# Introduction & Data science platforms

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

# Copyright declaration 版權說明

- Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani.
- The web site of the book
- The credit of individual is indicated in the bottom part of the slide.
  - ie.,

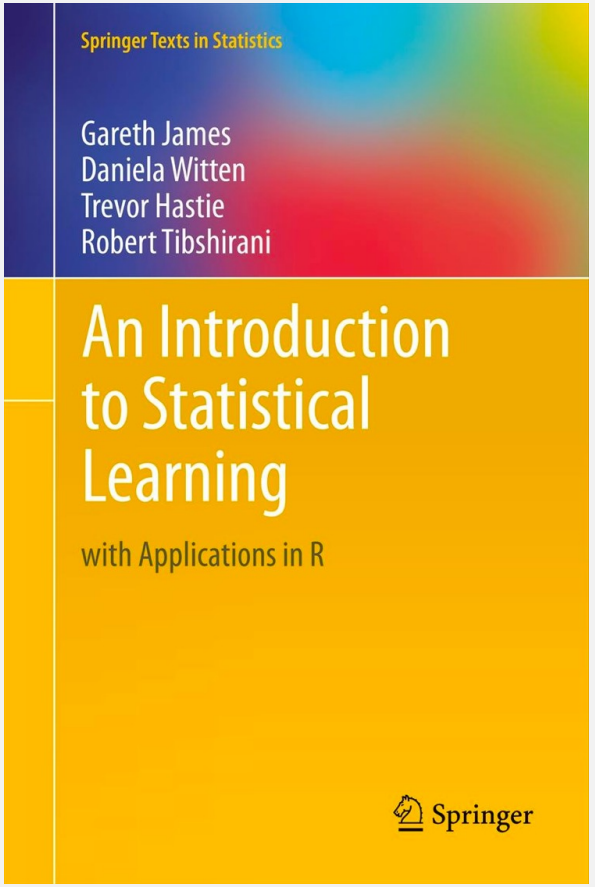



Figure 3.18, *An Introduction to Statistical Learning with Applications in R,* by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani

# Copyright declaration 版權說明

- Some of the figures in this presentation are taken from "Practical Data Science with R (Manning, 2019)"
- The web site of the book
- The credit of individual is indicated in the bottom part of the slide.
  - ie.,

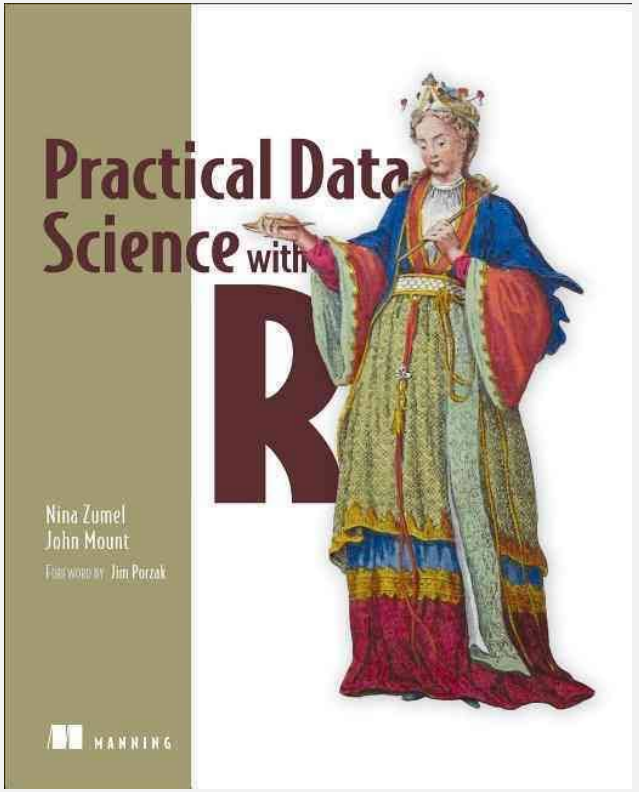
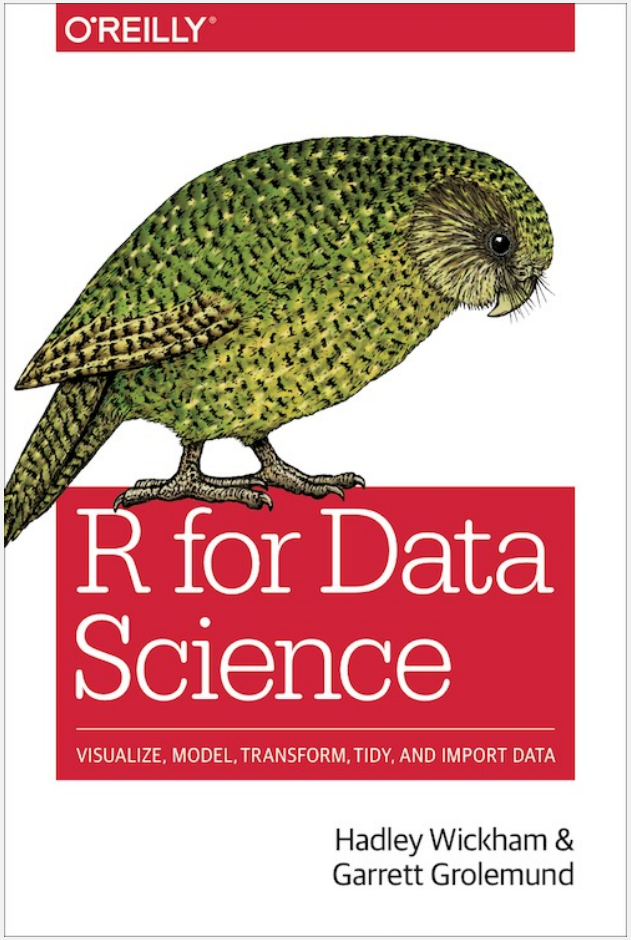



Figure 7.6, *Practical Data Science with R* by Nina Zumel and John Mount

# Copyright declaration 版權說明

- Some of the figures in this presentation are taken from "R for Data Science" under the Creative Commons Attribution-NonCommercial-NoDerivs 3.0 License.
- The web site of the book
- The credit of individual is indicated in the bottom part.
  - ie.,

*R for Data Science* by Garrett Grolemund, Hadley Wickham



# What is data science?

# Interdisciplinary

- The statistician *William S. Cleveland* defined data science as an interdisciplinary field larger than statistics itself.
  - statistics
  - machine learning
  - programming / computer science
  - data engineering

# Data to prediction

- as managing the process that can transform **hypotheses** and **data** into actionable **predictions**
  - predicting who will win an election
  - what products will sell well together
  - which loans will default
  - which advertisements will be clicked on

# Three components

- The data scientist is responsible for
  - Data : acquiring the data, managing the data
  - Modeling: choosing the modeling technique, writing the code
  - Evaluation: verifying the results

# Examples @ Job market

## Korean web giant Naver acquires makers of Whoscall

0 COMMENTS

Josh Horwitz
12:55 PM on Dec 9, 2013


Korean web giant Naver has acquired Gogolook, the Taiwan-based startup behind Whoscall, a popular app in East Asia that identifies the origins of unknown callers. The exact amount paid for the purchased has not yet been disclosed.

