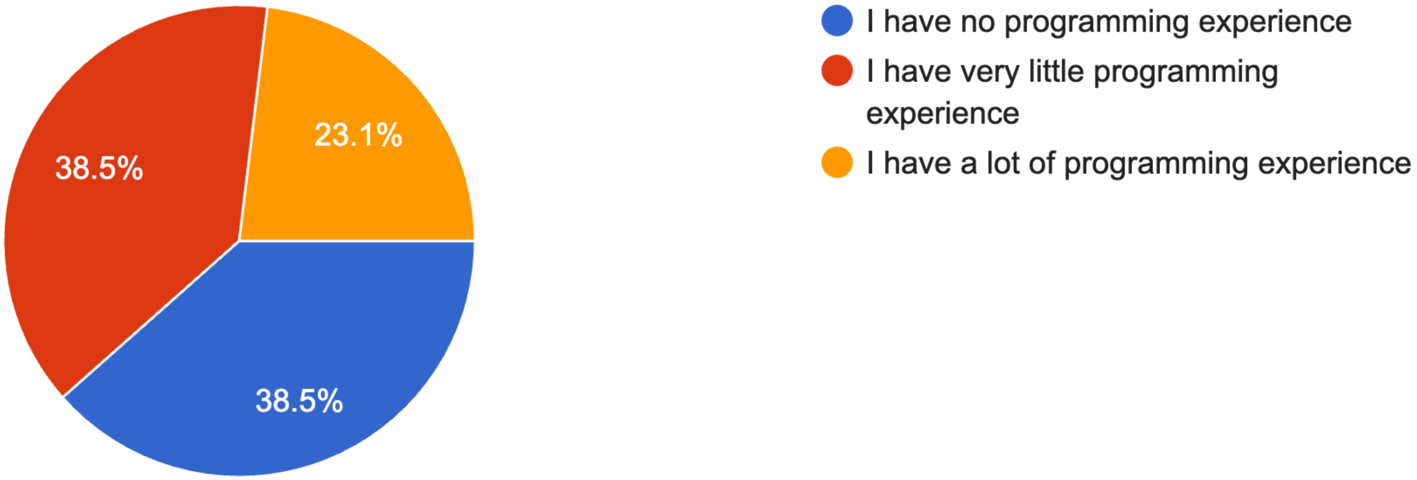
DATA 133 - Introduction to Data Science I

Instructor: Renzhi Cao
Computer Science Department
Pacific Lutheran University
Fall 2023

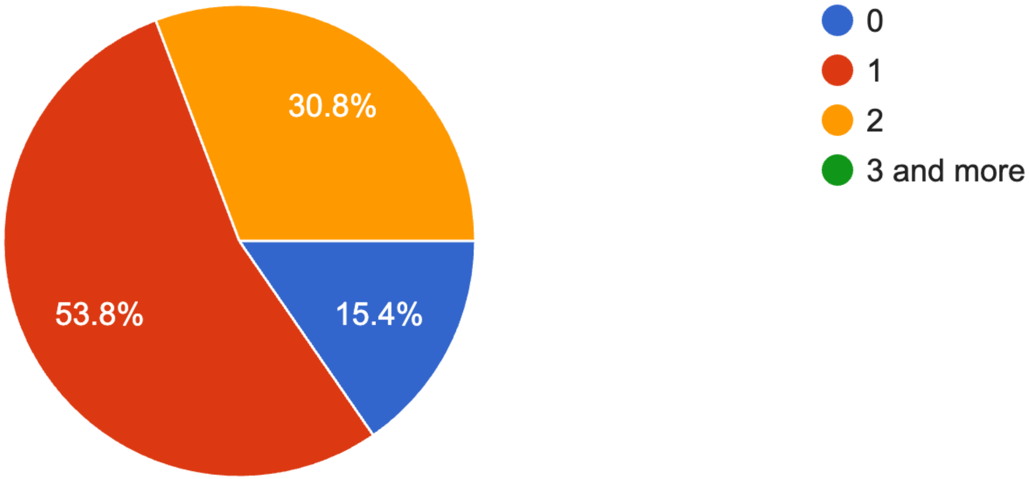
Please describe your programming experience (like any python or R or Java programming experience)

13 responses



How many time do you want to meet with Dr. Cao each week (NOT including meeting during class, but separate meetings in office hour or other type of one on one meeting)?

13 responses



Announcements

- Finish survey about your background (available on course website):
- https://docs.google.com/forms/d/e/1FAIpQLSe6dv_qgfTb08IvBk02RLUpP6Q5BSDF-seWc2qRji3HsjeWKA/viewform?vc=0&c=0&w=1&flr=0
- Request account:https://cs.plu.edu/hub/requests/new

user name: firstday
password: Fall23*Java

- Read books <<R Programming for Data Science>>: Page 1 - 12

Pictures from: https://www.google.com/search?

Reference book

- R Programming for Data Science. By Roger Peng. **ISBN-10:** 1365056821, April 20, 2016.

Pictures from: https://www.google.com/search?

