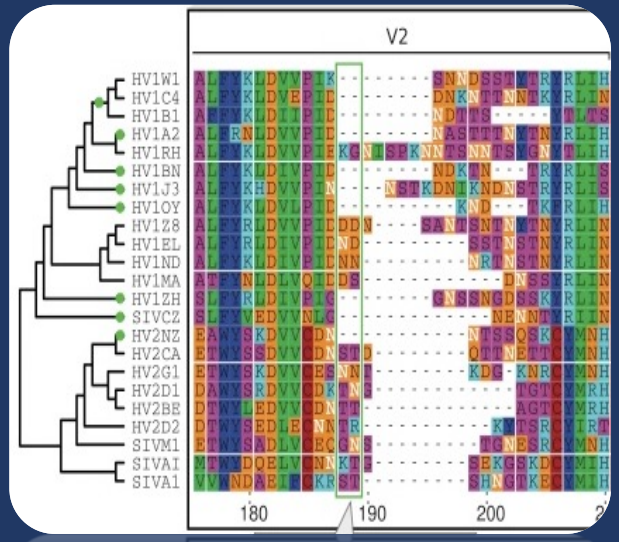
He is directly involved in the FP7 EpiGeneSys NoE as board member.
2008 ERC Advanced Investigator Grant, 2.2 million euro

Big/Biodata Science lab (BBS)
巨觀生物資料科學實驗室

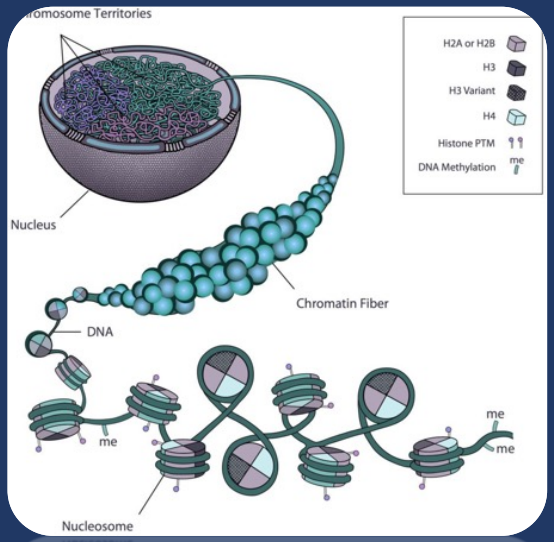
Computational protein function prediction



Large scale alignment & phylogenetic tree



3D organization of chromosomes



About the course

The course

- This course will introduce you to the work of data science
  - It is an introduction to an advanced topic.
  - We will concentrate on a portion of data science related to scoring and prediction.
- We will work examples with actual data using an analysis system called *R*
  - Lectures will be
    - Slides
    - On-hand programing

Scheduled progress

- https://www.changlabtw.com/1122-datascience.html
- https://www.changlabtw.com/1122-datascienceinservice.html

How many components?

Three components

1. Data
2. Modeling
3. Evaluation

Topic 1 – Evaluation

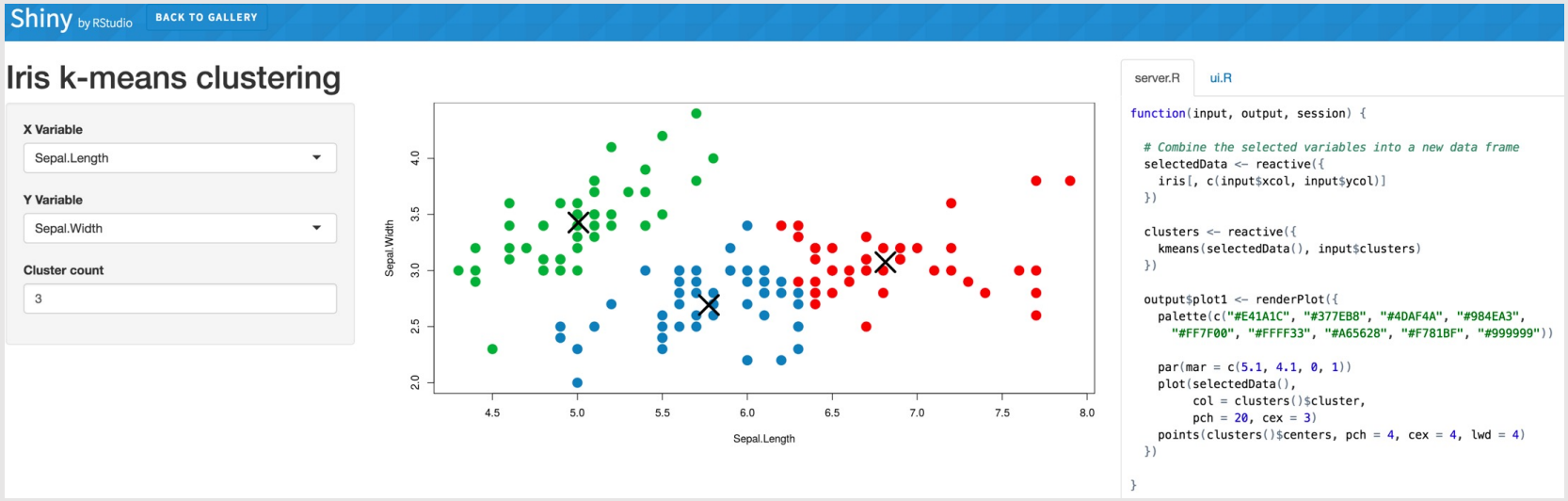
- Metrics
  - Specificity, sensitivity, RMSE, likelihood, NMI
  - ROC, AUC
  - Statistical significance : p-value, false discovery rate
- Perform evaluation
  - Overfitting: Bias-variance decomposition
  - Cross-validation
  - Statistical significance : p-value, false discovery rate