## Data Scientist - WhosCall 全職

職務分類：網頁程式設計師

工作地點：Taipei

薪資範圍： NT$ 60,000 - NT$ 90,000

張貼日期：2013-09-26

有效期限：2013-10-31

### World-Class Stage:

- It has been a year, Gogolook and our App "WhosCall" are now penetrating into Korea, Japan, South East Asia, India, and US. In order to expand our service spectrum and intensify the service quality, we need a Data Scientist like you to join our journey of disruption and innovation.

**We are also looking for Senior Web Engineer / Python Engineer，refer to Apply Information below.**

### Gogolook Introduction:

Gogolook is a fast growing mobile app startup located in Taipei, we are currently looking for passionate crews to re-invent the calling experience with us. We have a very connected team focusing on the most innovative ideas in the world and very clear vision to build a contact network of trust in every communication device.

WhosCall, the service we are focusing, has accumulated 5 million mobile users and identified billions of phone numbers. The service now has been expanded to Korea, Japan, South East Asia, and India and our incoming goal is to become the biggest collective phonebook and trusted contact network in Asia-Pacific. Therefore, we need a talent Data Scientist to join us.

### Must to have:

- Passion
- Passion
- Passion

### Optional Requirement:

- Solid stats background (familiar with various descriptive data analysis tools and hypothesis testing methods)
- Experience studying online user behavior (on top of exploratory/descriptive data analysis)
- Familiar with R language (capable of writing custom R functions when there is no built-in support in R)
- Familiar with Python, PHP, or any other scripting language (our goal is to standardize our data analysis toolchain)
- Familiar with NoSQL system

### Know more?

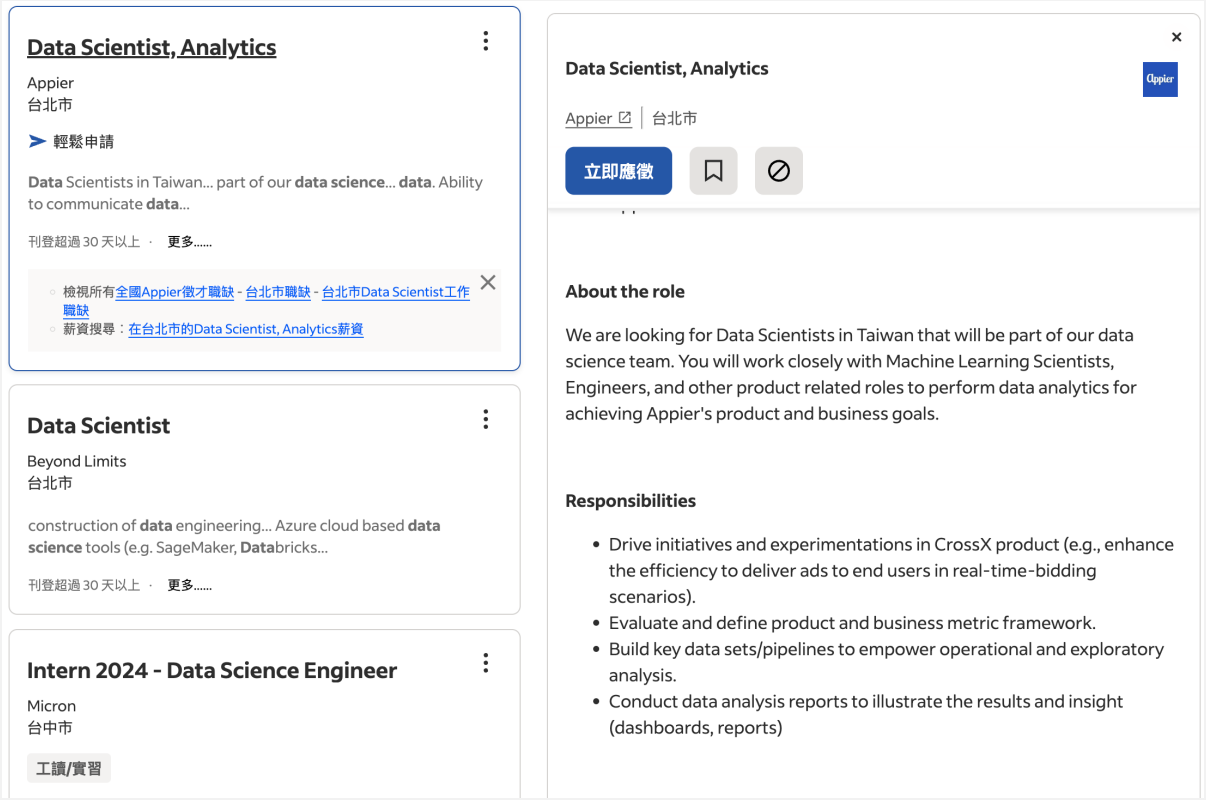
To know more about Gogolook and WhosCall, you can just Google us or refer to the URL below

Website：http://whoscall.com/
WhosCall Android: http://bit.ly/1548hqZ
WhosCall iOS: http://bit.ly/ZsQwy2

https://www.techinasia.com/korean-web-giant-naver-acquires-taiwanese-startup-gogolook

# Data science is the fastest growing industry

- https://tw.indeed.com/jobs?q=data+science&l=&from=searchOnHP&vjk=83a3a2c24e65e271



# Data scientist in Bioinformatics

- https://www.indeed.com/viewjob?jk=d244633e18f4d22c&tk=1hm42rdqkj210803&from=serp&vjs=3

**Research Software Engineer - Data Science**

Dana-Farber Cancer Institute ★★★★☆ 409 reviews

450 Brookline Avenue, Boston, MA 02215
Remote

Full-time

Job Company

You must create an Indeed account before continuing to the company website to apply

