CISC 520:
Data Engineering and Mining

Daqing Yun
Harrisburg University

Data Mining Tools - A Brief Review

- Data access tools
- Data integration tools
- Data exploration tools
- Modeling management tools
- Modeling analysis tools
- Miscellaneous

Data mining tools

Data access tools
SQL and other database query language

Data integration tools
Extract-Transform-Load (ETL) tools to access, modify, and load data from different structures and formats into a common output format

Data exploration tools
Basic descriptive statistics, particularly frequency tables

Model management tools
Data mining workspace libraries, templates, and projects

Modeling analysis tools
Feature selection; model evaluation tools

Miscellaneous tools
In-place Data Processing (IDP) tools, rapid deployment tools, model monitoring tools

Data access tools

Structured Query Language (SQL) Tools

Microsoft SQL server

https://www.microsoft.com/en-us/sql-server/sql-server-2016

Linux SQL tools

A list: https://www.linas.org/linux/db.html

e.g., Squirrel SQL: http://squirrel-sql.sourceforge.net/

MySQL

https://www.mysql.com/

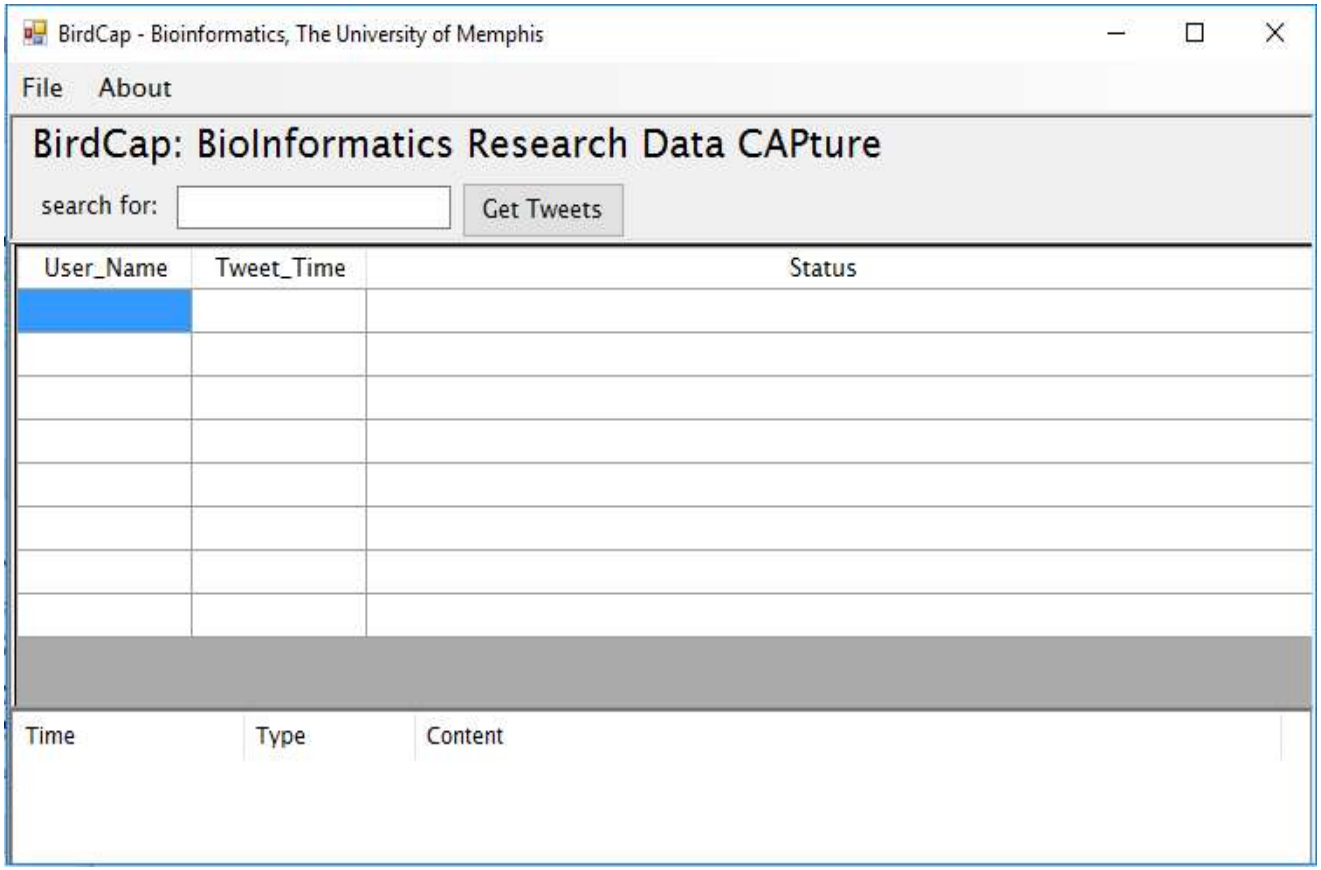
phpMyAdmin

(try demo: https://demo.phpmyadmin.net/master-config/)

...

Data access tools

Extraction, transformation, and loading of data can be performed in native SQL programs (C#)



Data access tools

Extraction, transformation, and loading of data can be performed in native SQL programs (C#)

```
OleDbConnection dbCon = null;
