

# More data structures and R packages



David Li

# Matrix

---

- There are several ways to make a matrix
- To make a 2x3 (2 rows, 3 columns) matrix of 0's:  
`>mat<-matrix(0,2,3)`
- To make the following matrix:

71	172
73	169
69	160
65	130

```
>mat2<-rbind(c(71,172),c(73,169),c(69,160),c(65,130))  
>mat3<-cbind(c(71,73,69,65),c(172,169,160,130))
```

- To make the following matrix:
  - `mat4<-matrix(1:10,2,5, byrow=T)`

1	2	3	4	5
6	7	8	9	10 <sub>2</sub>

# Revisit vectors: access data

---

- Accessing individual observations

  - `>whales[2]`

- Slicing

  - `>whales[2:5]`

- Negative indices

  - `>whales[-1]`

- Logical values

  - `>whales[whales > 100]`

  - `>which(whales > 100)`

  - `>which.max(whales)`

# Indexing of vector/matrix

---

## □ `x=1:10`

`ith` element

all *but* `ith` element

first `k` elements

specific elements.

all greater than some value

bigger than or less than some values

`x[2]` ( $i = 2$ )

`x[-2]` ( $i = 2$ )

`x[1:5]` ( $k = 5$ )

`x[c(1,3,5)]` (First, 3rd and 5th)

`x[x>3]` (the value is 3)

`x[ x< -2 | x > 2]`

## □ `mat=matrix(1:24, nrow=4)`

`mat[,2]` # 2<sup>nd</sup> column

`mat[2,]` # 2<sup>nd</sup> row

`mat[c(2,4),]` # 2<sup>nd</sup> and 4<sup>th</sup> row

`mat[1:3,1]` # 1 to 3 element in column 1

`mat[-c(2,4),]` # all but row 2 and 4

# Create logical vectors by conditions

---

- ❑ Logical operators: `<`, `<=`, `>`, `>=`, `==`, `!=`
- ❑ Comparisons
  - Vectors: `AND &`; `OR |`
  - Longer forms `&&`, `||`: return a single value
  - `all()` and `any()`
- ❑ Examples
  - `X=1:5`
  - `X<5`; `X>1`
  - `X >1 & X <5`; `X >1 | X <5`;
  - `all(X<5)`; `any(X>1)`; `all(X<5) && any(X>1)`
- ❑ `%in%` operator: `x %in% c(2,4)`

# Missing values

---

- ▣ R codes missing values as NA
- ▣ `is.na(x)` is a logical function that assigns a T to all values that are NA and F otherwise

```
>x[is.na(x)]<-0
```

```
>mean(x, na.rm=TRUE)
```

# Reading in other sources of data

---

## ▣ Use R's built-in libraries and data sets

```
> range(lynx) #lynx is a built-in dataset  
> library(MASS) # load a library  
> data(survey) # load a dataset in the library  
> data(survey, package="MASS") #load just data  
> head(survey)  
> tail(survey)
```

## ▣ Copy and paste by scan()

```
> whales=scan()  
1: 74 122 235 111 292 111 211 133 156 79  
11:  
Read 10 items
```

# Read formatted data

---

- Read data from formatted data files, e.g. a file of numbers from a single file, a table of numbers separated by space, comma, tab etc, with or without header

```
>whale=scan(file="whale.txt")
```

```
"whale.txt":
```

```
74 122 235 111 292 111 211 133 156 79
```

```
>whale=read.table(file="whale.txt", header=TRUE)
```

```
"whale.txt":
```

```
    texas florida
```

```
1  74      89
```

```
2 122    254
```

```
3 ....    ....
```

```
>read.table(file=file.choose()) # specify the file
```

```
>read.table(file="http://statweb.stanford.edu/~rag/stat141/exs/whale.txt"  
            ,header=T)    # read from internet
```



# Data frame

---

- ❑ A “data matrix” or a “data set”
  - it likes a matrix (rectangular grid)
  - But unlike matrix, different columns can be of different types
  - Row names have to be unique
- ❑ `>alphabet<-data.frame(index=1:26, symbol=LETTERS)`
- ❑ `read.table()` stores data in a *data frame*
- ❑ Access var in a dataset: `$`, `attach()`, `with()`
  - `>library(ISwR) #load the package that provides thuesen data`
  - `>data(thuesen)`
  - `>names(thuesen) #variable names`
  - `> blood.glucose # not visible`
  - `>length(thuesen$blood.glucose)`
  - `>with(thuesen, range(blood.glucose))`
  - `>attach(thuesen)`
  - `>range(blood.glucose)`
  - `>detach(thuesen)`