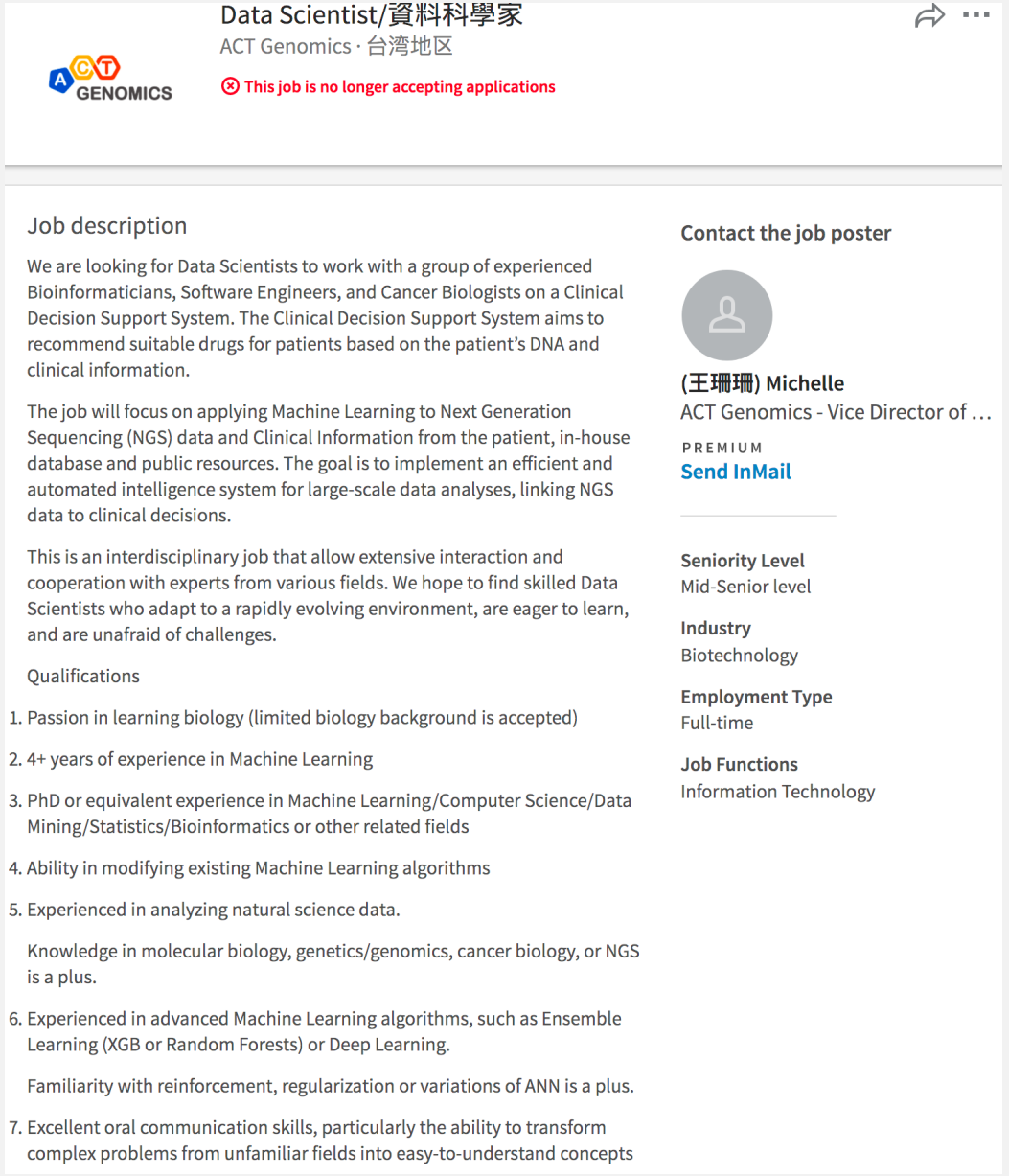
Apply now

Overview

The Department of Data Science at the Dana Farber Cancer Institute (DFCI) seeks candidates with a strong R programming background. As part of the department's mission to collaborate with basic biologists and clinical researchers to better understand cancer and improve treatment, our department develops new statistical methods and data analysis pipelines and implements these as R packages or shiny dashboards. We need help extending and improving these, as well as training our students, postdoctoral faculty, and collaborators in best practices and new developments related to R. We are seeking a software engineer to help with these challenges. The department chair will help prioritize projects and compartmentalize them into manageable units.

The successful candidate will have a unique opportunity to work in an exceptional collaborative environment with experts in a wide range of areas including clinical trials, cancer genetics, immunology, epigenetics, machine learning, Bayesian methods, and alignment algorithms. There is room for growth in this position as the career ladder permits promotion to levels that lead groups. We offer salaries that are competitive with the biotech industry. Remote work is a possibility. Please contact chair@ds.dfci.harvard.edu with questions about this role.

# Data scientist in Bioinformatics



Data Scientist/資料科學家

ACT Genomics · 台湾地区

This job is no longer accepting applications

## Job description

We are looking for Data Scientists to work with a group of experienced Bioinformaticians, Software Engineers, and Cancer Biologists on a Clinical Decision Support System. The Clinical Decision Support System aims to recommend suitable drugs for patients based on the patient's DNA and clinical information.

The job will focus on applying Machine Learning to Next Generation Sequencing (NGS) data and Clinical Information from the patient, in-house database and public resources. The goal is to implement an efficient and automated intelligence system for large-scale data analyses, linking NGS data to clinical decisions.

This is an interdisciplinary job that allow extensive interaction and cooperation with experts from various fields. We hope to find skilled Data Scientists who adapt to a rapidly evolving environment, are eager to learn, and are unafraid of challenges.

Qualifications

1. Passion in learning biology (limited biology background is accepted)
2. 4+ years of experience in Machine Learning
3. PhD or equivalent experience in Machine Learning/Computer Science/Data Mining/Statistics/Bioinformatics or other related fields
4. Ability in modifying existing Machine Learning algorithms
5. Experienced in analyzing natural science data.

   Knowledge in molecular biology, genetics/genomics, cancer biology, or NGS is a plus.
6. Experienced in advanced Machine Learning algorithms, such as Ensemble Learning (XGB or Random Forests) or Deep Learning.

   Familiarity with reinforcement, regularization or variations of ANN is a plus.
7. Excellent oral communication skills, particularly the ability to transform complex problems from unfamiliar fields into easy-to-understand concepts

Contact the job poster

(王珊珊) Michelle

ACT Genomics - Vice Director of …

PREMIUM

Send InMail

**Seniority Level**
Mid-Senior level

**Industry**
Biotechnology

**Employment Type**
Full-time

**Job Functions**
Information Technology

https://www.linkedin.com/jobs/view/370881654/?fref=gc&dti=183664932063716

# What is *R*?

- a programming language and software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. R is an implementation of the S programming language. (wikipedia)
- https://www.youtube.com/watch?v=XcBLEVknqvY

# Why choose *R* programming language?

- R's strong package ecosystem and charting benefits
  - https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis
- Pros and Cons of R vs Python Sci-kit learn by Kaggle

# Which is better for data analysis: R or Python?

- Both R and Python are excellent choices for data analysis, and which one you choose may depend on a variety of factors, including your specific needs, preferences, and prior experience.
- R is a language specifically designed for statistical computing and graphics. It has a rich set of built-in libraries for data manipulation, statistical analysis, and visualization. R's syntax is generally more concise and expressive than Python's, making it easier to perform complex data analyses with fewer lines of code.
- Python, on the other hand, is a more general-purpose language that is widely used in many different fields, including data science. It has a vast ecosystem of libraries and tools for data analysis, including NumPy, Pandas, and Scikit-learn. Python is also a more versatile language than R, allowing you to perform a wide range of tasks beyond data analysis, such as web development and automation.
- In summary, both R and Python are excellent choices for data analysis, and the choice ultimately depends on your specific needs and preferences. If you are interested in statistical analysis and data visualization, R may be the better choice. If you are interested in a more versatile language that can be used for a wide range of tasks, including data analysis, Python may be the better choice.