Topic 2 – Data

- Feature reduction
  - PCA, SVD, CA
  - PLSA
- Imbalance data

Topic 3: Visualization

- http://shiny.rstudio.com/gallery/kmeans-example.html



Topic 4: supervised v.s. unsupervised

- supervised statistical learning
  - involves building a statistical model for predicting, or estimating, an output based on one or more inputs
- unsupervised statistical learning
  - there are inputs but no supervising output; nevertheless we can learn relationships and structure from such data

Supervised : output?

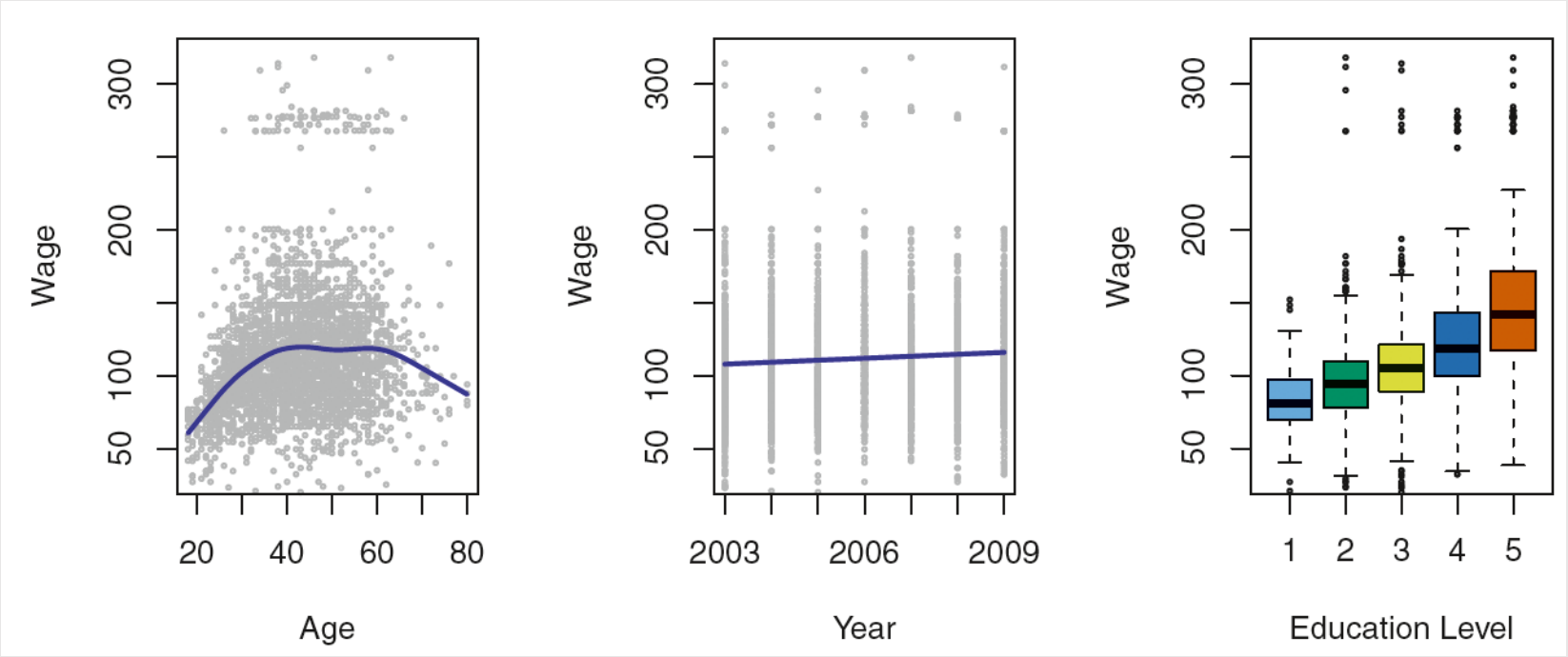
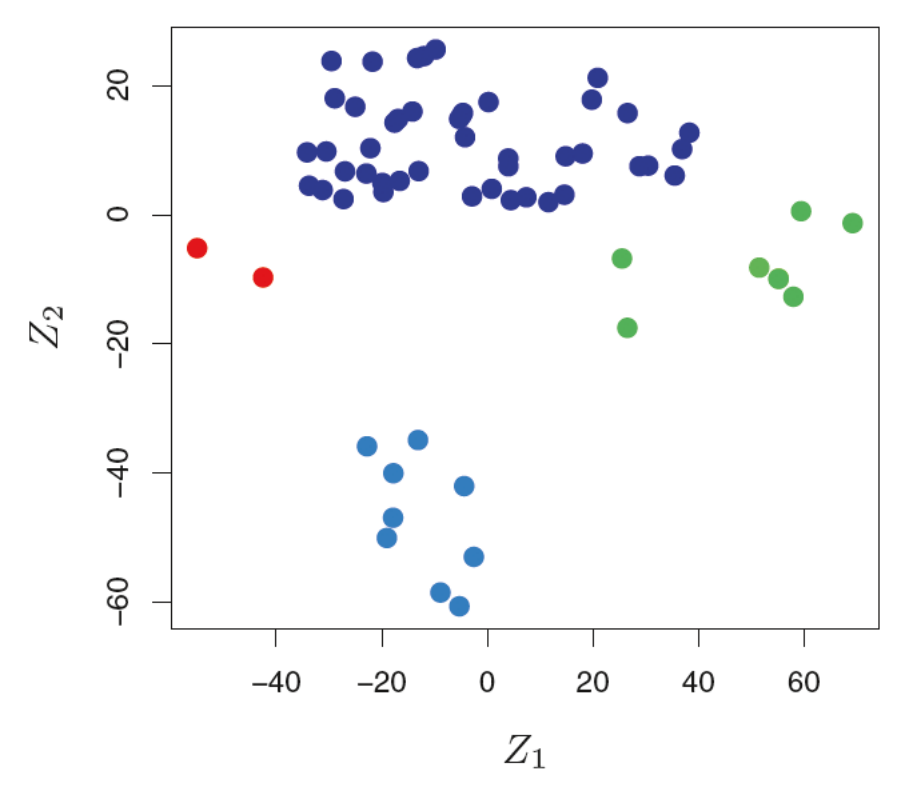


