

Statistics 202: Statistical Aspects of Data Mining

Professor David Mease



Tuesday, Thursday 9:00-10:15 AM Terman 156

Lecture 1 = Course web page and chapter 1

Agenda:

- 1) Go over information on course web page**
- 2) Lecture over chapter 1**
- 3) Discuss necessary software**
- 4) Take pictures**

Statistics 202: Statistical Aspects of Data Mining

Professor David Mease

Course web page:

www.stats202.com

This page is linked from the SCPD web page

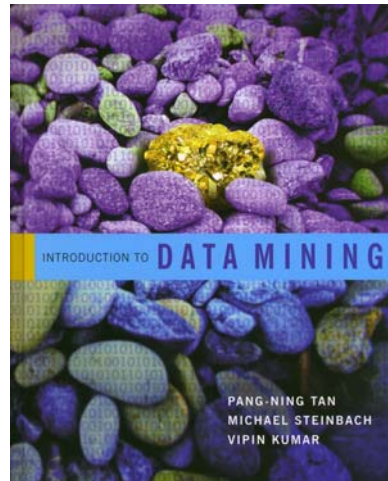
It is also linked from my personal page

www.davemease.com

which is easily found by querying “David Mease” or simply “Mease” on any search engine

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**

In class exercise #1:

**Find a different definition of data mining online.
How does it compare to the one in the text on the
previous slide?**

Data Mining Examples and Non-Examples

Data Mining:

-Certain names are more prevalent in certain US locations (O'Brien, O'Rourke, 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:

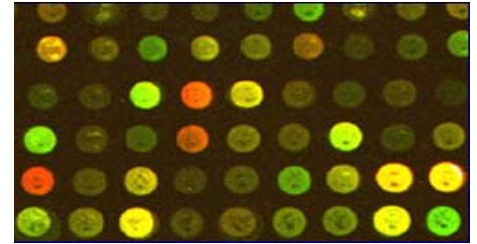
-Look up phone number in phone directory

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

Why Mine Data? Scientific Viewpoint

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

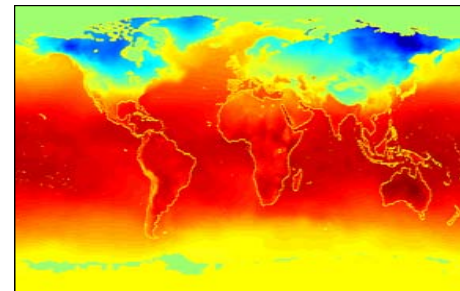
- remote sensors on a satellite
- telescopes scanning the skies
- microarrays generating gene expression data
- scientific simulations generating terabytes of data



- Traditional techniques infeasible for raw data

- 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 cheaper and more powerful

- Competitive pressure is strong

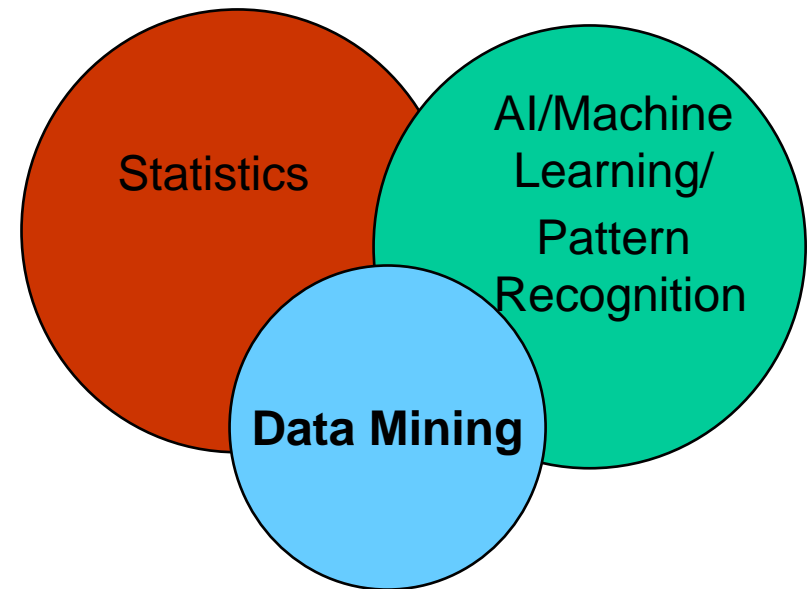
- Provide better, customized services for an edge

In class exercise #2:

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)

- **Prediction Methods:**

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

- **Description 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:

You should make sure you have access to the following two software packages for this course

- **Microsoft Excel**

- **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 Forward Stop Refresh Home 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 packages 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](#)
- [Source of R alpha and beta releases \(daily snapshots, created on](#)


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/>



CRAN

- [Mirrors](#)
- [What's new?](#)
- [Task Views](#)
- [Search](#)

About R

- [R Homepage](#)

Software

- [R Sources](#)
- [R Binaries](#)
- [Packages](#)

This directory contains binaries for a base distribution and packages

Note: CRAN does not have Windows systems and cannot check the

Subdirectories:

