Statistics 202: Statistical Aspects of Data Mining

Professor Rajan Patel

Lecture 1 = Course web page and Chapters 1+2

Agenda:

1) Go over information on course web page
2) Lecture over Chapter 1
3) Discuss necessary software
4) Start lecturing over Chapter 2 (Data)

Statistics 202: Statistical Aspects of Data Mining

Professor Rajan Patel

Course web page:

http://sites.google.com/site/stats202
(linked from stats202.com)

Course e-mail address:
stats202@gmail.com

Google group for general discussion:
stats202

Statistics 202: Statistical Aspects of Data Mining

Professor Rajan Patel

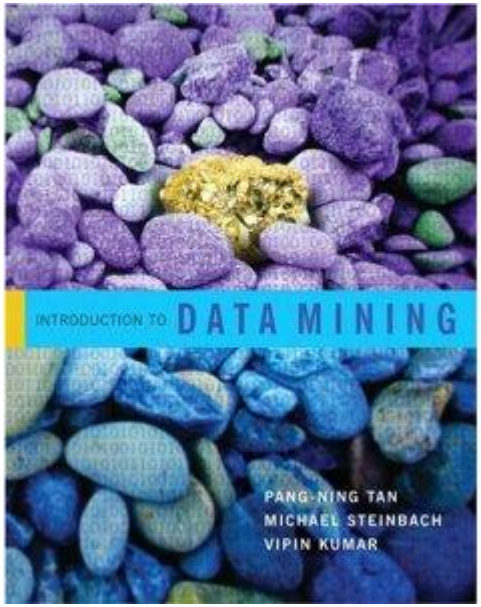
Who are you?

131 students enrolled
- 8 biomedical informatics grad students
- 50 visiting students (from around the world)
- 10 high school students
- 20 undergrad visiting students
- and students from

medical school, computer science, electrical engineering, economics, physics, psychology, statistics, mechanical engineering, materials science, immunology, geophysics, genetics, education, developmental biology, chemistry, bioengineering, and more.

Introduction to Data Mining

by
Tan, Steinbach, Kumar

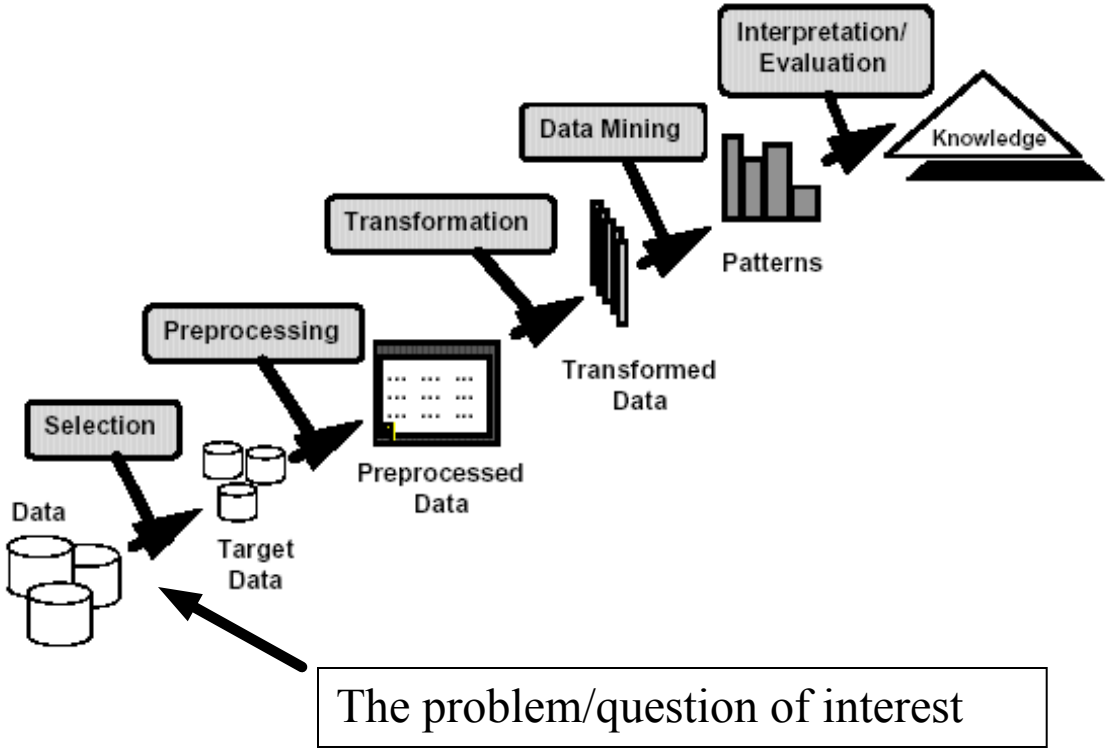


Chapter 1: Introduction

What is Data Mining?