Which is better for data analysis: R or Python?

Both R and Python are excellent choices for data analysis, and which one you choose may depend on a variety of factors, including your specific needs, preferences, and prior experience.

R is a language specifically designed for statistical computing and graphics. It has a rich set of built-in libraries for data manipulation, statistical analysis, and visualization. R's syntax is generally more concise and expressive than Python's, making it easier to perform complex data analyses with fewer lines of code.

Python, on the other hand, is a more general-purpose language that is widely used in many different fields, including data science. It has a vast ecosystem of libraries and tools for data analysis, including NumPy, Pandas, and Scikit-learn. Python is also a more versatile language than R, allowing you to perform a wide range of tasks beyond data analysis, such as web development and automation.

In summary, both R and Python are excellent choices for data analysis, and the choice ultimately depends on your specific needs and preferences. If you are interested in statistical analysis and data visualization, R may be the better choice. If you are interested in a more versatile language that can be used for a wide range of tasks, including data analysis, Python may be the better choice.

# Data science in R is only a small subset of data science

- We are mostly teaching in an R context so we have a specific simple shared platform
- Most data scientists work using multiple platforms
- Other platforms include:
  - SAS
  - Python (pandas, scikit-learn)
  - Hadoop (Mahout)
  - SQL analytics
  - Microsoft Azure
  - And many others

# Data Science project

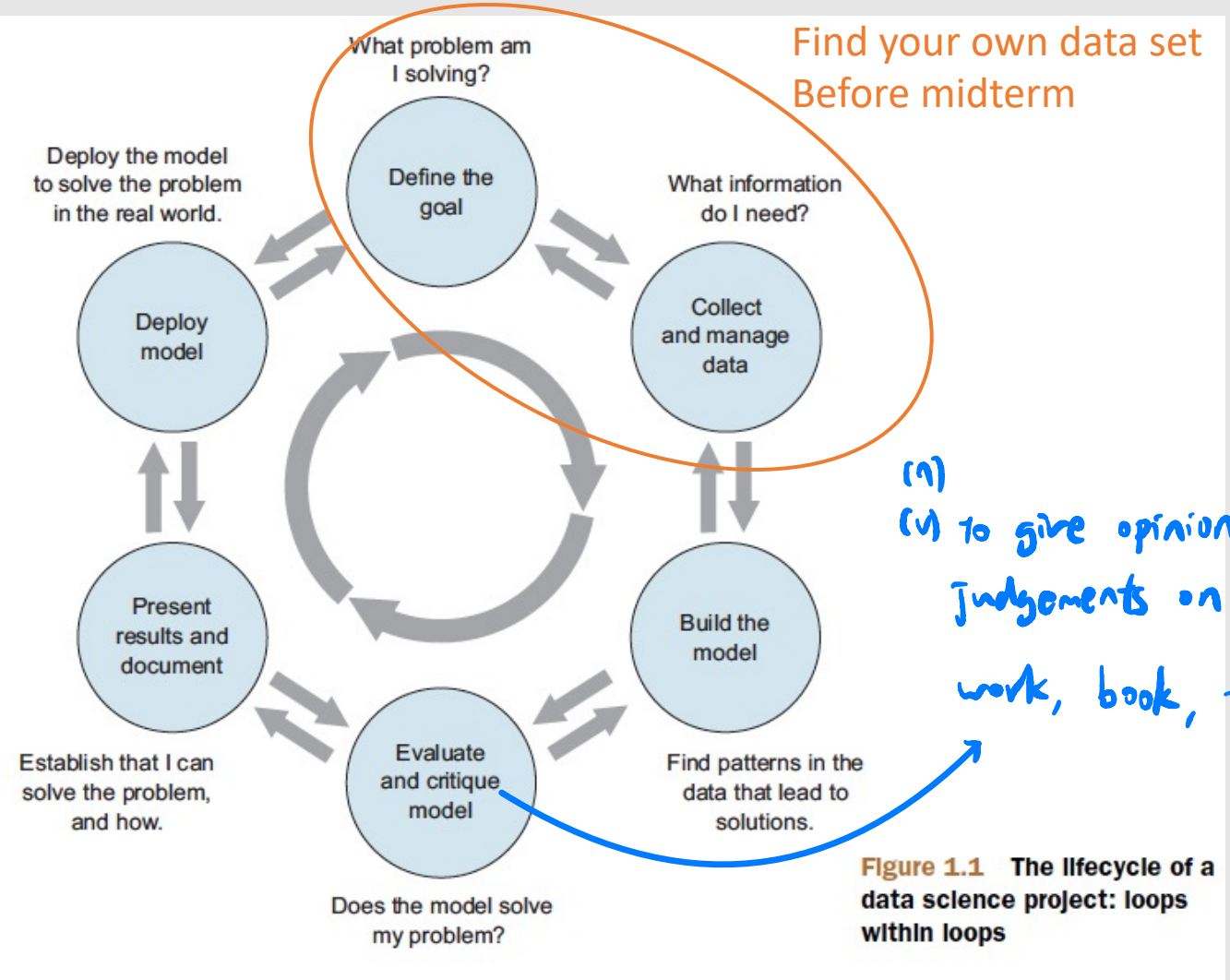



**Figure 1.1 The lifecycle of a data science project: loops within loops**

Figure 1.1, *Practical Data Science with R* by Nina Zumel, John Mount

# Modeling

- The most common data science modeling tasks are these:
  - Classification—Deciding if something belongs to one category or another
  - Scoring—Predicting or estimating a numeric value, such as a price or probability
  - Ranking—Learning to order items by preferences
  - Clustering—Grouping items into most-similar groups
  - Finding relations—Finding correlations or potential causes of effects seen in the data
  - Characterization—Very general plotting and report generation from data

(n) ① the way ppl are described in a film, book, play, etc. so they seem real and natural