# More about data frame

---

- ❑ Indexing of data frames is the same as that of vector and matrix  
`>energy[energy$stature== "lean",]`
- ❑ Sorting rows by `order()`  
`>energy[order(energy$expend),]`  
`>energy[with(energy, order(stature, expend)),]`
- ❑ Selecting subsets of data by `subset()`  
`>subset(energy, stature=="lean" & expend>8)`
- ❑ Splitting data  
`>split(energy$expend, energy$stature)`

# Lists

---

- A larger composite object for combining a collection of objects

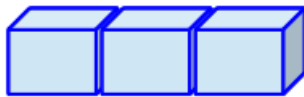
- Different from data frame, each object can be of different length, in addition to being of different types

```
>a=list(whales=c(74,122,235,111,292,111,211,  
133,16,79),  
simpsons=c("Homer", "Marge", "Bart", "Lisa",  
"Maggie"))
```

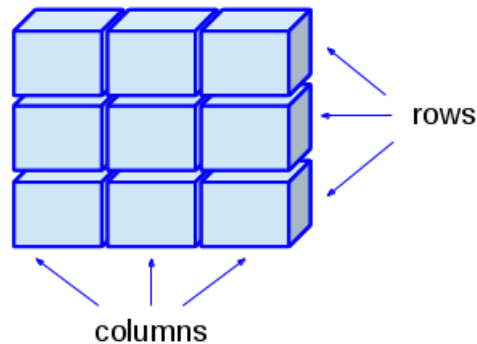
- Access by `$` or `[[ ]]`: `a$simpsons` or `a[[2]]`

# Summary of data structures

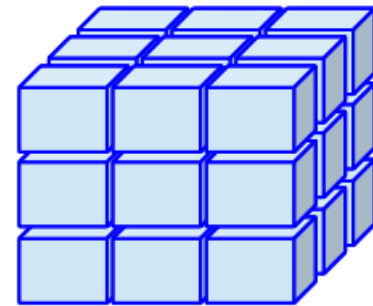
Vector



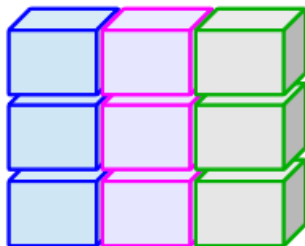
Matrix



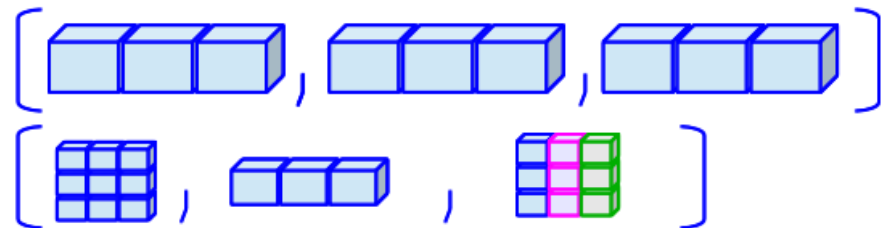
Array



Data Frame  
(Table)



Lists



# Manage the work environment

---

- ❑ What if there are more variables defined than can be remembered?
- ❑ `ls()` list all the objects(var, fun, etc) in a given environment
- ❑ `rm(a, b)`: delete variables a and b
  - `rm(list=ls())` will ?
- ❑ Get and set working directory
  - >`getwd()`
  - >`setwd("working/directory/path")`
- ❑ Save and load working environment
  - >`save.image(file="filename.RData")`
  - >`load(file="filename.RData")`

# scripting

---

- ❑ Edit your commands using your favorite text editors
- ❑ How to run

Inside R: `> source(filename)`

- Takes the input and runs them
- Do syntax-check before anything is executed
- Set `echo=T` to print executed commands