Figure 1.1, *An Introduction to Statistical Learning with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

Regression v.s. Classification



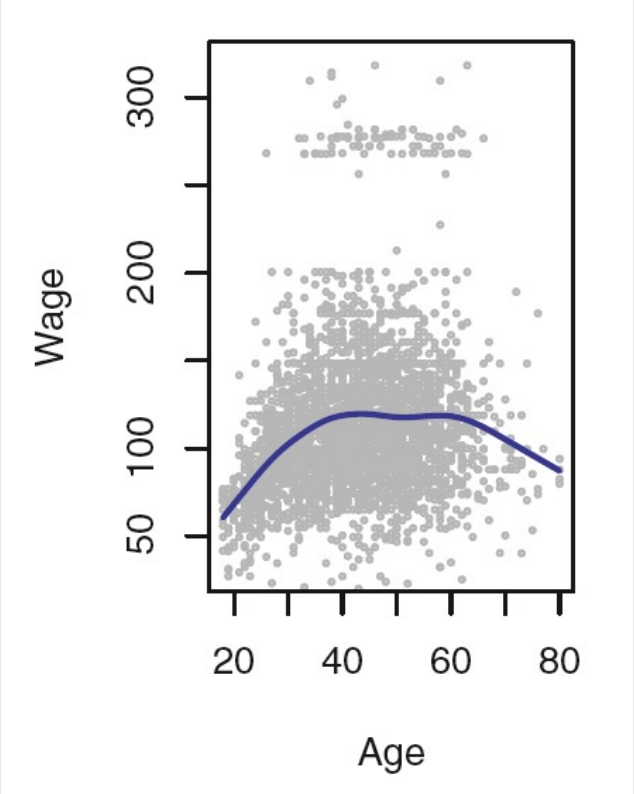
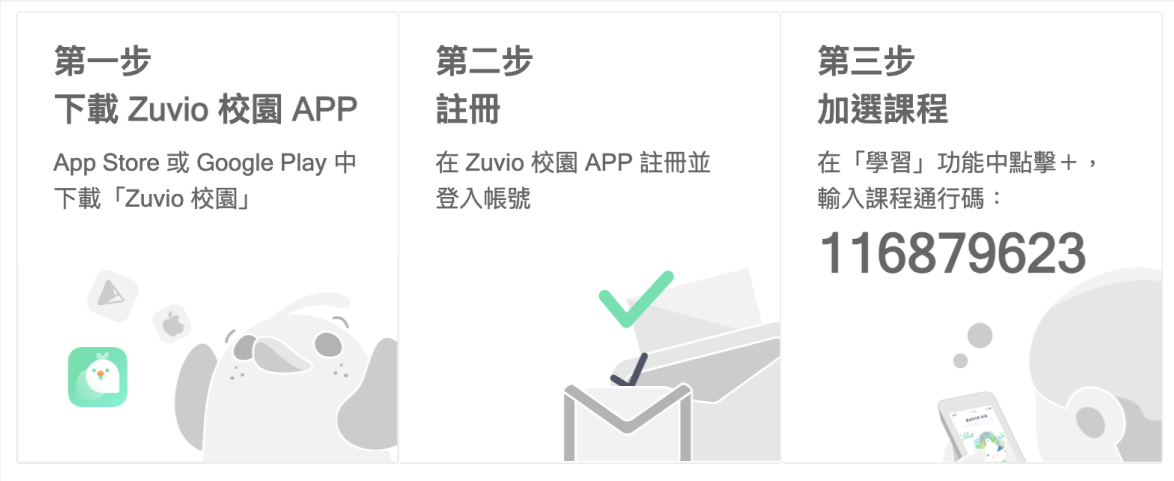


Figure 1.1 & 1.4, *An Introduction to Statistical Learning with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

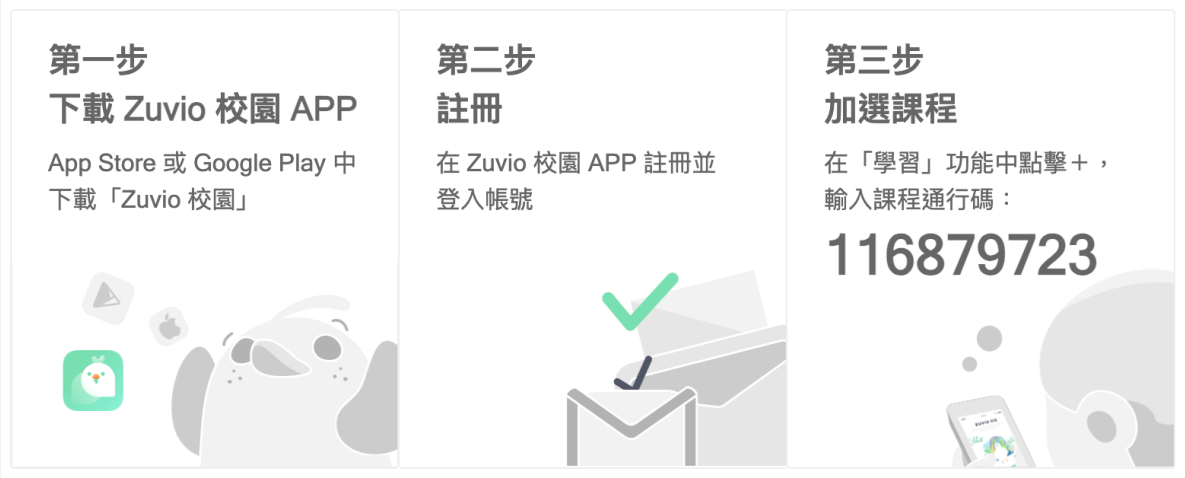
Course online

- Web site
  - https://coursedatascienechanglabtw.readthedocs.io/
- Moodle
- FB group
  - https://www.facebook.com/groups/nccu.datascience/
- Zuvio
  - 如果 nccu.edu.tw 登入，預設密碼身分證後5碼



Course online - in service

- Web site
  - https://coursedatascienechanglabtw.readthedocs.io/
- Moodle
- FB group
  - https://www.facebook.com/groups/nccu.datascience/
- Zuvio
  - 如果 nccu.edu.tw 登入，預設密碼身分證後5碼



Teaching methods

- Lecture
- MOOCs
- Hands-on Programming

What is not in the course?

What is not in this course?

- Big data (engineering) or hardware implementation
- How to implement your own machine learning algorithms
  - Except for one example we emphasize exploring
    and using already available machine learning libraries
    => thanks rich R package libraries

What is not in this course?