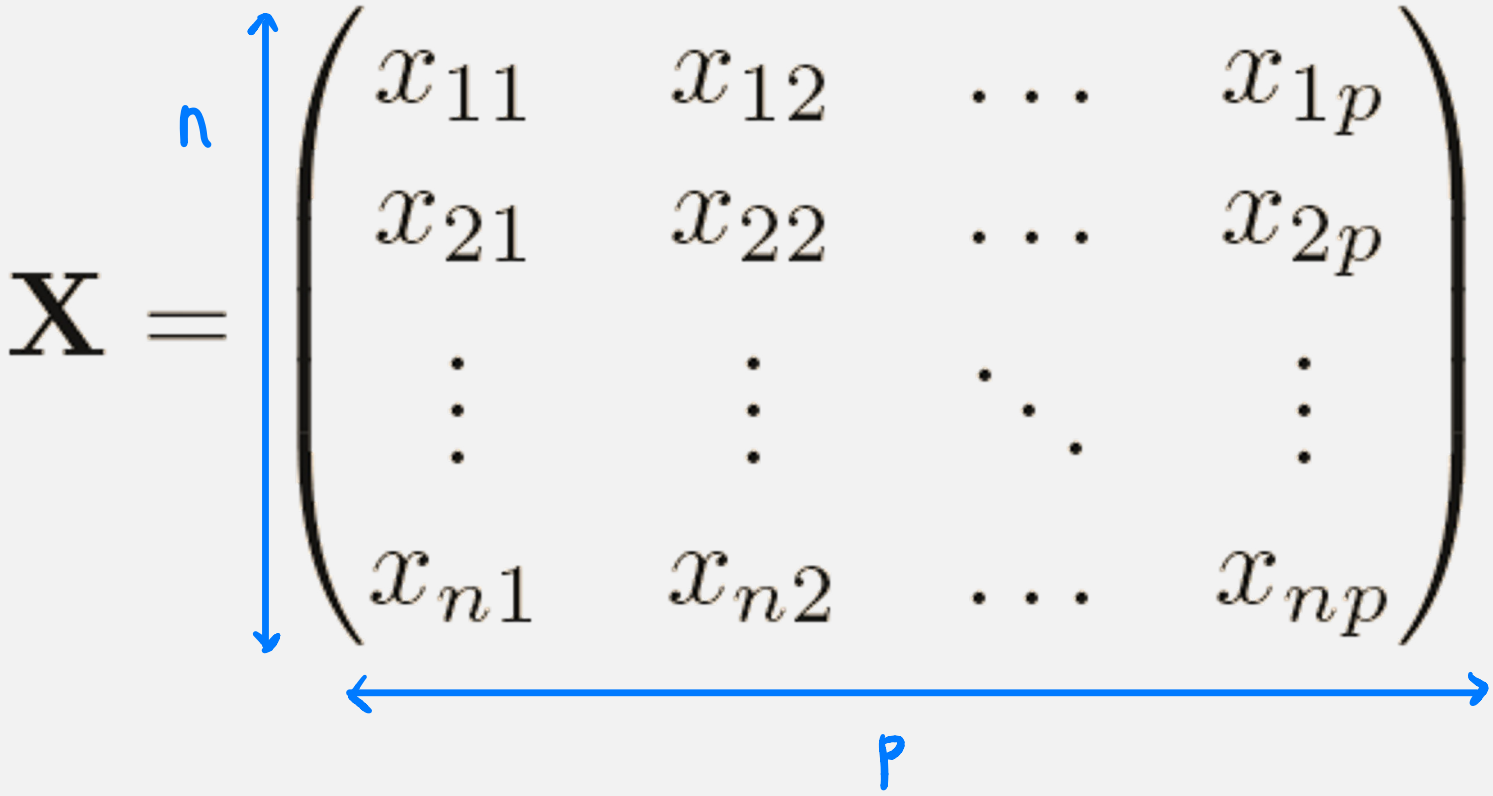
② the way in which sth is described by stating its qualities

# Statistical Learning

# Notation and Simple Matrix Algebra

- $n$ : the number of distinct data points, or observations, in our sample
- $p$ : the number of variables that are available for use in making predictions
- Wage data set consists of 12 variables for 3,000 people
  - $n$=? 3000
  - $p$=? 12

# $X$ denote a $n$ x $p$ matrix



# A person will be?

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix} \qquad \mathbf{x}_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix}$$

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix}.$$

$$\mathbf{AB} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix} = \begin{pmatrix} 1 \times 5 + 2 \times 7 & 1 \times 6 + 2 \times 8 \\ 3 \times 5 + 4 \times 7 & 3 \times 6 + 4 \times 8 \end{pmatrix} = \begin{pmatrix} 19 & 22 \\ 43 & 50 \end{pmatrix}$$

$$(\mathbf{AB})_{ij} = \sum_{k=1}^{d} a_{ik} b_{kj}.$$

# What is input? Output?

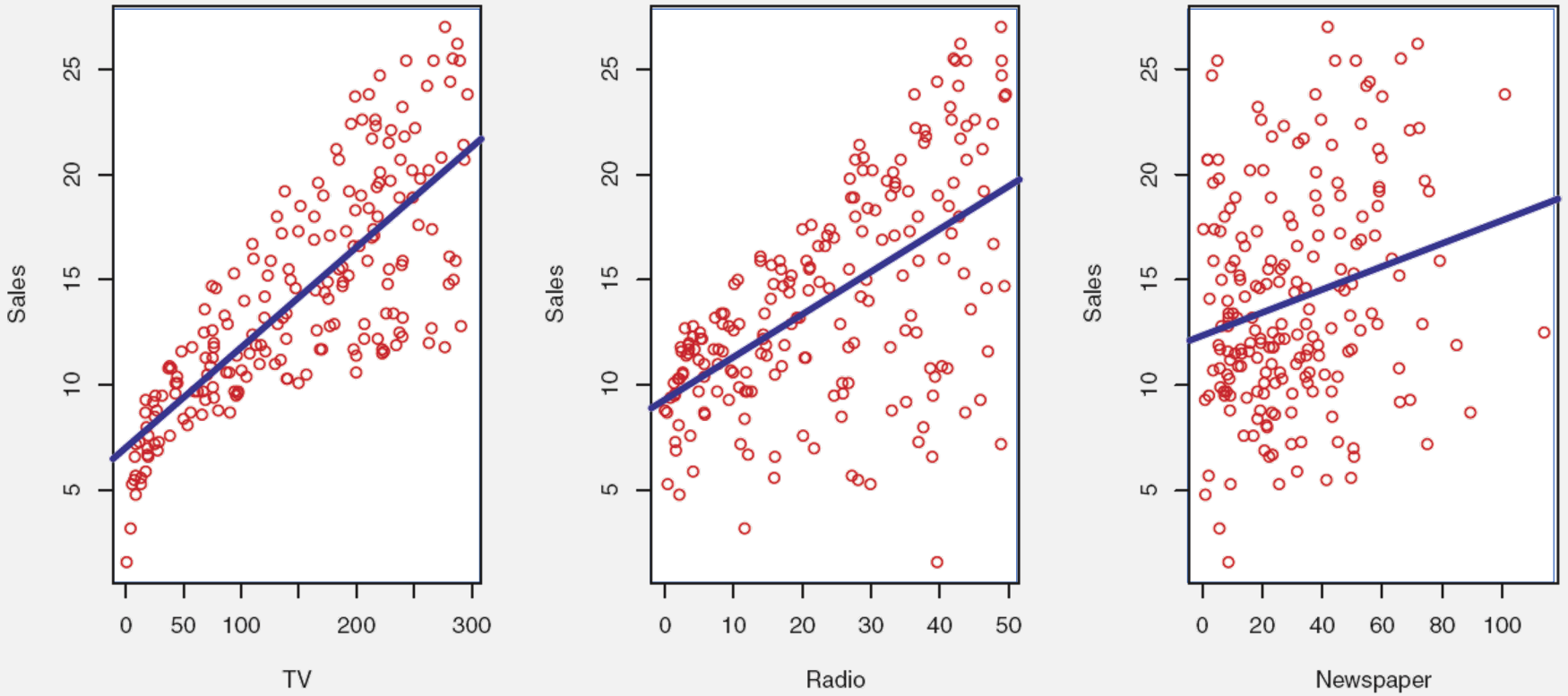



Figure 2.1, *An Introduction to Statistical Learning with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie,Robert Tibshirani

# The advertising data set

- The plot displays sales, in thousands of units, as a function of TV, radio, and newspaper budgets, in thousands of dollars, for 200 different markets.