Learning in today

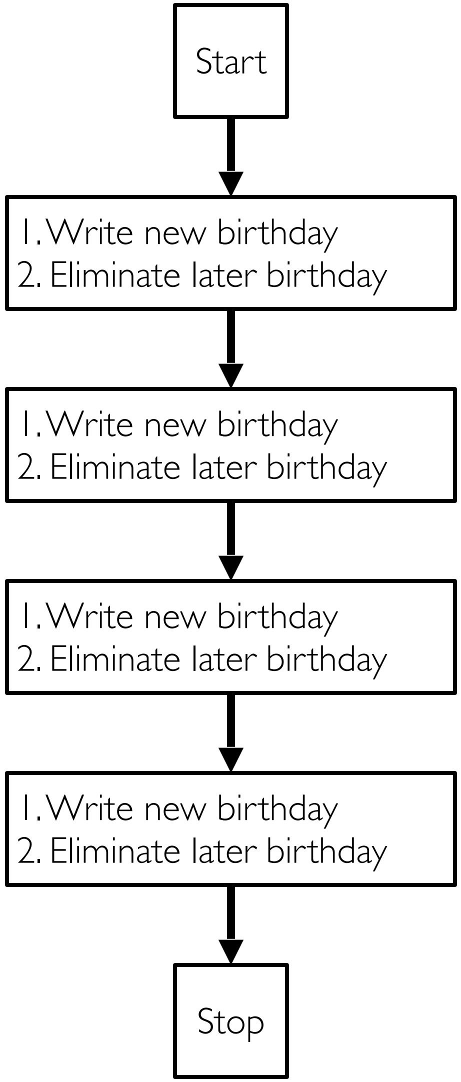
- Introduction to data science
- Understanding File Systems and department file server
- Basics of R environment

Pictures from: https://www.google.com/search?

Review - Problem solving

Finding the earliest birthday - method 1

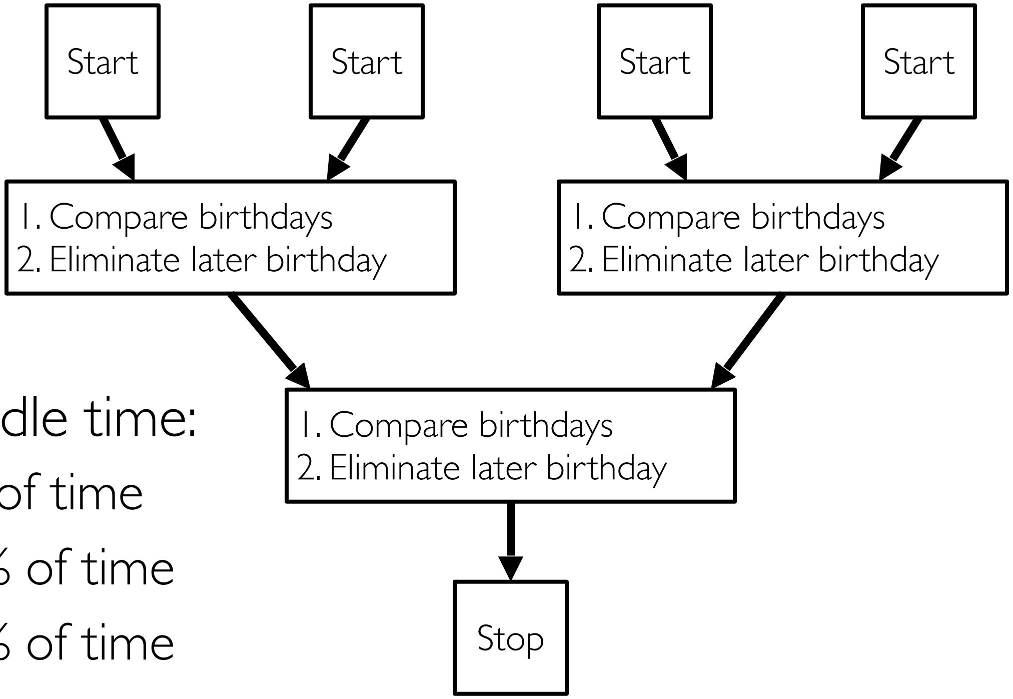
- Requires as many steps as people:
  - 4 people – 4 steps
  - 16 people – 16 steps
  - 32 people – 32 steps
- Each person spends most of their time sitting idle:
  - 4 people – Each person idle 75% of the time
  - 16 people – Each person idle 94% of the time
  - 32 people – Each person idle 97% of the time


Start
1. Write new birthday
2. Eliminate later birthday
1. Write new birthday
2. Eliminate later birthday
1. Write new birthday
2. Eliminate later birthday
1. Write new birthday
2. Eliminate later birthday
Stop

Review - Problem solving

Finding the earliest birthday - method 2

- Simultaneous events mean fewer steps:
  - 4 people – 2 steps
  - 16 people – 4 steps
  - 32 people – 5 steps



- Fewer steps mean less idle time:
  - 4 people – idle ≤ 50% of time
  - 16 people – idle ≤ 75% of time
  - 32 people – idle ≤ 80% of time

Conclusion #1: Computers can't see the "big picture" – only the immediate task at hand.
Conclusion #2: Not all programs are equal – some are faster or more flexible than others.

Review - Problem-Solving

A. Understand the Problem
- Do you understand all the words & terms that are being used?
- What are you being asked to find or show?
- Is there enough information to solve the problem?
- Can you draw a picture that might help?

B. Come Up With a Plan
- Guess and check, make a list, or draw a picture.
- Look for a pattern, or find a key equation.
- Try solving a simplified version of the problem.
- Work backwards.

C. Carry Out the Plan
- Be aware that you may run into roadblocks or dead-ends!
- Check to see if your results make sense.
- Don't be afraid to start over!

D. Make Your Solution Computer-Friendly
- Imagine you are writing to a student not in this class.
- Keep things brief… but make sure that you don't leave anything out.
- Write a step-by-step list of instructions… like writing a recipe.

Data science

What comes to mind when I say the word "DATA"?

Data presence in our daily life

- Websites track user’s clicks
- Smart phones are tracking your location, searches, patterns
- Smart watches
- Smart cars
- Amazon collects purchase habits
- Databases
- Government
- Sports

What can we do with all of this data?