Data mining is the process of automatically discovering useful information in large data repositories. (page 2)

There are many other definitions



Data Mining Examples and Non-Examples

Data Mining:

-Certain names are more prevalent in certain US locations (O'Brien, O' Rurke, O'Reilly… in Boston area)

-Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com, etc.)

NOT Data Mining:

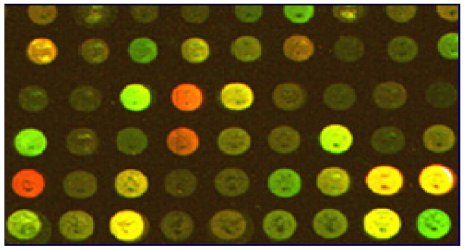
-Look up phone number in phone directory

-Query a Web search engine for information about "Amazon"

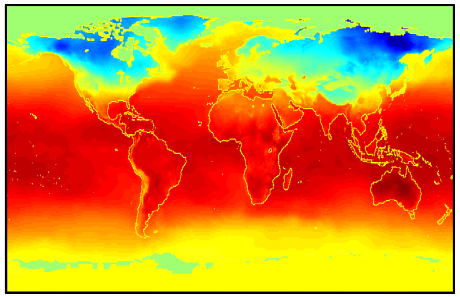
Why Mine Data? Scientific Viewpoint

Data collected and stored at enormous speeds (GB/hour)

- remote sensors on a satellite
- telescopes scanning the skies
- microarrays generating gene expression data
- scientific simulations generating terabytes of data



Traditional techniques infeasible for large data sets



Data mining may help scientists

- in classifying and segmenting data
- in hypothesis formation

Why Mine Data? Commercial Viewpoint

Lots of data is being collected and warehoused

- **Web data, e-commerce**
- **Purchases at department / grocery stores**
- **Bank/credit card transactions**
- **Computers have become more powerful**
- **Competitive pressure is strong**
- **Provide better, customized services for an edge**

In class exercise #1:

Give an example of something you did yesterday or today which resulted in data which could potentially be mined to discover useful information.

Origins of Data Mining (page 6)

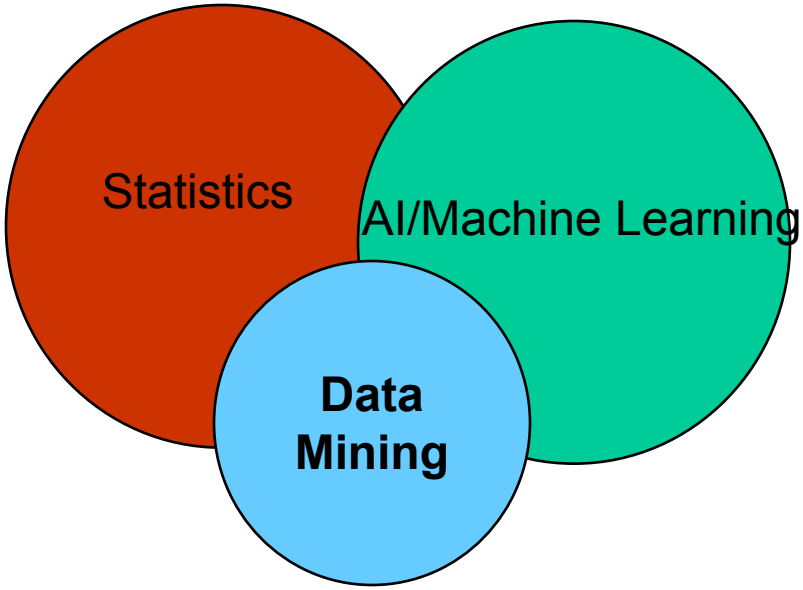
Draws ideas from machine learning, AI, pattern recognition and statistics

Traditional techniques may be unsuitable due to

- enormity of data

- high dimensionality of data

- heterogeneous, distributed nature of data



2 Types of Data Mining Tasks (page 7)

Predictive Methods:

Use some variables to predict unknown or future values of other variables.