OR copy & paste

Outside R: `R CMD BATCH filename`

output is in `*.Rout`

Or: `Rscript filename`

# How to install packages

---

- To install CRAN packages, execute from the R console the following command:

```
> install.packages('UsingR')
```

OR download the package and install it directly

```
R CMD INSTALL aPackage_0.1.tar.gz
```

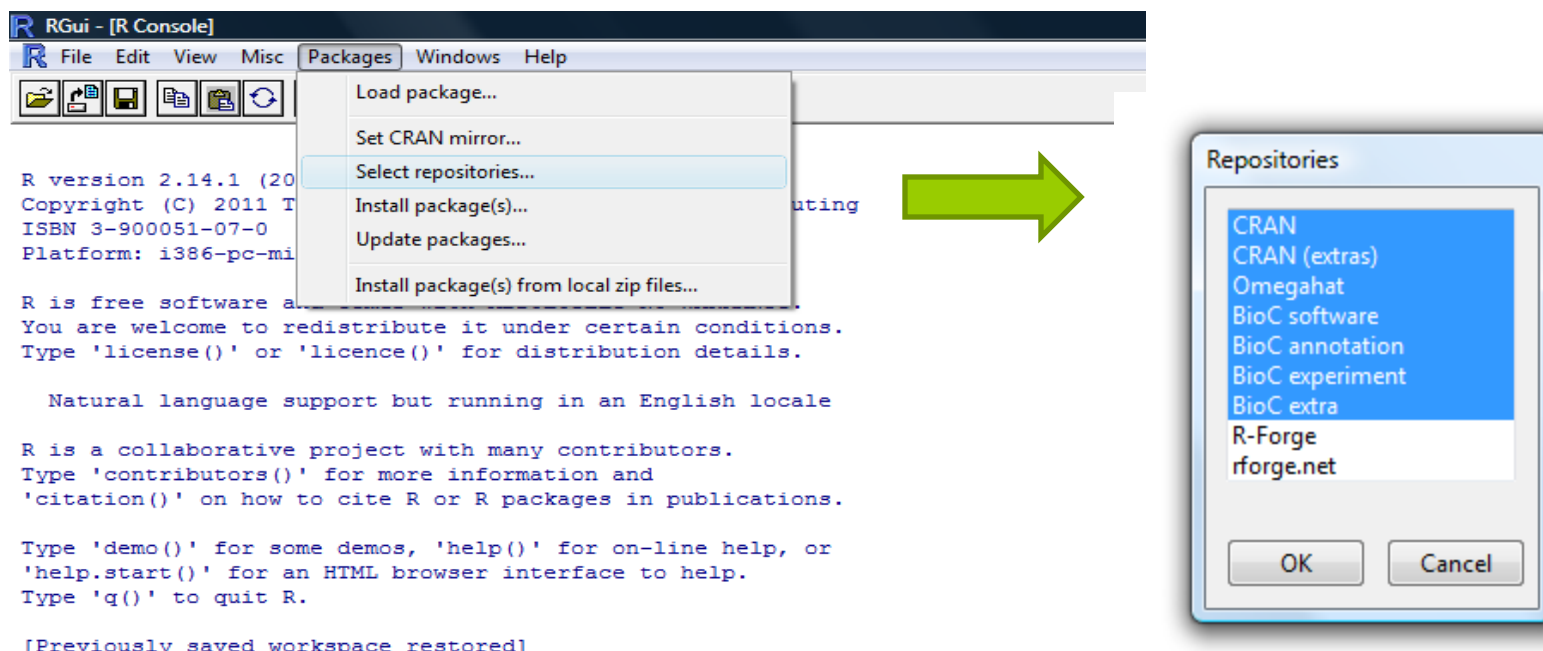
- Load a library

```
> library("UsingR")
```