Data presence in our daily life

What is Data Science?

Book defines a data scientist as: “Data scientist is someone who knows more statistics than a computer scientist and more computer science than a statistician”

Better definition for data scientist: individual that extracts insights from unorganized data.

Facebook: https://www.facebook.com/notes/facebook-data-science/nfl-fans-on-facebook/10151298370823859

Target: http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?_r=0

Government: http://www.marketplace.org/2014/08/22/tech/beyond-ad-clicks-using-big-data-social-good

First problem with data

Assume a list of users:

ID	Name
1	Hero
2	Dunn
3	Sue
4	Chi
5	Thor
6	Clive
7	Hicks
8	Devin
9	Kate
10	Klein

First problem with data

Assume a list of users:

ID	Name
1	Hero
2	Dunn
3	Sue
4	Chi
5	Thor
6	Clive
7	Hicks
8	Devin
9	Kate
10	Klein

We know something about their friendships

Friendships
Hero-Dunn
Hero-Sue
Dunn-Sue
Dunn-Chi
Sue- Chi
Chi – Thor
Thor – Clive
Clive – Hicks
Clive – Devin
Hicks – Kate
Devin – Klein
Kate - Klein

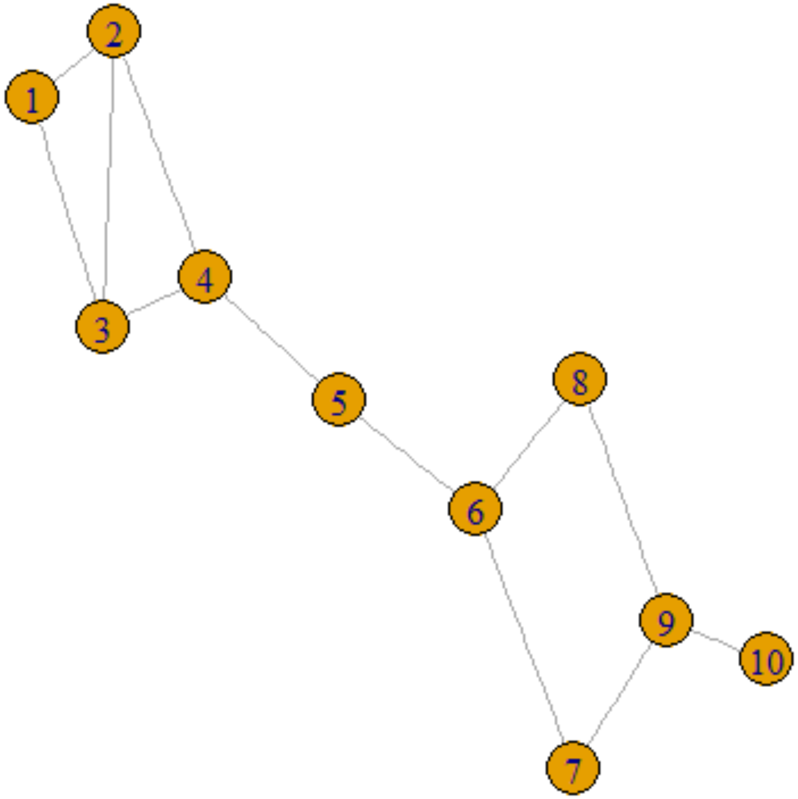
First problem with data

Assume a list of users:

ID	Name
1	Hero
2	Dunn
3	Sue
4	Chi
5	Thor
6	Clive
7	Hicks
8	Devin
9	Kate
10	Klein

Easier to read:

Friendships
1 - 2
1 - 3
2 - 3
2 - 4
3 - 4
4 - 5
5 - 6
6 - 7
6 - 8
7 - 9
8 - 9
9 - 10


1
2
3
4
5
6
7
8
9
10

Data presence in our daily life

Let’s analyze our graph

- What can we learn by looking at it?
  - What is the average number of friends per person?
  - Who is the most popular person?
  - Who is the most important person in the network?

Data presence in our daily life

A little taste of R

We will cover R in the future in much more detail, but this is a taste of the things you can do.

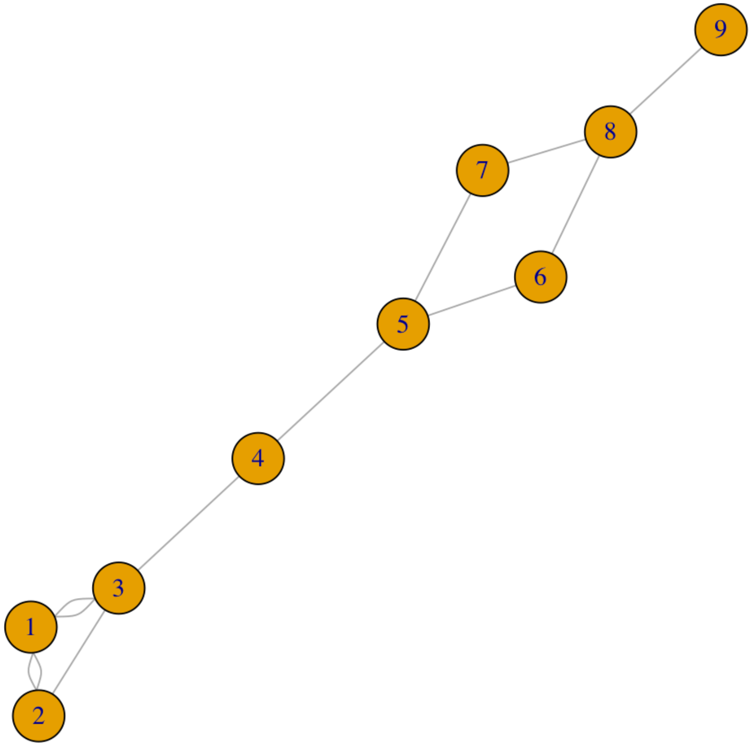
Open R “as administrator”

```
> install.packages("igraph")
> library(igraph)
> graph.non <- graph(c(1,2, 1,3, 1,2, 1,3, 2,3, 3,4, 4,5, 5,6,
5,7, 6,8, 7,8, 8,9),directed=FALSE)
➢ plot(graph.non)
➢ tkplot(graph.non,layout=layout.kamada.kawai)
➢
```