Descriptive Methods:

Find human-interpretable patterns that describe the data.

Examples of Data Mining Tasks

- Classification [Predictive] (Chapters 4,5)
- Regression [Predictive] (covered in stats classes)

- Visualization [Descriptive] (in Chapter 3)
- Association Analysis [Descriptive] (Chapter 6)
- Clustering [Descriptive] (Chapter 8)
- Anomaly Detection [Descriptive] (Chapter 10)

Software We Will Use:

R

Can be downloaded from

http://cran.r-project.org/ for Windows, Mac or Linux

Downloading R for Windows:

The Comprehensive R Archive Network - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites

Address http://cran.r-project.org/

The Comprehensive

Frequently used pages

CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals

Download and Install R

Precompiled binary distributions of the base system and contributed pack
versions of R:

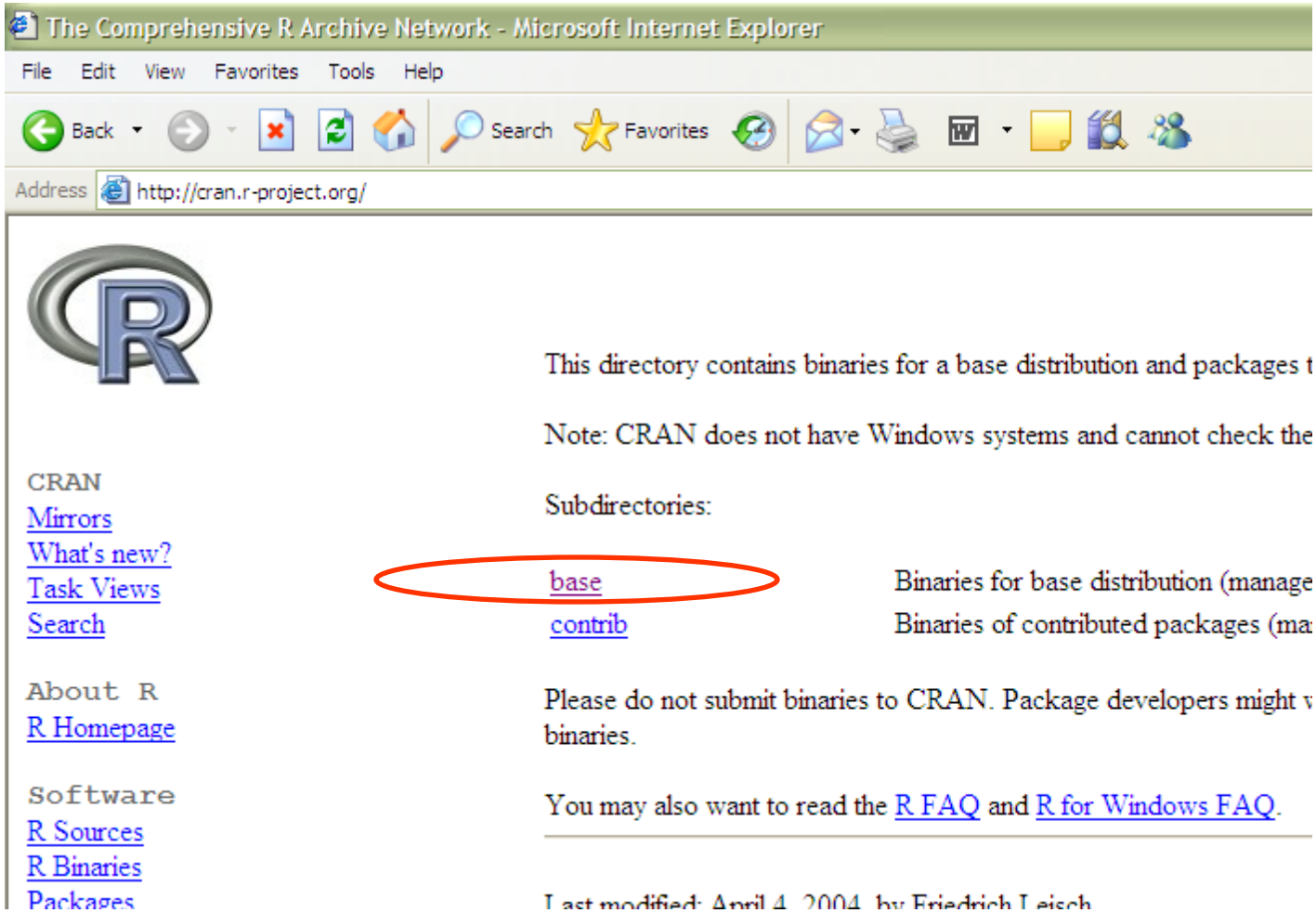
- Linux
- MacOS X
- Windows (95 and later)