- Deep Learning
  - https://www.deeplearning.ai/
  - http://bangqu.com/Wi6Zfj.html

What is not in this course?

- Python, Julia, and friends
  - You won't learn anything about Python, Julia, or any other programming language useful for data science. This isn't because we think these tools are bad. They're not! And in practice, most data science teams use a mix of languages, often at least *R* and *Python*.

Course Grading Criteria

Grading standards

- 50 Homework
- 10 Zuvio questions
- 15 Midterm
- 25 Final Project
  - 5 Joint poster
  - 20 Oral presentation
- < 5 Bonus
  - actively attend course
  - ask questions

Grading standards – in service

- 50 Homework
- 15 Zuvio questions
- 15 Midterm
- 20 Final project
- < 5 Bonus
  - actively attend course
  - ask questions

Grading standards

- rounding into integer: round(x)
  - round(89.5) = 90
  - round(89.4) = 89

Homework

- *n* assignments, where $6 \leq n \leq 8$
  - 1/per 2|3 weeks
- Announce on Moodle
- Only accept *R* version 4
- zero score for don't match requirement, no excuse
  - 100 Rscript hw1_55688.R –input input.tsv –output output.tsv
  - 0 Rscript hw1_55688.R –input input.tsv –output output.tsv –feature v1

Homework Late Policy

- original score * 0.9
- at most twice for each homework
- must email TA after updating your Github
  - DO NOT contact using message
- we will check homework once every two weeks and reply email
- deadline will be is final presentation

TA

- Office: Da Ren building (大仁樓) 200403
- 高語謙 111753130@g.nccu.edu.tw

Midterm

- One A4 page note only
  - Handwriting or typing with a computer
- Can't use calculator/computer
- You can answer in English or Chinese.
- Absent is not accepted
  - Medical certification if temporary illness
  - There will be a more rigorous retest if needed

Final Project

- Collect your data before the midterm
  - From your own research project
  - Public data set
- Oral presentation + (joint Poster)
- Upload your code/document into GitHub