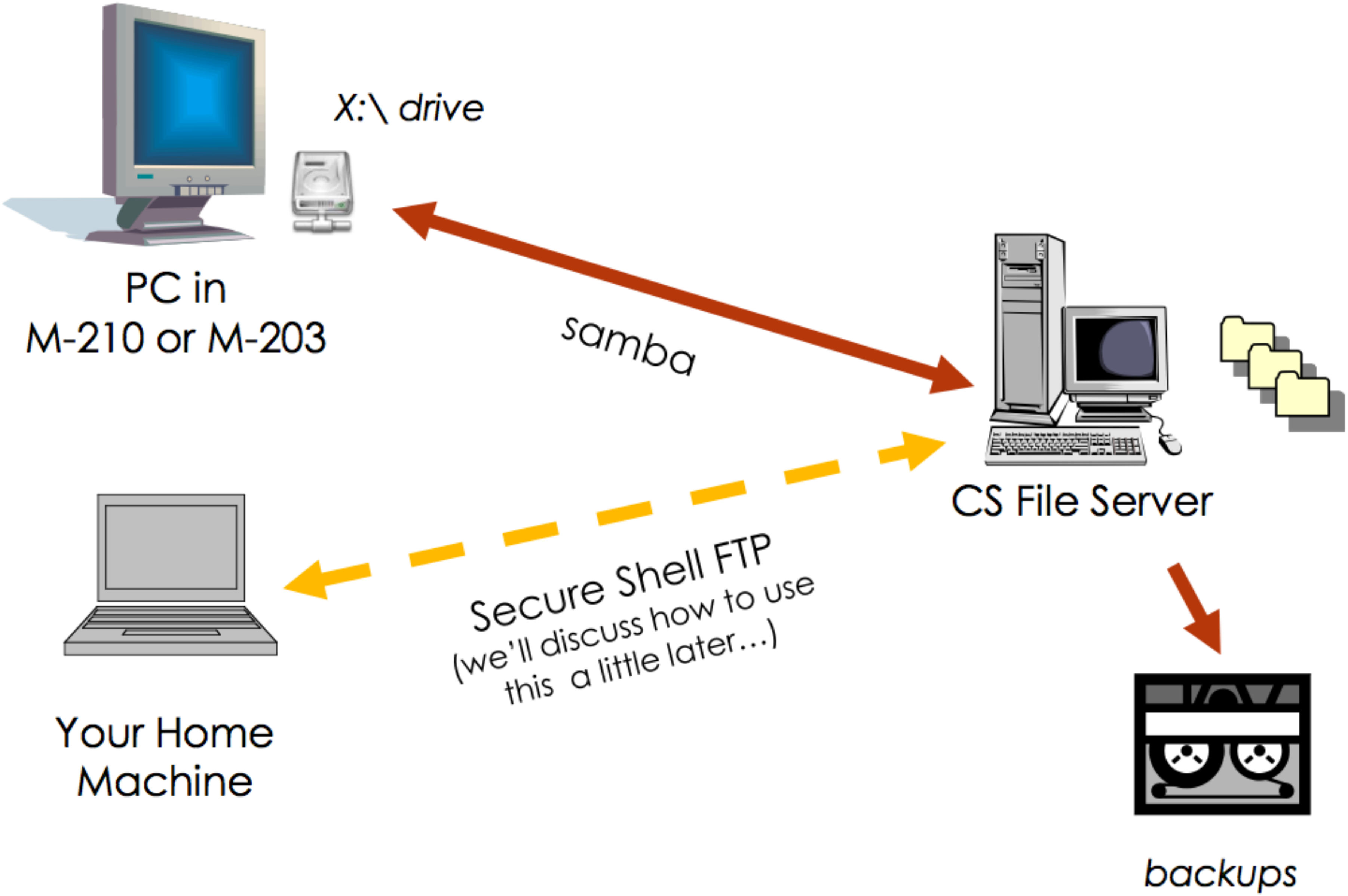
Disclaimer: Don’t worry if this looks too complex. It will all make sense at the end of the semester!