Source Code for all Platforms

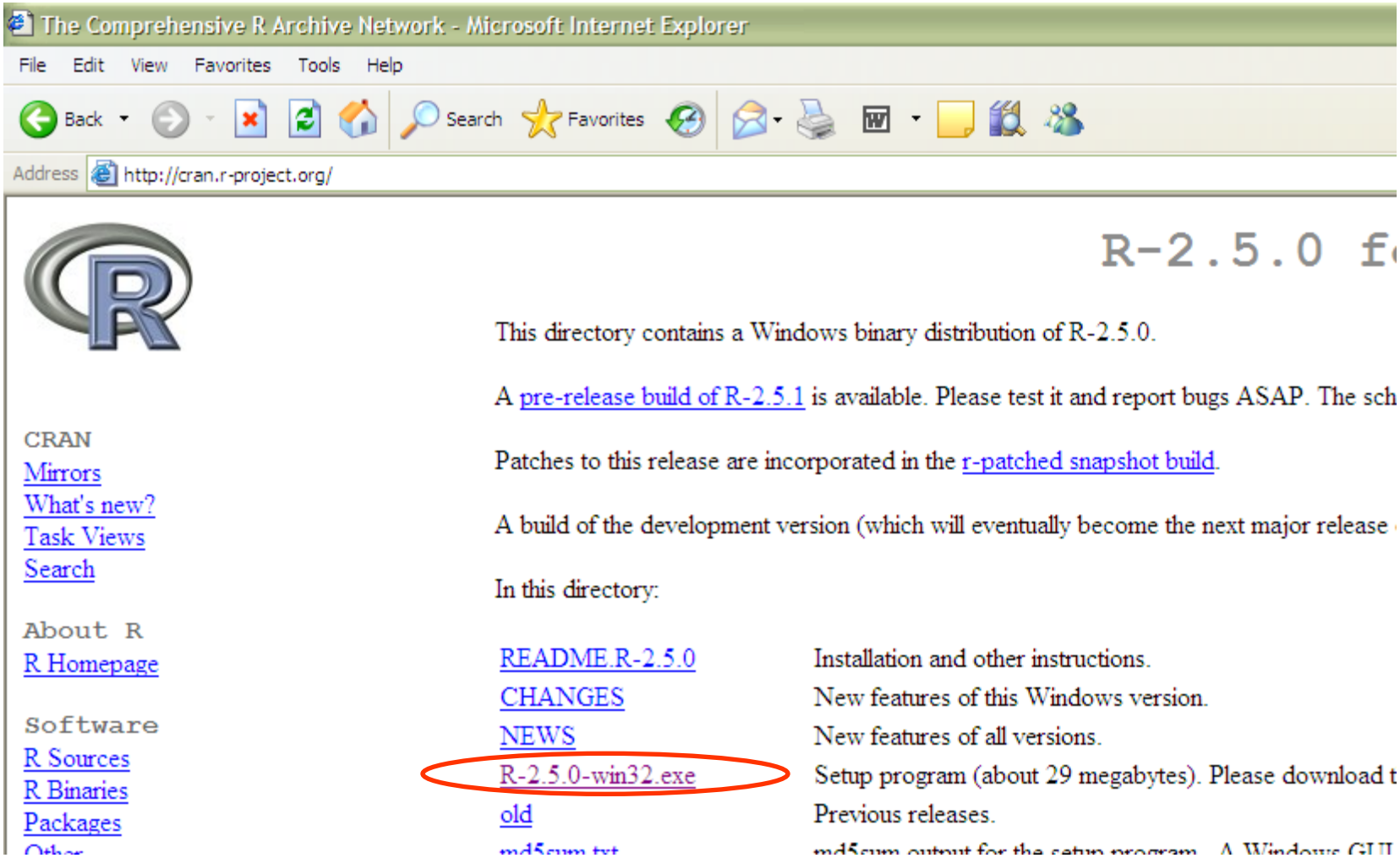
Windows and Mac users most likely want the precompiled binaries listed
compiled before you can use them. If you do not know what this means,