Final Project - Oral presentation

- 10 iVoting
- 10 Internal-group judgement

Final Score

國立政治大學

張家銘資料科學調分

政治大學 · 2021年11月26日 00:29

想請問張家銘老師的資料科學會調分嗎

今天拿到考卷

覺得跟預期的分數有點差距

想問老師以前會調分嗎

學期初有說的上課表現加分大概是怎麼算

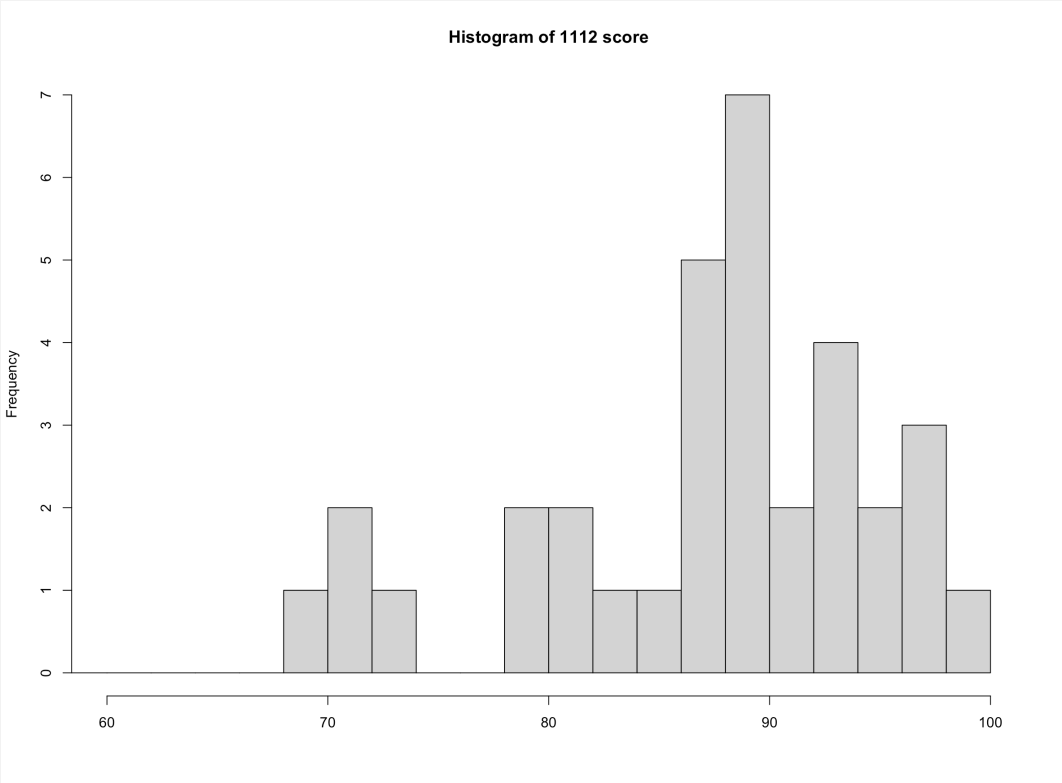
希望有人可以解答 謝謝

Final grade notification letter

If you have any questions about the scores, please contact me before 01/21 23:59, the results will be sent out on 01/22 (Friday).
Please refer to Moodle for the detailed results.
We will only correct the calculation errors of the scores and will not provide additional supplementary submissions.

111-2 score distribution

```
s<-read.csv("~/Dropbox/13_NCCU/courses/../1112_資料科學_期末總成績.csv")
hist(s$score, breaks = 40, xlim=c(60,100), xlab = "", main="Histogram of 1112 score")
```



recommendation letter / 推薦信

- Strong reference letter for top 3 students if you need.
- Examples
  - I came to know ??? by teaching him Data Science in the fall of 2021. Her course grade was as high as 95, the second highest in the class.