or

```
> library(UsingR)
```

# Windows: Set repositories

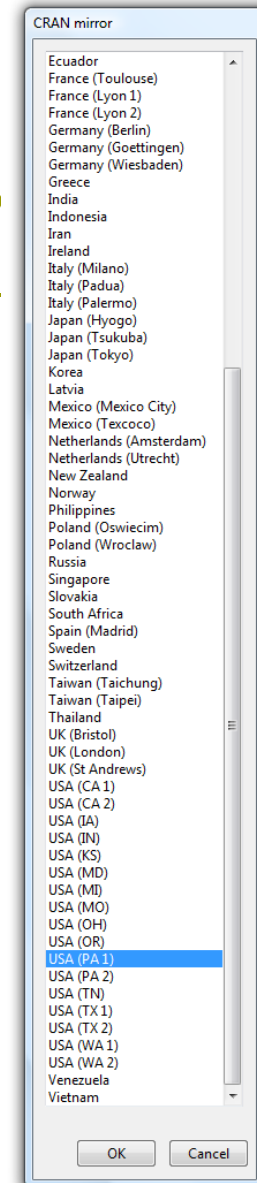
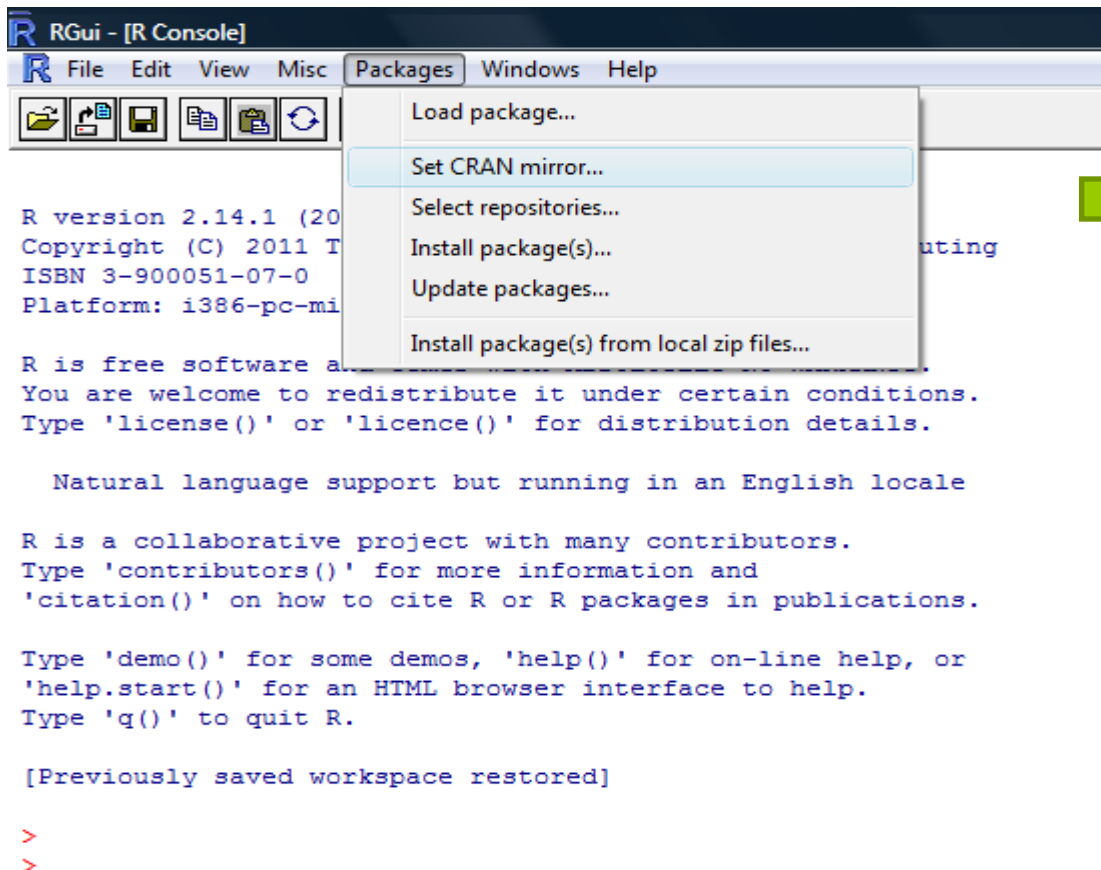


> |

- Make sure you include necessary repositories (you may simply select all of them)