Data presence in our daily life

A little taste of R



Navigating Drives & Directories…



X:\ drive

PC in
M-210 or M-203

samba

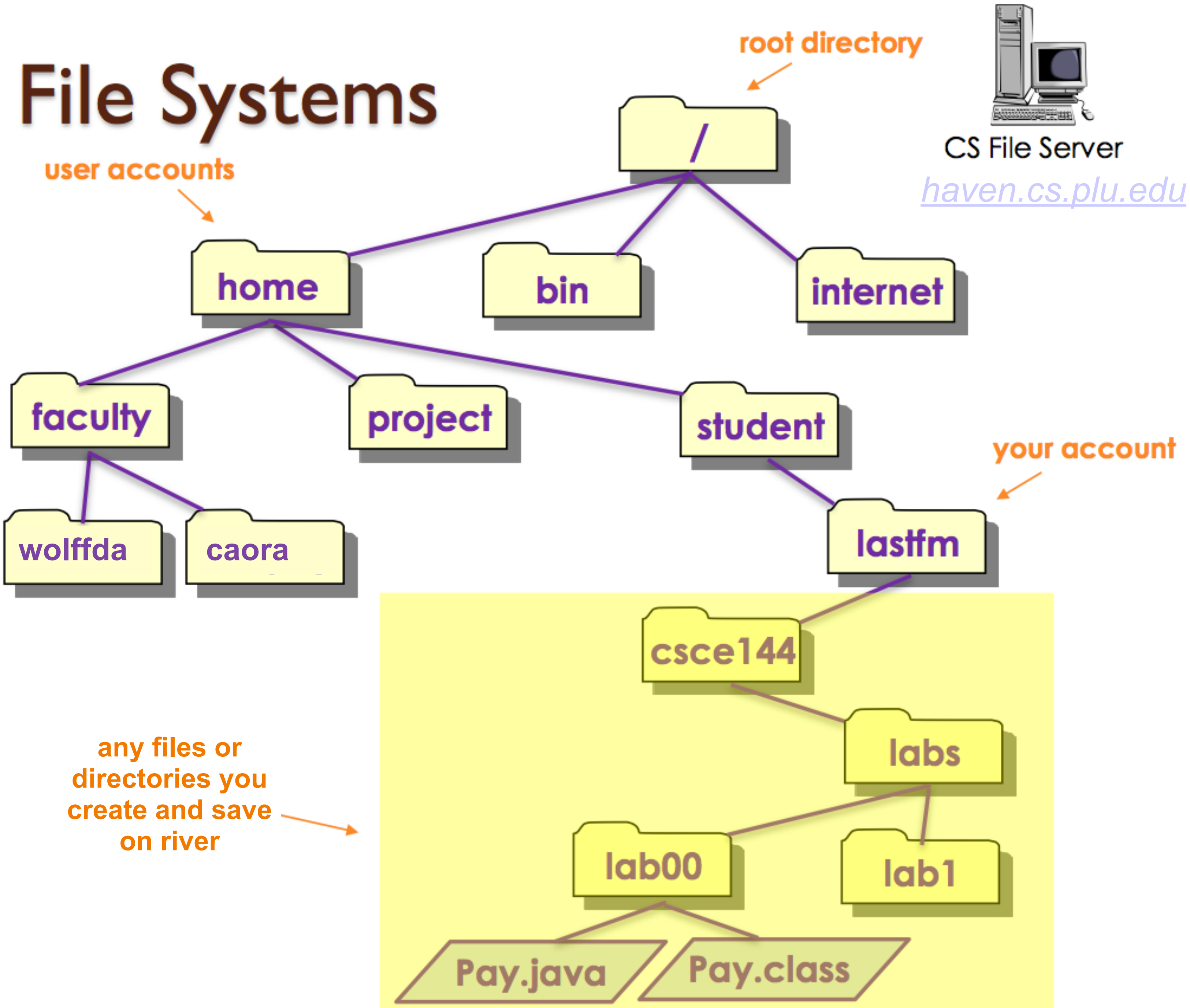
CS File Server

Your Home
Machine

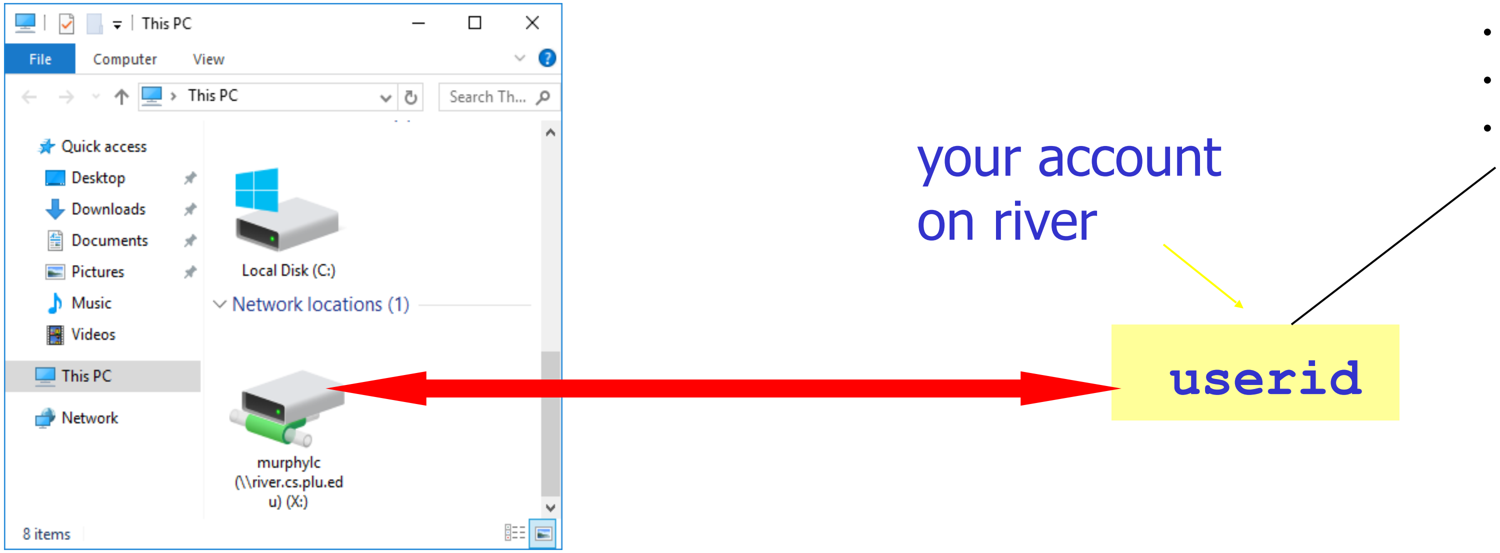
Secure Shell FTP
(we'll discuss how to use
this a little later…)

backups

File Systems



When you logon to the CSCI lab machines in Morken 203 or 210 using your epass and password the PC's "X" drive is automatically mapped to your river account



your account
on river

userid

Data presence in our daily life

Demo on creating folder DATA133 in X drive.