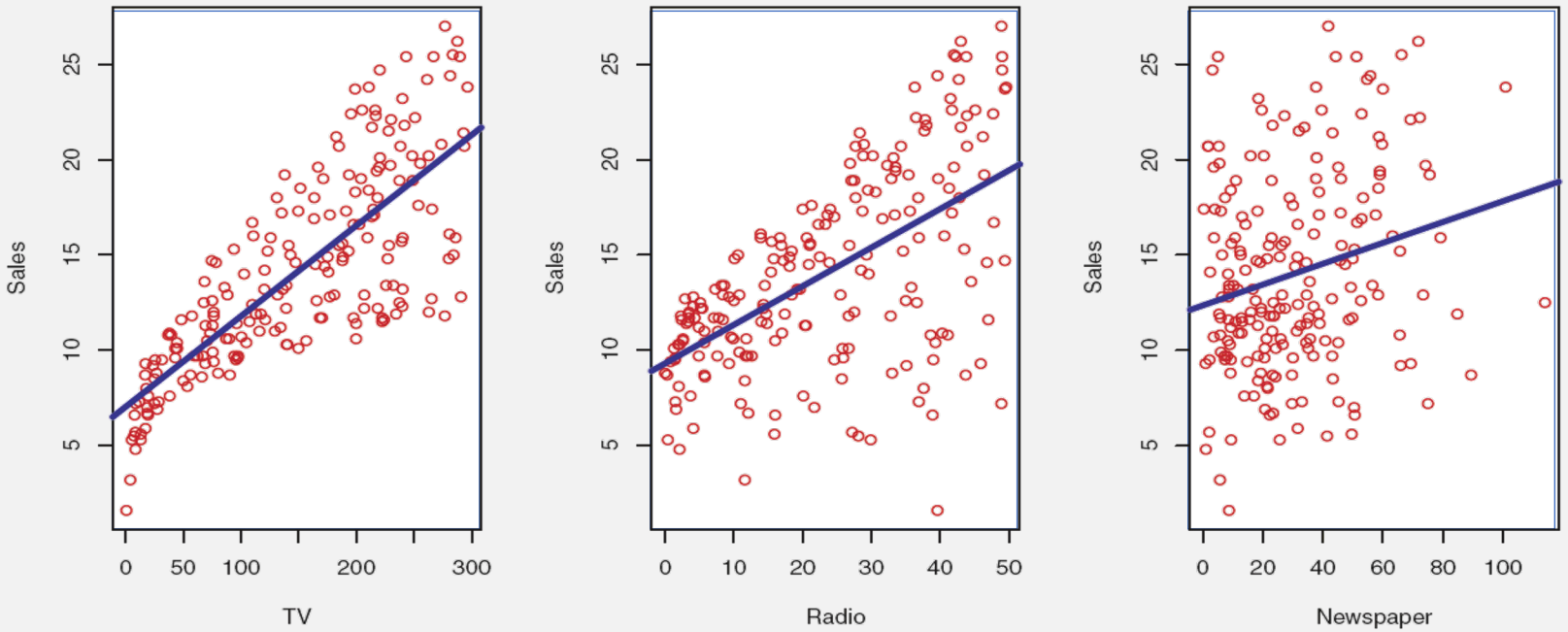



Figure 2.1, *An Introduction to Statistical Learning with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

# The advertising data set

- TV, radio, newspaper
  - predictors , independent variables , features , variable
  - typically denoted using the variable symbol X
- Sales
  - response , dependent variable
  - typically denoted using the symbol Y

# Modeling

- Given $Y$ and $p$ different predictors, $X_1$ , $X_2$ ,…,$X_p$
  - $f$ : some fixed but unknown function of $X_1$,…,$X_p$
  - $\varepsilon$ : a random error term, which is independent of $X$ and has mean zero
  - $f$ represents the systematic information that $X$ provides about $Y$.

# The income data set

The red dots are the observed values of income (in tens of thousands of dollars) and years of education for 30 individuals.

- What is *f*?
- overall, the errors have approximately mean zero???

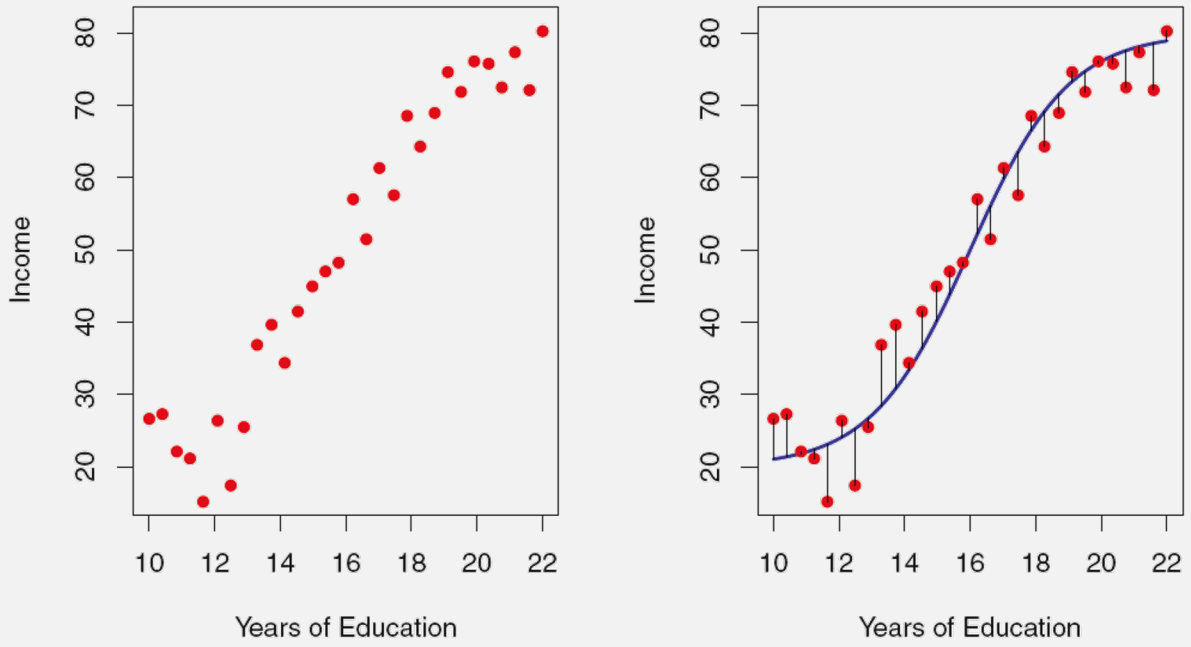



Figure 2.2, *An Introduction to Statistical Learning with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie,Robert Tibshirani