  - I came to know ??? by teaching him Data Science in the fall of 2021. His course grade was 87, ranking 13th in the class.

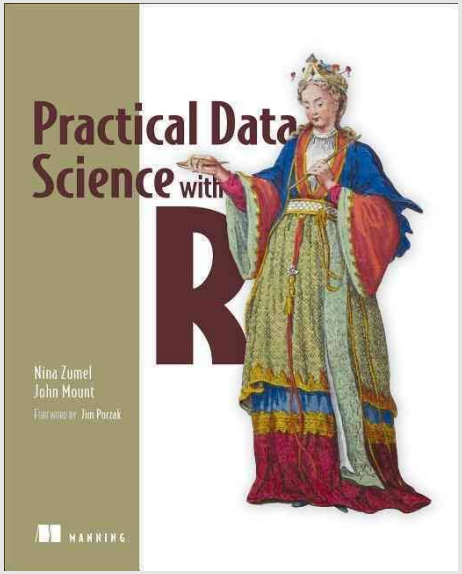
How to contact me?

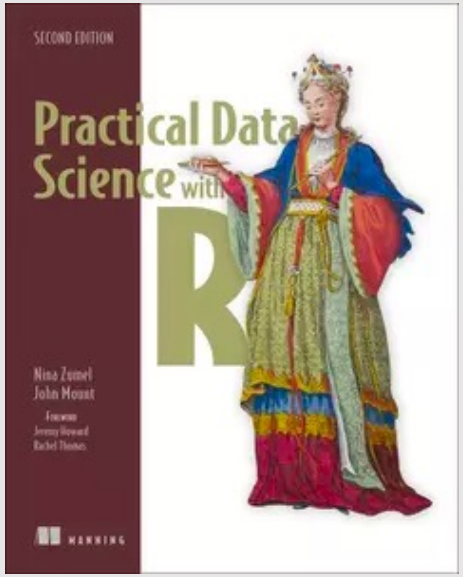
- Room 313, DaRen building (大仁樓)
- Email: chang.jiaming@gmail.com
- Subject:
  - [1122DataScience] yourname

Reference Books

Practical Data Science with R

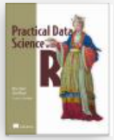
- Zumel, N. & Mount, J. (Manning)
  - 2th, 2019 @ 天瓏書局 1530
  - 1th, 2014 @ 天瓏書局 1320
- Example R scripts and data https://github.com/WinVector/zmPDSwR





Practical data science with R @ NCCU library

1



圖書

Practical data science with R / Nina Zumel, John Mount.

Zumel, Nina, author.; Mount, John (Computational scientist), author.

Shelter Island, NY : Manning Publications Co.; [2014]

可在 政大總圖/NCCU Main Library 總圖四樓西文圖書區 (006.3 Z94)獲得 ›

2

圖書

Practical data science with R / Nina Zumel and John Mount.

Zumel, Nina, author.; Mount, John (Computational scientist), author.