/* open database connection */
try {
    dbCon = new OleDbConnection(@"Provider=Microsoft.Jet.OLEDB.4.0;Data Source=F:\tweetsDB.mdb");
    dbCon.Open();
    ...
} catch (Exception e) {
    ...
}
/* construct sql statement */
try {
    OleDbCommand insertCmd = new OleDbCommand();
    insertCmd.Connection = dbCon;

    insertCmd.CommandText = "insert into tweets (id_str,"
                          + "created_at,"
                          + "from_user_name,"
                          ...)";
        insertCmd.Parameters.AddWithValue("@id_str", t.id_str);
        insertCmd.Parameters.AddWithValue("@created_at", t.created_at);
        insertCmd.Parameters.AddWithValue("@from_user_name", t.from_user_name);
        ...
    /* execute the sql command */
    insertCmd.ExecuteNonQuery();
    /* clear the parameters for the command */
    insertCmd.Parameters.Clear();
}

dbCon.Close();
...
```

Data integration tools

ETL functions

Extract, Transform, and Load

STATISTICA Data Miner

http://www.statsoft.com/products/statistica/data-miner

- ✓ **Extracting data**: stores the metadata describing the nature of the tables that are queried
- ✓ **Transforming data**: supports operations for transposing, sorting, and ranking of data, in addition to standardizing, transforming and stacking variables
- ✓ **Loading data**: automate the process of validating and aligning multiple diverse data sources into a single source suitable for ad hoc or automated analyses