# the function $f$ may involve more than one input variable

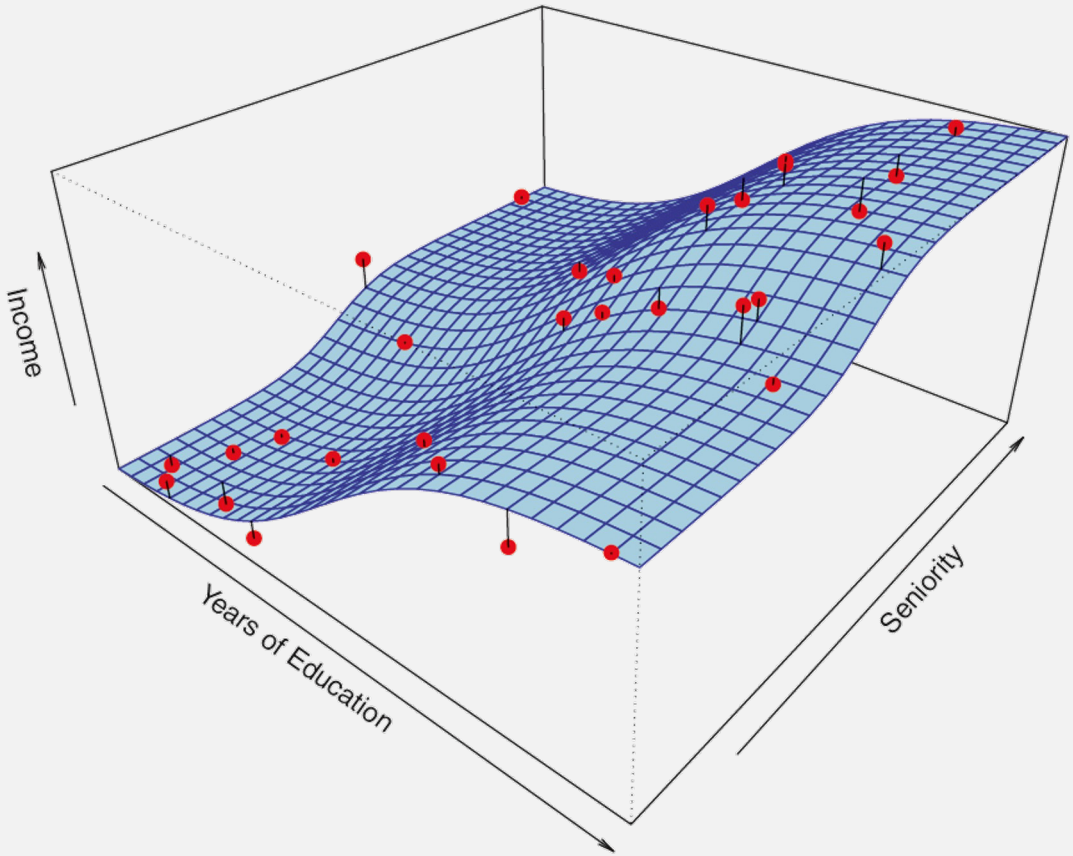



Figure 2.3, *An Introduction to Statistical Learning with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie,Robert Tibshirani

# Practice Time

# Installing R

- CRAN : the comprehensive R archive network
  - the central repository for the most popular R libraries & serves the central role for R
  - A new major version of R comes out once a year, and there are 2-3 minor releases each year.
  - https://cloud.r-project.org
    - Instead use the cloud mirror, which automatically figures it out

# Installing RStudio

- An integrated development environment, or IDE, for R programming.
  - https://www.rstudio.com/products/rstudio/download/

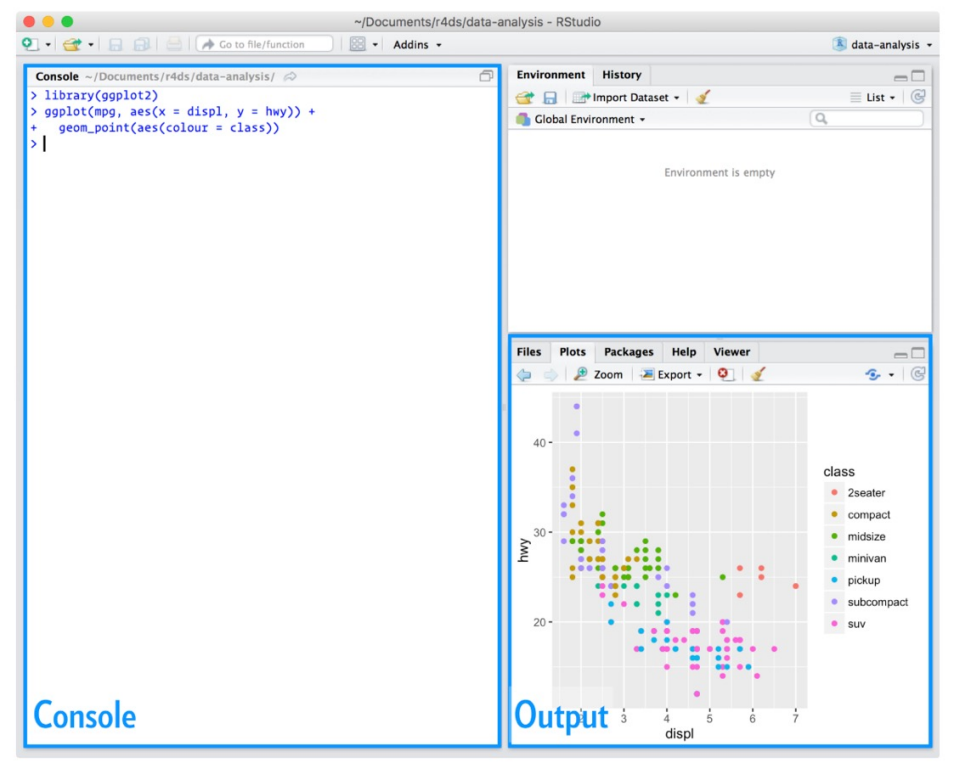



Figure 1.4.2, *R for Data Science* by Garrett Grolemund, Hadley Wickham

# RStudio IDE Cheat Sheet

- https://www.rstudio.com/wp-content/uploads/2016/01/rstudio-IDE-cheatsheet.pdf
- The RStudio IDE is the most popular integrated development environment for R.
  - Do you want to write, run, and debug your own R code?
  - Work collaboratively on R projects with version control?
  - Build packages or create documents and apps?
- No matter what you do with R, the RStudio IDE can help you do it faster. This cheat sheet will guide you through the most useful features of the IDE, as well as the long list of keyboard shortcuts built into the RStudio IDE. Updated 01/16.