Shelter Island, NY : Manning Publications; [2020]

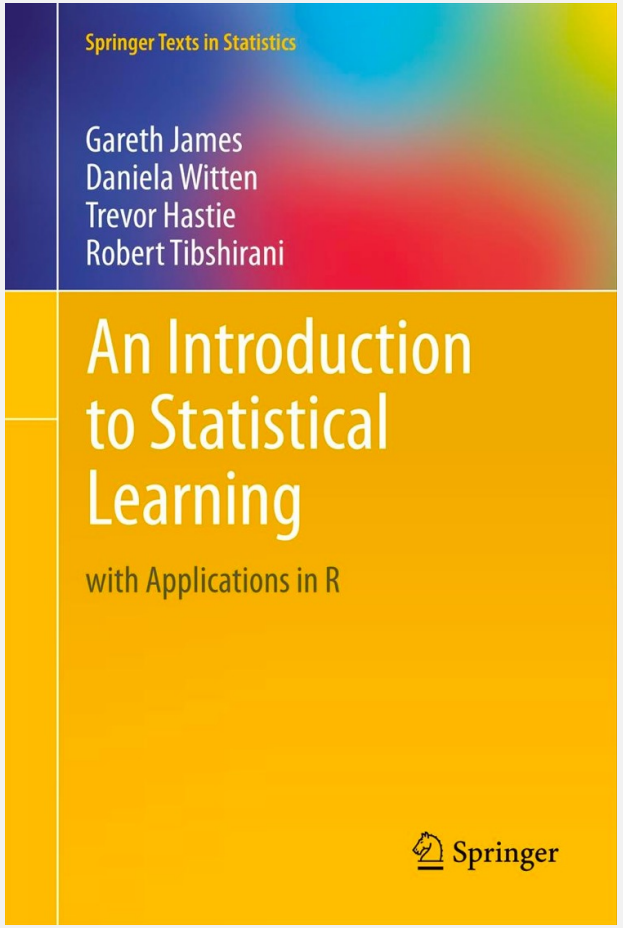
EBSCOhost ebooks

可在 政大總圖/NCCU Main Library 總圖四樓西文圖書區 (006.3 Z94 2020)獲得 ›

線上可獲得 ›

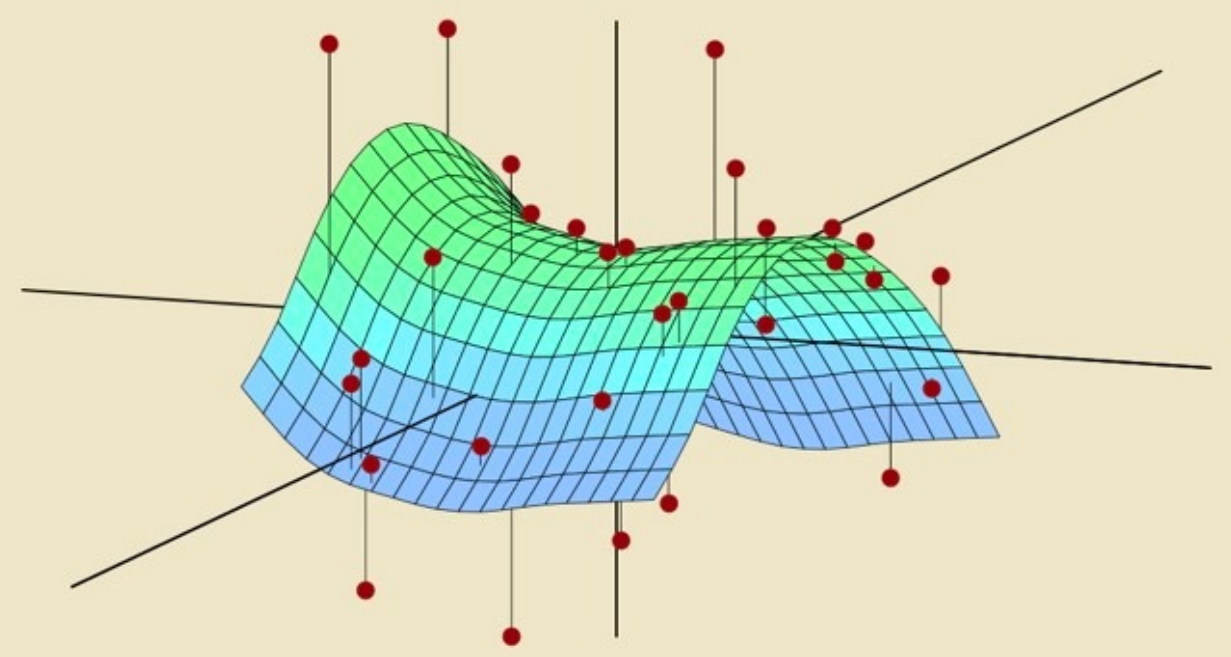
An Introduction to Statistical Learning with Applications in *R*

- by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani
- Hastie and Tibshirani coined the term *generalized additive models* in 1986 for a class of non-linear extensions to generalized linear models
- Download the book PDF