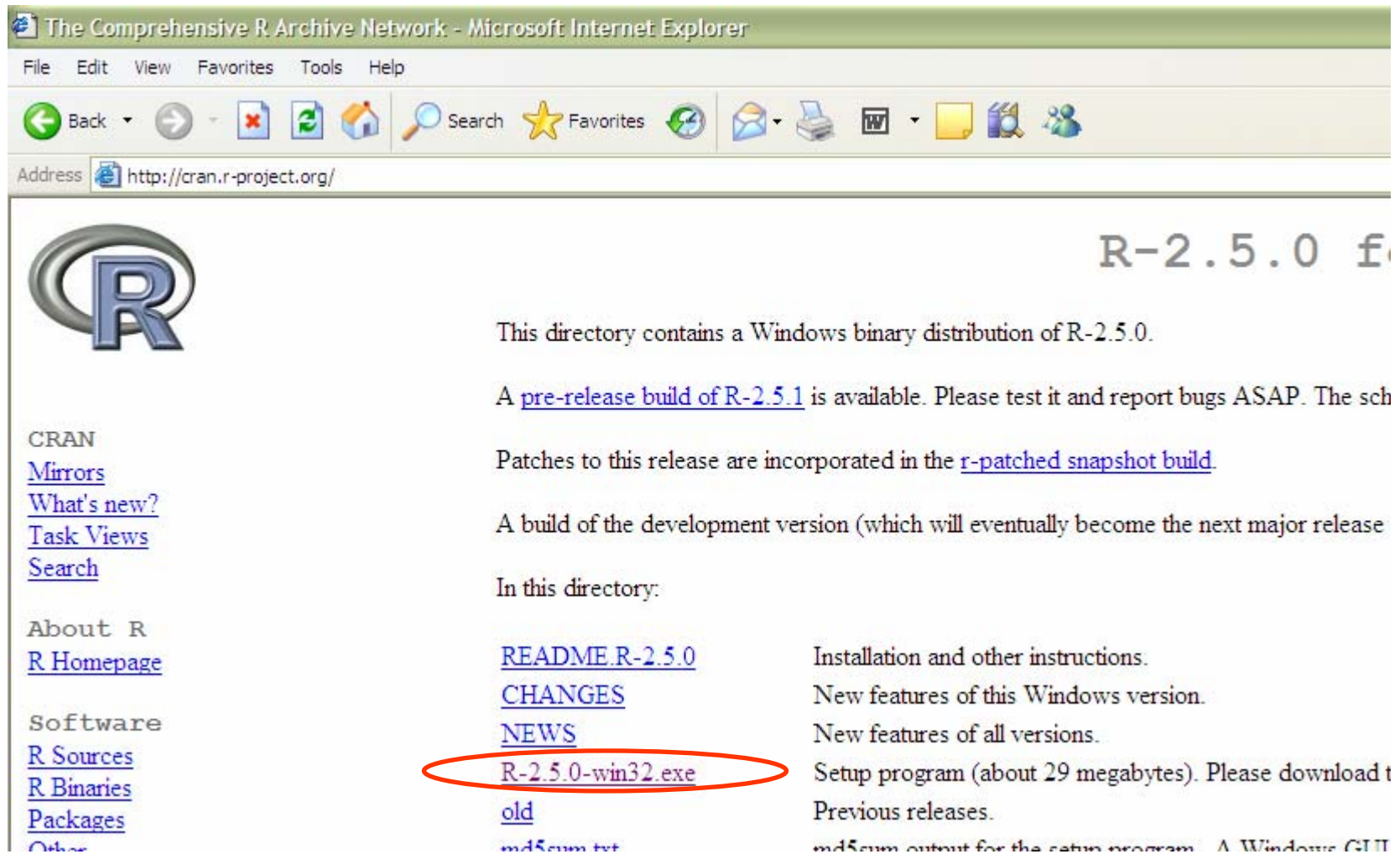
base	Binaries for base distribution (manage
contrib	Binaries of contributed packages (ma

Please do not submit binaries to CRAN. Package developers might v
binaries.

You may also want to read the [R FAQ](#) and [R for Windows FAQ](#).

Last modified: April 4, 2004, by Friedrich Leisch

Downloading R for Windows:




The Comprehensive R Archive Network - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites Home Mail Print View Stop Home People

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



R-2.5.0

This directory contains a Windows binary distribution of R-2.5.0.

A [pre-release build of R-2.5.1](#) is available. Please test it and report bugs ASAP. The schedule of releases is available at [CRAN](#).

Patches to this release are incorporated in the [r-patched snapshot build](#).

A build of the development version (which will eventually become the next major release) is available at [CRAN](#).

In this directory:

README.R-2.5.0	Installation and other instructions.
CHANGES	New features of this Windows version.
NEWS	New features of all versions.
R-2.5.0-win32.exe	Setup program (about 29 megabytes). Please download the
old	Previous releases.
md5sum.txt	md5sum output for the setup program. A Windows GUI

Pictures:

This is just to help me remember your names.

No one will see these but me.

If you don't want your picture taken please let me know when I come to your seat.

Remote students may email me pictures if you like, but there is no need if I will never see you.