Current working directory
Path

Some commonly used command in Ubuntu (from chatGPT)

Break

Background of R

- What is R?

  - *A dialect of S, S is a language that was developed by John Chambers and others at the old Bell Telephone Laboratories, originally part of AT&T Corp*

Pictures from: https://www.google.com/search?

Background of R

- What is basic features of R?

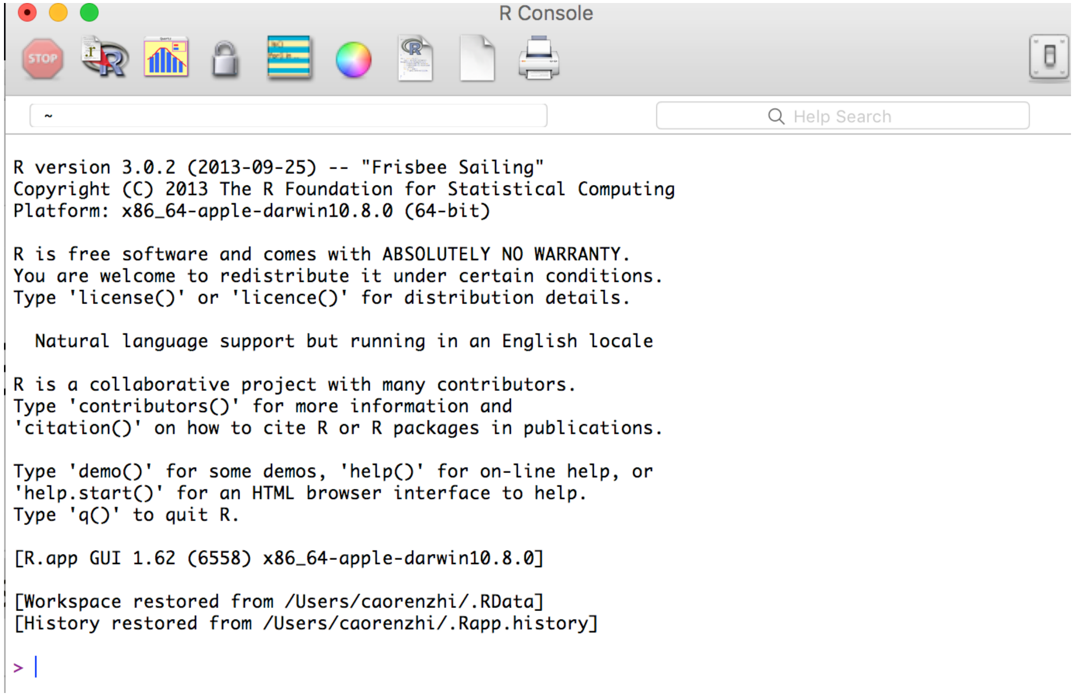
- *R is free and open source.*
- *R runs on most standard operating system*
- *R has frequent releases.*
- *R has sophisticated graphic capabilities.*
- *R is both useful for interactive work and powerful programming language for developing new tools.*

Pictures from: https://www.google.com/search?