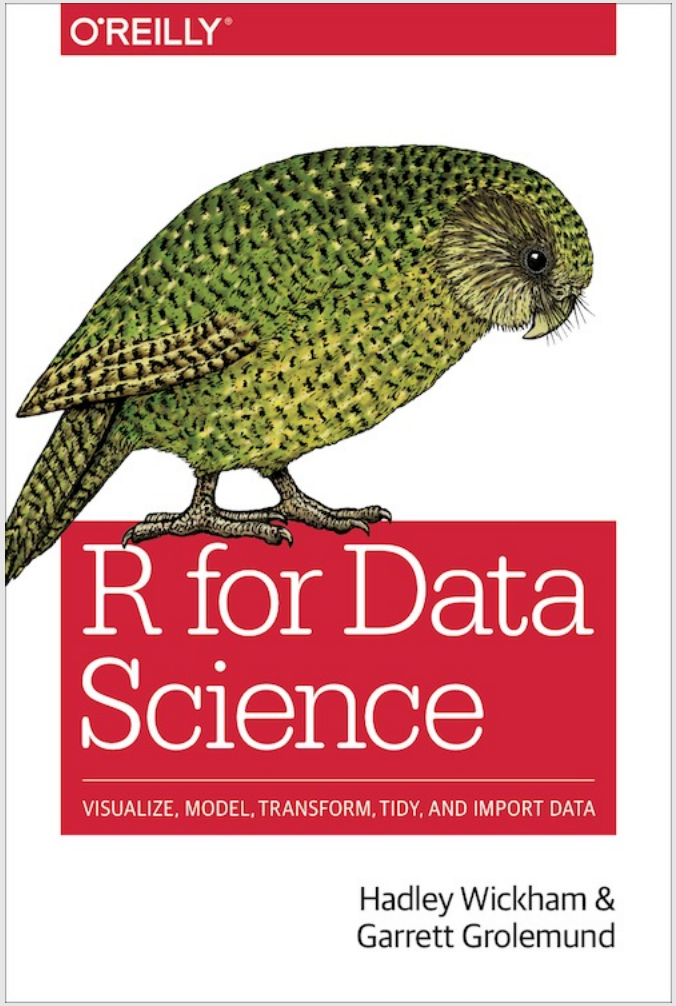
An Introduction to Statistical Learning with Applications in *R*

- Statistical Learning MOOC covering the entire ISL book offered by *Trevor Hastie* and *Rob Tibshirani*.
  - https://lagunita.stanford.edu/courses/HumanitiesSciences/StatLearning/Winter2016/about



R for Data Science

- Import, Tidy, Transform, Visualize, and Model Data
  - by Hadley Wickham, Garrett Grolemund
  - Hadley Wickham：一个改變了R的人
- Where?
  - Amazon
  - On-line
    - http://r4ds.had.co.nz/index.html



Science Misconduct

PubPeer

- https://pubpeer.com/

- G9a/RelB regulates self-renewal and function of colon-cancer-initiating cells by silencing Let-7b and activating the K-RAS/β-catenin pathway
  - Shih-Ting Cha, Ching-Ting Tan, Cheng-Chi Chang, Chia-Yu Chu, Wei-Jiunn Lee, Been-Zen Lin, Ming-Tsan Lin, Min-Liang Kuo, Nature Cell Biology (2016)
- https://pubpeer.com/publications/E3105C953203929608360C56F52950

Academic Ethics Guidelines for Researchers by the Ministry of Science and Technology

- Research misconduct: The scope of improper conduct of research covers a wide area. These guidelines are mainly concerned with the primary issue of violations of academic ethics, namely : fabrication, falsification, plagiarism, duplicate publication of research results, improper citations, illegal or inappropriate means is used to influence the scientific review of the paper, and listing the name of improper authors.

科技部對研究人員學術倫理規範

- 違反學術倫理的行為：研究上的不當行為包含範圍甚廣，本規範主要涵蓋 核心的違反學術倫理行為，即造假、變造、抄襲、研究成果重複發表或未 適當引註、以違法或不當手段影響論文審查、不當作者列名等。

Cheating

- Academic Integrity in Assignments
  - Any instance of sharing or utilizing someone else's code, such as allowing others to view your own work or inspecting another’s source code, is a violation of academic integrity. It is imperative that students safeguard their assignments to prevent unauthorized copying. Should there be evidence of code duplication, all individuals involved will be held equally accountable for the breach of conduct.
- Academic Integrity during Examinations
  - Exposing your exam responses to peers or viewing another student's work during an examination is strictly prohibited and will be treated as an academic offense.

Do not cheat !!!

- The first time
  - the assignment = 0
  - Final score -20
- The second occurrence
  - Final score = 0

AITC
教育部人工智慧技術及應用人才培育計畫
Artificial Intelligence Talent Cultivation Program