Data integration tools

STATISTICA Data Miner

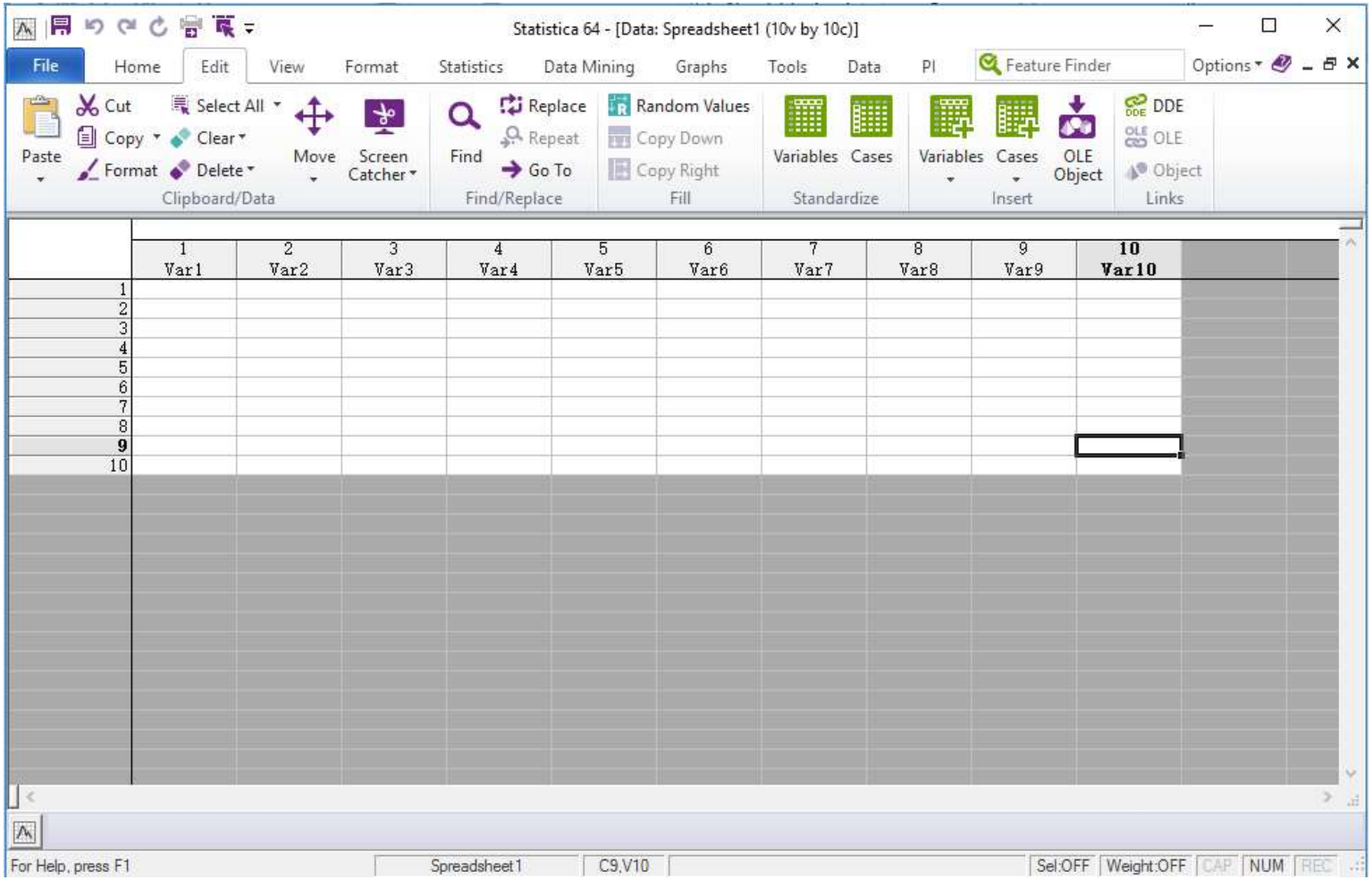
http://www.statsoft.com/products/statistica/data-miner