Installation of R

https://cran.r-project.org

- *1. Windows*
- *2. Mac*

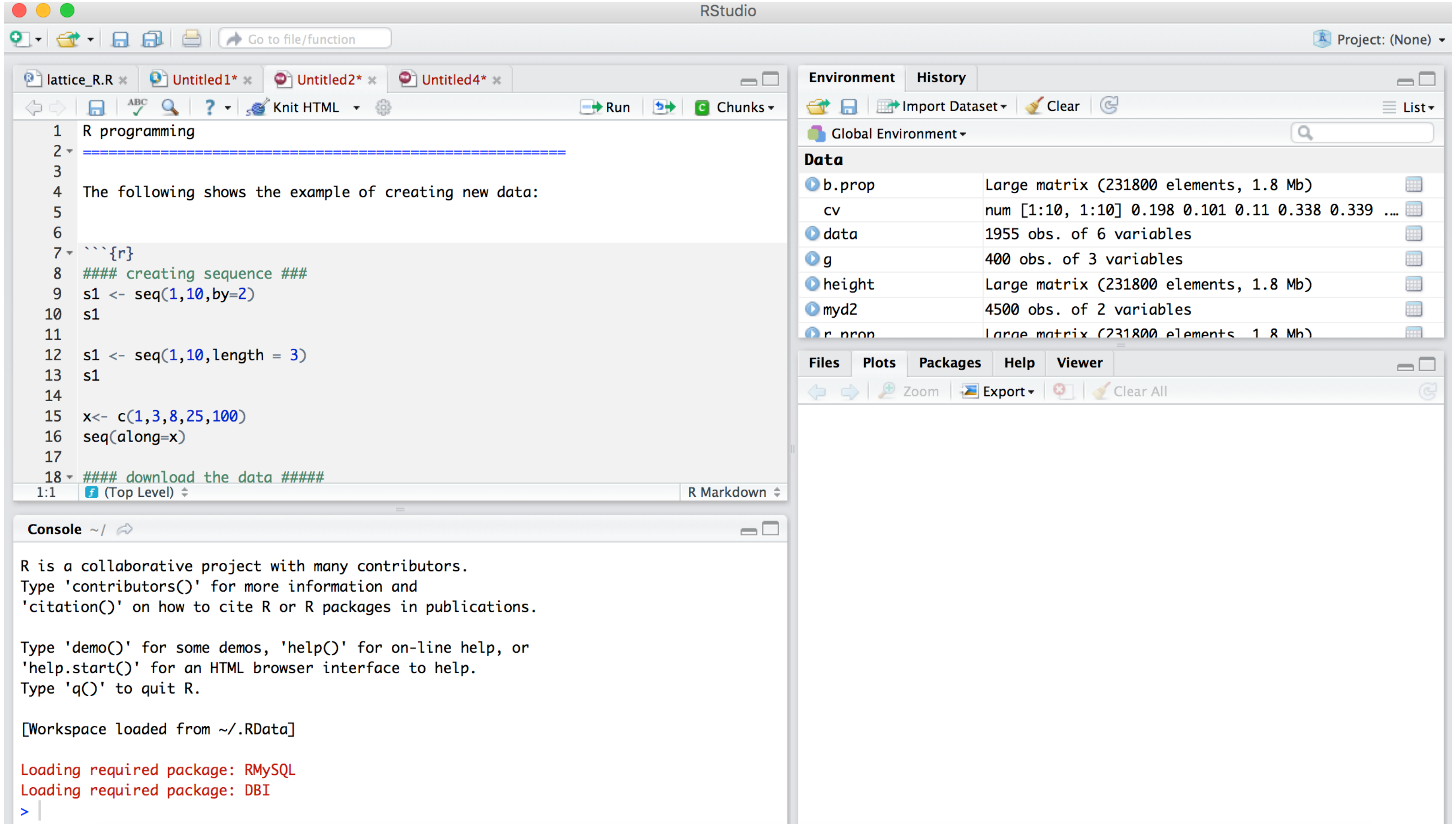


```
R version 3.0.2 (2013-09-25) -- "Frisbee Sailing"
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin10.8.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
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.

[R.app GUI 1.62 (6558) x86_64-apple-darwin10.8.0]

[Workspace restored from /Users/caorenzhi/.RData]
[History restored from /Users/caorenzhi/.Rapp.history]

> |
```

Installation of R

Useful IDE for R only: Rstudio

https://www.rstudio.com



~~~
R programming
=========================================================

The following shows the example of creating new data:


```{r}
#### creating sequence ###
s1 <- seq(1,10,by=2)
s1

s1 <- seq(1,10,length = 3)
s1

x<- c(1,3,8,25,100)
seq(along=x)

#### download the data #####
~~~

Data

b.prop	Large matrix (231800 elements, 1.8 Mb)
cv	num [1:10, 1:10] 0.198 0.101 0.11 0.338 0.339 ...
data	1955 obs. of 6 variables
g	400 obs. of 3 variables
height	Large matrix (231800 elements, 1.8 Mb)
myd2	4500 obs. of 2 variables
r.prop	Large matrix (231800 elements, 1.8 Mb)

```
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.

[Workspace loaded from ~/.RData]