# Work clean

- To start with an empty workspace and explicitly bring in the packages, code, and data you want. This ensures you know how to get into your ready-to-go state (as you have to perform or write down the steps to get there) and you aren't held hostage to state you don't know how to restore (what we call the "no alien artifact" rule).
- RStudio > Preferences, Tools > Global Options, Tools > Options

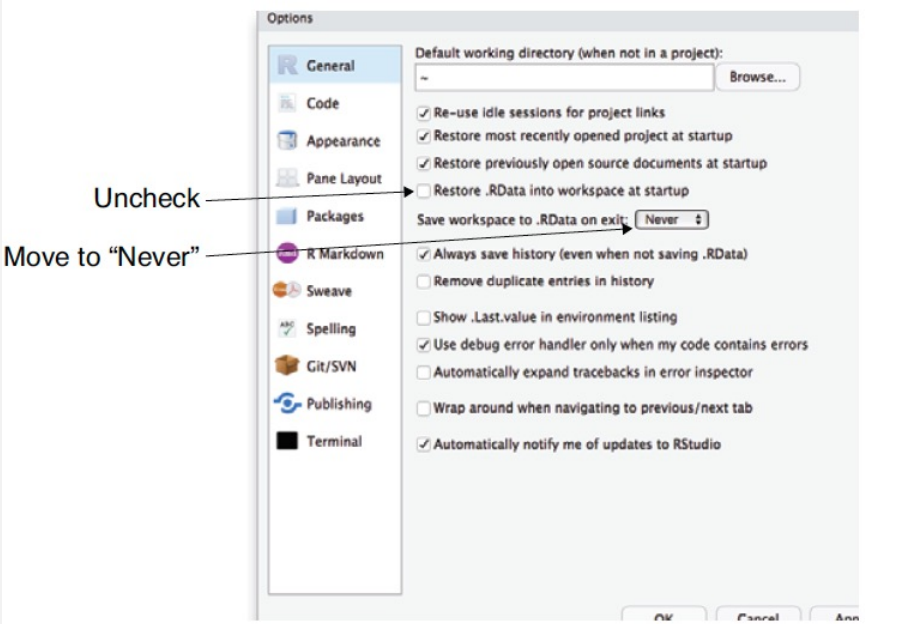



Figure A.4 RStudio options

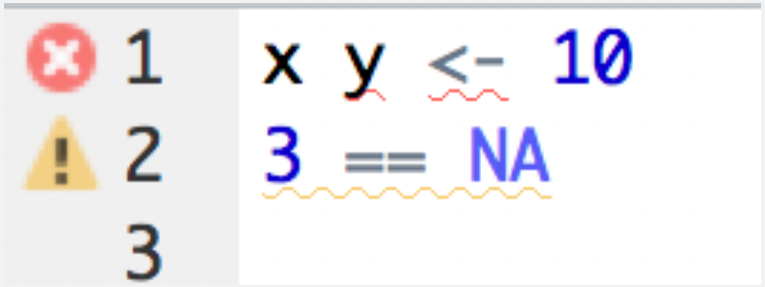
# Running Code

- Cmd/Ctrl + Enter : executes the current R expression in the console
- runningCode.R

```
1  library(dplyr)
2  library(nycflights13)
3
4  not_cancelled <- flights %>%
5    filter(!is.na(dep_delay), !is.na(arr_delay))
6
7  not_cancelled %>%
8    group_by(year, month, day) %>%
9    summarise(mean = mean(dep_delay))
```

# RStudio diagnostics

- The script editor will also highlight syntax errors with a red squiggly line and a cross in the sidebar
- diagnostics.R



# Basic Commands

- R uses functions to perform operations.
- funcname(input1, input2)

# Try the help command

- Start R or RStudio and type help(ls) to get
  - documentation on the ls command used in our example
  - ?funcname

# Starting with R

- How to use package?
  - install.package('ctv')
  - library('ctv')
- How many packages?
  - https://cran.r-project.org/web/views/

package : a collection of R functions, data and compiled code

library : the location where a package is stored

# Basic Commands

vector

- x <- c(1,3,2,5)
- x=c(1,6,2)
- y=c(1,4,3)
- length(x)
- length(y)
- x+y

return a vector of character strings containing all the variables and functions defined in the current working directory

# Basic Commands

- `ls()`
- `rm(x,y)` : delete variables from a workspace
- `ls()`
- `rm(list=ls())` : clear all objects from a workspace

# Basic Commands

- ?matrix
- x=matrix(data=c(1,2,3,4) , nrow=2, ncol =2)
- x=matrix(c(1,2,3,4) ,2,2)
- matrix(c(1,2,3,4) ,2,2,byrow =TRUE)
- sqrt(x) square root
- x^2

$\begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix}$

$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$

# Basic Commands

# of observations

- `x=rnorm(50)`
- `y=x+rnorm(50, mean=50, sd=.1)`
- `cor(x,y)`

x, y: vectors or matrices

- var: variance
- cov: covariance
- cor: correlation

# Basic Commands

- `set.seed(1303)`
- `rnorm(50)`

# Basic Commands

- `set.seed (3)`
- `y=rnorm (100)`
- `mean(y)`
- `var(y)`
- `sqrt(var(y))`
- `sd(y)`

# *sessionInfo()*

- what packages are present in your session
- Very information for reproducing your analysis

=> keep essential information when writing paper

# R語言翻轉教室

- http://datascienceandr.org/
  - By Wush Wu、Chih Cheng Liang、Johnson Hsieh

# References

- https://twitter.com/rstudiotips
  - RStudio Tips twitter account
  - find one tip that looks interesting. Practice using it!
- https://support.rstudio.com/hc/en-us/articles/205753617-Code-Diagnostics
  - What other common mistakes will RStudio diagnostics report?

# Getting help and learning more

- If you get stuck, start with Google. Typically adding “R” to a query is enough to restrict it to relevant results
- If Google doesn’t help, try stackoverflow. Start by spending a little time searching for an existing answer, including [R] restrict your search to questions and answers that use R.

# Getting help and learning more

- If you get stuck, start with Google. Typically adding “R” to a query is enough to restrict it to relevant results
- If Google doesn’t help, try stackoverflow. Start by spending a little time searching for an existing answer, including [R] restrict your search to questions and answers that use R.

- Webs
  - Stack Overflow R section : A Q&A site: http://stackoverflow.com/questions/tagged/r
  - LearnR : A translation of all the plots from Lattice: Multivariate Data Visualization with R (Use R!) (by D. Sarker; Springer, 2008) into ggplot2: http://learnr.wordpress.com
  - R-bloggers : A high-quality R blog aggregator: http://www.r-bloggers.com
  - Courses http://dataology.blogspot.tw/
- R programming
  - Norman Matloff, The Art of R Programming
  - Garrett Grolemund, Hands-On Programming with R
- R plus statistics
  - Robert Kabacoff R in Action (2nd edition) Quick-R http://www.statmethods.net/
  - Jared P. Lander R for Everyone
- Data Science
  - Cathy O'Neil, Rachel Schutt Doing Data Science
  - Nina Zumel, John Mount Practical Data Science with R
- Machine Learning
  - James et. al. An Introduction to Statistical Learning
  - Haste et. al. The Elements of Statistical Learning
- Free ebooks @ http://dataology.blogspot.tw/2015/09/60.html


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