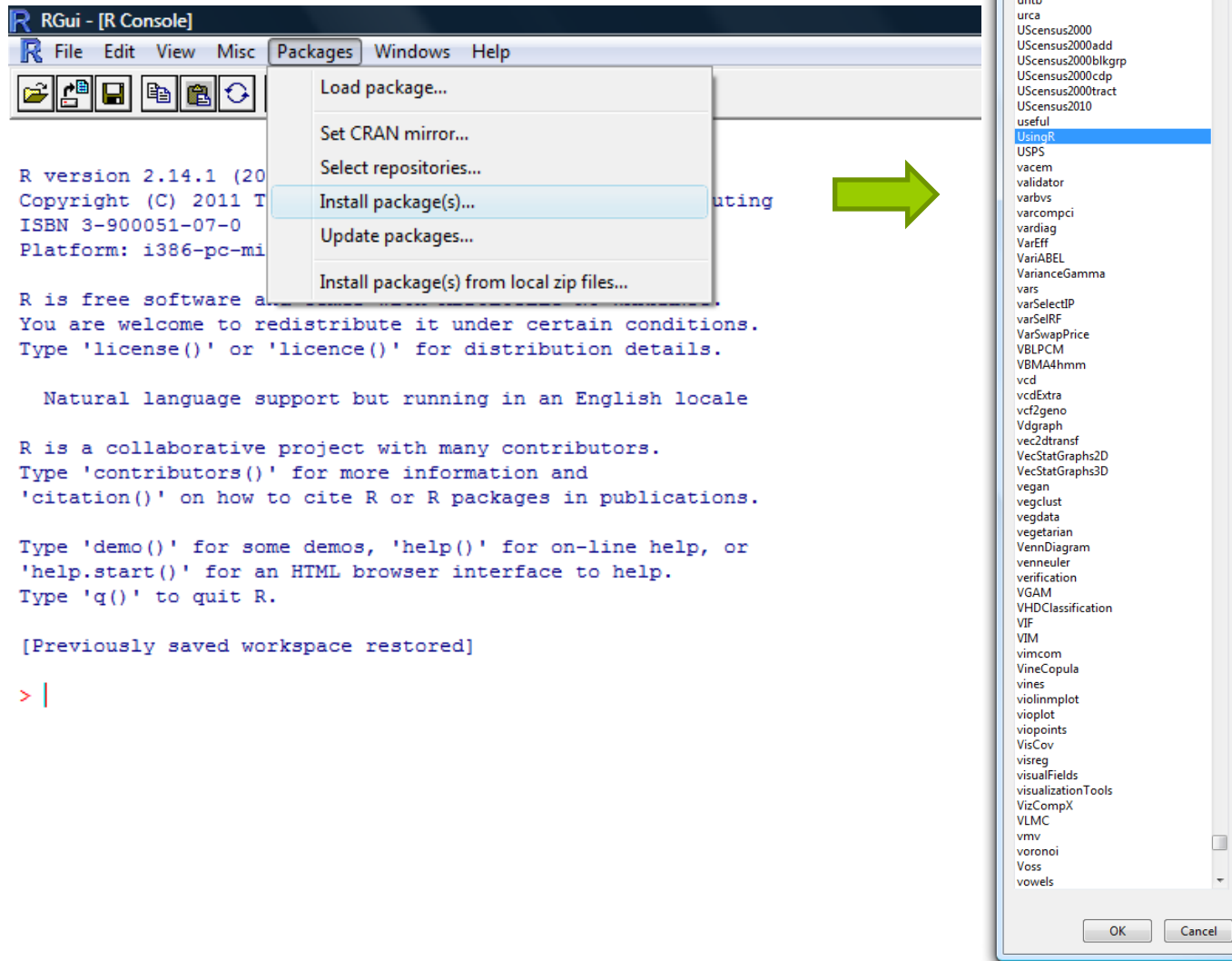


# Windows: Set CRAN mirror



- You can choose anyone but physically close ones are preferred

# Windows: install packages



The screenshot shows the RGui - [R Console] window. The 'Packages' menu is open, displaying options: 'Load package...', 'Set CRAN mirror...', 'Select repositories...', 'Install package(s)...', 'Update packages...', and 'Install package(s) from local zip files...'. The 'Install package(s)...' option is highlighted. A green arrow points from this menu item to a separate window titled 'Packages'. This window displays a list of installed packages, including 'unknownR', 'unmarked', 'unrb', 'urca', 'UScensus2000', 'UScensus2000add', 'UScensus2000blkgrp', 'UScensus2000cdp', 'UScensus2000tract', 'UScensus2010', 'useful', 'UsingR' (which is highlighted), 'USPS', 'vacem', 'validator', 'varbvs', 'varcompai', 'vardiag', 'VarEff', 'VariABEL', 'VarianceGamma', 'vars', 'varSelectIP', 'varSelfRF', 'VarSwapPrice', 'VBLPCM', 'VBMA4hmm', 'vcd', 'vcdExtra', 'vcf2geno', 'Vdgraph', 'vec2dtransf', 'VecStatGraphs2D', 'VecStatGraphs3D', 'vegan', 'vegclust', 'vegdata', 'vegetarian', 'VennDiagram', 'venneuler', 'verification', 'VGAM', 'VHDClassification', 'VIF', 'VIM', 'vimcom', 'VineCopula', 'vines', 'violinmplot', 'vioplot', 'viopoints', 'VisCov', 'visreg', 'visualFields', 'visualizationTools', 'VizCompX', 'VLMC', 'vmv', 'voronoi', 'Voss', and 'vowels'. At the bottom of the 'Packages' window are 'OK' and 'Cancel' buttons.

RGui - [R Console]

File Edit View Misc Packages Windows Help

Load package...  
Set CRAN mirror...  
Select repositories...  
Install package(s)...  
Update packages...  
Install package(s) from local zip files...

R version 2.14.1 (2011-05-22)  
Copyright (C) 2011 The R Foundation for Statistical Computing  
ISBN 3-900051-07-0  
Platform: i386-pc-mingw32

R is free software and is distributed under the terms of the GNU General Public License.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.