Loading required package: RMySQL
Loading required package: DBI
>
```

Demo of R

- *Setting work directory and edit R code*
- *Demo: ls(), dir(), getwd(), setwd()*
- *Edit R code*

R console input

- *<- as assignment operator*
- *# indicate comment*

Demo of R

- *When a complete expression is entered, it is evaluated and the result of the evaluated expression is returned.*

- *> x<- 100   # nothing printed*
- *> x   # auto-printing*
- *>print(x).   # explicit printing*

- *> x <- 1*
- *> print(x)*
- *> x*

Demo of R

- *Use source() to run R script*

Demo of R

What is the output?

- *(1). i <- 100*

- *(2). i <- 100   # assign 100 to i*
- *i*

- *(3). i <- 10 / 2*
- *i*

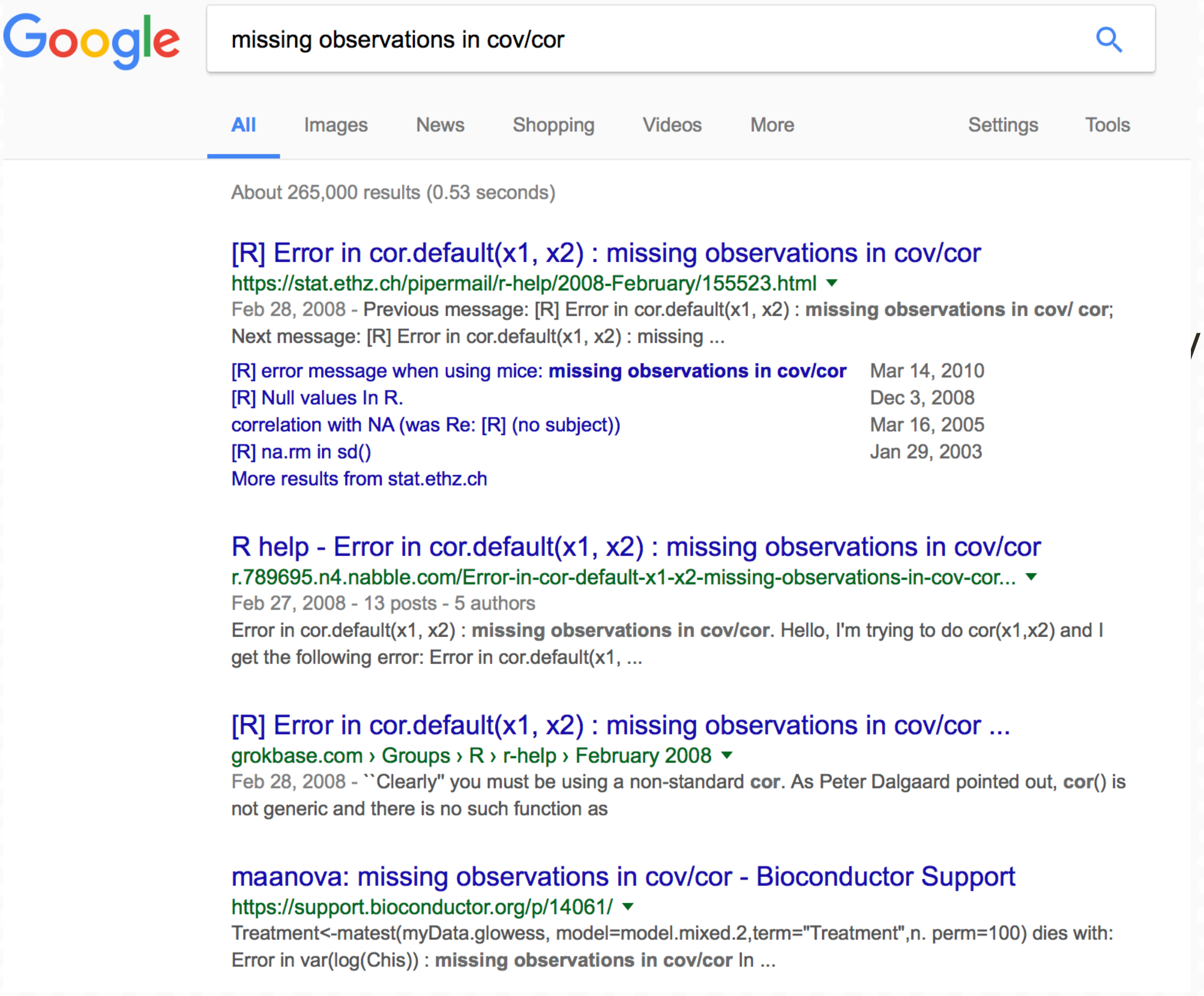
- *(4). 10/2*

- *(5). a <- 100*
- *print(a)*

Getting help

- *Try to find answer by contacting Dr. Cao*
- *Try to find answer by searching the web*
- *Try to find answer by reading the manual*
- *Try to find answer by reading the FAQ*
- *Try to find answer by inspection or experimentation*
- *Try to find answer by asking a skilled friend*
- *Try to find answer by reading the source code*

Getting help



About 265,000 results (0.53 seconds)

[R] Error in cor.default(x1, x2) : missing observations in cov/cor
https://stat.ethz.ch/pipermail/r-help/2008-February/155523.html ▾
Feb 28, 2008 - Previous message: [R] Error in cor.default(x1, x2) : **missing observations in cov/ cor**; Next message: [R] Error in cor.default(x1, x2) : missing ...

[R] error message when using mice: **missing observations in cov/cor**	Mar 14, 2010
[R] Null values In R.	Dec 3, 2008
correlation with NA (was Re: [R] (no subject))	Mar 16, 2005
[R] na.rm in sd()	Jan 29, 2003

More results from stat.ethz.ch

R help - Error in cor.default(x1, x2) : missing observations in cov/cor
r.789695.n4.nabble.com/Error-in-cor-default-x1-x2-missing-observations-in-cov-cor... ▾
Feb 27, 2008 - 13 posts - 5 authors
Error in cor.default(x1, x2) : **missing observations in cov/cor**. Hello, I'm trying to do cor(x1,x2) and I get the following error: Error in cor.default(x1, ...

[R] Error in cor.default(x1, x2) : missing observations in cov/cor ...
grokbase.com › Groups › R › r-help › February 2008 ▾
Feb 28, 2008 - ``Clearly'' you must be using a non-standard **cor**. As Peter Dalgaard pointed out, **cor**() is not generic and there is no such function as

maanova: missing observations in cov/cor - Bioconductor Support
https://support.bioconductor.org/p/14061/ ▾
Treatment<-matest(myData.glowess, model=model.mixed.2,term="Treatment",n. perm=100) dies with: Error in var(log(Chis)) : **missing observations in cov/cor** In ...

Getting help

- *What steps will reproduce the problem?*
- *What is the expected output?*
- *What do you see instead?*
- *What version of R do you use?*
- *What operating system?*
- *What other information?*

Getting help

- *Stupid: "Help! Cann't fit linear model"*
- *Smart: "R 3.0.2 lm() function produces seg fault with large data frame, MAC OS X 10.9.1"*
- *Smarter: "R 3.0.2 lm() function on MAC OS X 10.9.1 — seg fault on large data frame"*

Getting help in R

- > *?print*
- > *help(print)*

Pair-programming

Try the following R script from command mode:

```
> install.packages("igraph")
> library(igraph)
> graph.non <- graph(c(1,2, 1,3, 1,2, 1,3, 2,3, 3,4, 4,5, 5,6,
5,7, 6,8, 7,8, 8,9),directed=FALSE)
➢ plot(graph.non)
➢ tkplot(graph.non,layout=layout.kamada.kawai)
```

Pair-programming

➢ Here is the R code to calculate the sum of the first 20 integers

```
20*(20+1)/2
```

➢ However, we can define a variable to use the formula for other values of n n <- 20 n*(n+1)/2

```
n <- 20
n*(n+1)/2
```

Pair-programming

First work on command mode and then save the script and run it using source command. Print it out and turn it in as in-class exercise

- Now, write code to calculate the sum of the first 100 integers
- How about the sum of the first 10000 integers?

For next time

- Read book Page 12-22
- Quiz 0 on the Syllabus
- Quiz 1 on reading and lecture

Pictures from: https://www.google.com/search?