- **The latest release** (2007-04-24): R-2.5.0.tar.gz (read what's ne
- Sources of R alpha and beta releases (daily snapshots, created onl

Downloading R for Windows:

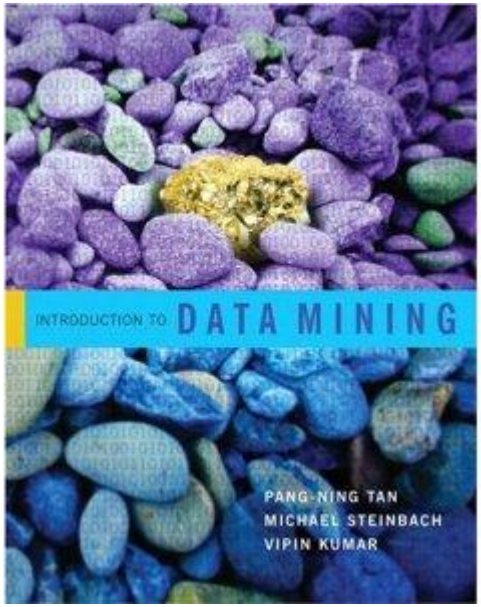


Downloading R for Windows:



Introduction to Data Mining

by
Tan, Steinbach, Kumar



Chapter 2: Data

What is Data?

Attributes

An attribute is a property or characteristic of an object

Examples: eye color of a person, temperature, etc.

Objects

An Attribute is also known as variable, field, characteristic, or feature

A collection of attributes describe an object

An object is also known as record, point, case, sample, entity, instance, or observation

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Reading Data into R

Download it from the web at

http://sites.google.com/site/stats202/data/weblog2.txt

What is your working directory?

```
> getwd()
```

Change it to your deskop:

```
> setwd("/Users/rajan/Desktop")
```

Read it in:

```
> data<-read.csv("weblog2.txt", sep=" ",header=F)
```

Reading Data into R

Look at the first 5 rows:

```
>data[1:5,]
```

```
            V1 V2 V3                 V4     V5                          V6  V7   V8
1 122.178.203.210  -  - [20/Jun/2011:00:00:25 -0400]        GET /favicon.ico HTTP/1.1 404 2294
2  70.105.172.121  -  - [20/Jun/2011:00:01:03 -0400]                GET / HTTP/1.1 200  736
3  70.105.172.121  -  - [20/Jun/2011:00:01:03 -0400]        GET /favicon.ico HTTP/1.1 404 2290
4  70.105.172.121  -  - [20/Jun/2011:00:01:03 -0400]        GET /favicon.ico HTTP/1.1 404 2290
5  70.105.172.121  -  - [20/Jun/2011:00:01:32 -0400] GET /original_index.html HTTP/1.1 200 3897
             V9                                V10
1 www.stats202.com http://www.stats202.com/original_index.html
2     stats202.com                                -
3     stats202.com                                -
4     stats202.com                                -
5 www.stats202.com                    http://stats202.com/
                                                 V11 V12
1  Opera/9.80 (X11; Linux x86_64; U; en) Presto/2.8.131 Version/11.11   -
2 Mozilla/5.0 (Windows NT 5.1; rv:2.0.1) Gecko/20100101 Firefox/4.0.1   -
3 Mozilla/5.0 (Windows NT 5.1; rv:2.0.1) Gecko/20100101 Firefox/4.0.1   -
4 Mozilla/5.0 (Windows NT 5.1; rv:2.0.1) Gecko/20100101 Firefox/4.0.1   -
5 Mozilla/5.0 (Windows NT 5.1; rv:2.0.1) Gecko/20100101 Firefox/4.0.1   -
```

Look at the first column:

```
> data[,1]
```