[Previously saved workspace restored]

> |

Packages

unknownR  
unmarked  
unrb  
urca  
UScensus2000  
UScensus2000add  
UScensus2000blkgrp  
UScensus2000cdp  
UScensus2000tract  
UScensus2010  
useful  
UsingR  
USPS  
vacem  
validator  
varbvs  
varcompai  
vardiag  
VarEff  
VariABEL  
VarianceGamma  
vars  
varSelectIP  
varSelfRF  
VarSwapPrice  
VBLPCM  
VBMA4hmm  
vcd  
vcdExtra  
vcf2geno  
Vdgraph  
vec2dtransf  
VecStatGraphs2D  
VecStatGraphs3D  
vegan  
vegclust  
vegdata  
vegetarian  
VennDiagram  
venneuler  
verification  
VGAM  
VHDClassification  
VIF  
VIM  
vimcom  
VineCopula  
vines  
violinmplot  
vioplot  
viopoints  
VisCov  
visreg  
visualFields  
visualizationTools  
VizCompX  
VLMC  
vmv  
voronoi  
Voss  
vowels

OK Cancel

# Additional references

---

- ❑ Beginners should print out the [R Reference Card](http://cran.r-project.org/doc/contrib/Short-refcard.pdf)  
<http://cran.r-project.org/doc/contrib/Short-refcard.pdf>
- ❑ [The R-FAQ](http://cran.r-project.org/doc/FAQ/R-FAQ.html) (Frequently Asked Questions on R)  
<http://cran.r-project.org/doc/FAQ/R-FAQ.html>
- ❑ A rather terse [introduction to R](http://cran.r-project.org/doc/manuals/R-intro.html) online  
<http://cran.r-project.org/doc/manuals/R-intro.html>
- ❑ [Bioconductor](http://www.bioconductor.org/)  
<http://www.bioconductor.org/>
- ❑ A useful online manual for R & Bioconductor  
[http://manuals.bioinformatics.ucr.edu/home/R\\_BioCondManual](http://manuals.bioinformatics.ucr.edu/home/R_BioCondManual)

# Computing Lab Ex.

---

## □ Lab 1

- Install UsingR and ISwR packages to your own directory
- Using **R** for Introductory Statistics, Page 18: 1.1-1.12
- 1.18, 1.19

# Acknowledgments

---

- ▣ Jeff Solka: for some of the slides adapted or modified from his lecture slides at George Mason University
- ▣ Brian Healy: for some of the slides adapted or modified from his lecture slides at Harvard University

# Next week

---

- Getting Data In and Out of R
  - Homework 1 due
  - First Quiz