You can try the free trial

Data integration tools

STATISTICA Data Miner

Looks like…

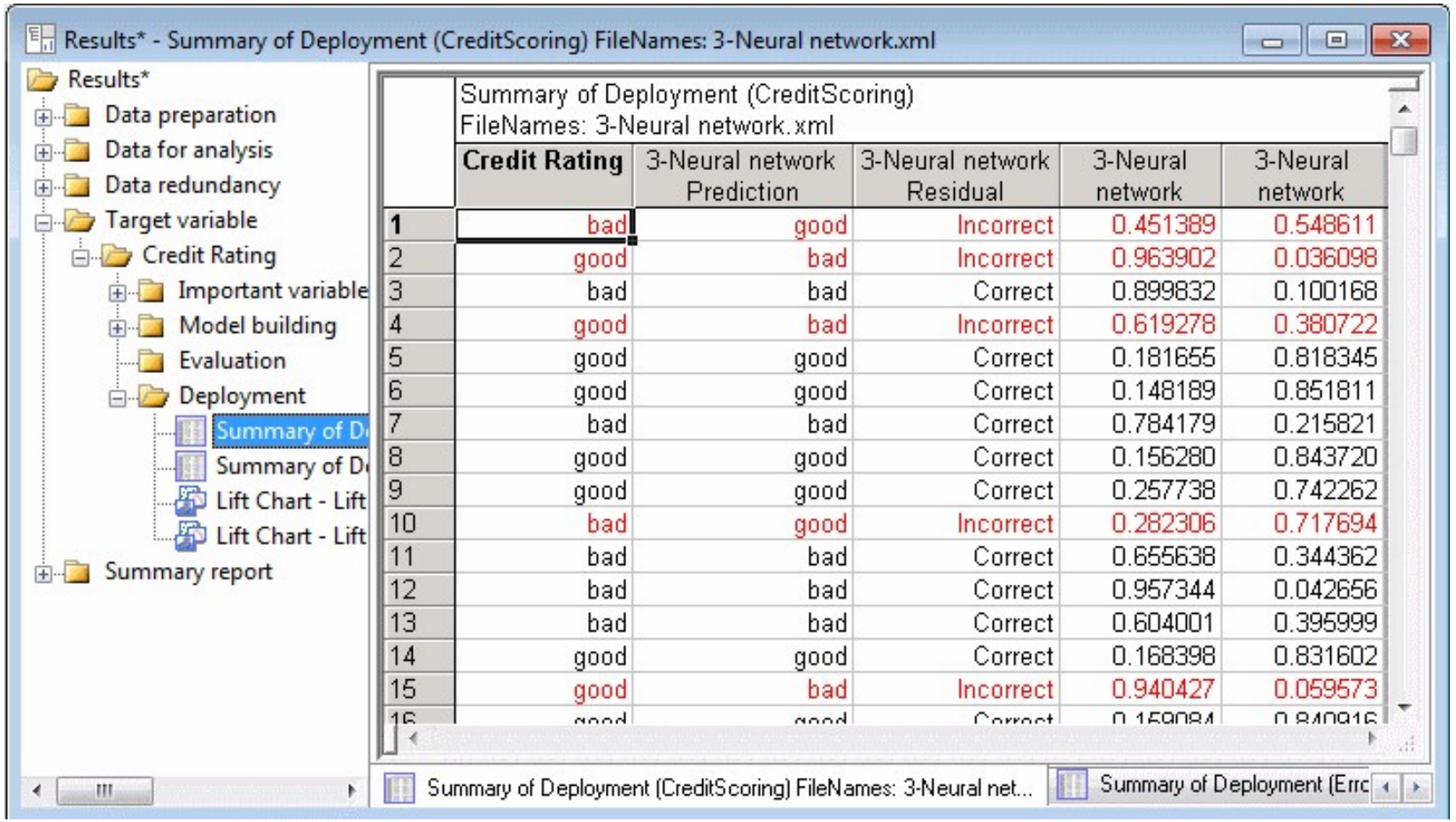


Data integration tools

STATISTICA Data Miner

Example: credit scoring applications - a step by step example

http://documentation.statsoft.com/STATISTICAHelp.aspx?path=DMR/DataMinerRecipeExample



Data exploration tools - Basic descriptive statistics

We have discussed this in last time lecture

Measures of location

- ✓ **Mean**: average for all observations in the range of a variable
- ✓ **Median**: middle observation in a sorted list of values in the range for a given variable
- ✓ **Mode**: most frequently occurring value

Measures of dispersion

- ✓ **Variance**: a measure of the variability of squared values around the mean
- ✓ **Standard deviation**: square root of the variance

Range

- ✓ **Maximum**: highest value in the range of a variable
- ✓ **Minimum**: lowest value in the range of a variable

Data exploration tools - Basic descriptive statistics

Measures of position

- ✓ **Quantiles**: a portion of the total number of observations (quartiles, percentiles)
- ✓ **The PTH percentile**: value where at least *p* percent of the items are less than or equal to this value
- ✓ **Median percentile**: 50th percentile
- ✓ **Q1**: 1st quartile = 25th percentile
- ✓ **Q3**: 3rd quartile = 75th percentile

Measures of shape

- ✓ **Skewness**: degree to which the distribution of data for a variable is largely to one side of the mean
- ✓ **Kurtosis**: degree to which distribution of the data for a variable is closely arranged around the mean

Robust measures of location

- ✓ **Trimmed mean**: calculated by removing a percentage of values from both ends of the data set
- ✓ **Winsorized mean**: the mean computed after the x-percentage highest and lowest values are replaced by the next adjacent value in the distribution

Data exploration tools - Basic descriptive statistics

Frequency tables

For example, in a survey research, frequency table can show:

- ✓ the number of males and females who participated in the survey
- ✓ the number of respondents from particular ethnic and racial backgrounds
- ✓ Etc.

Frequency or **one-way tables** represent the simplest method for analyzing categorical (nominal) data

- ✓ For example, review how different categories of values are distributed in the sample
- ✓ For example, a survey of spectator interest in different sports (Table 6.1 in textbook)
- ✓ STATISTICA Data Miner can automatically generate frequency tables and histograms for both continuous and categorical variables

	Frequency Table: Football: "Watching Football"			
Category	Count	Cumulative Count	Percent	Cumulative Percent
Always: Always interested	39	39	39.00000	39.0000
Usually: Usually interested	16	55	16.00000	55.0000
Sometimes: Sometimes interested	26	81	26.00000	81.0000
Never: Never interested	19	100	19.00000	100.0000
Missing	0	100	0.00000	100.0000

Data exploration

Basic Descriptive Statistics - MS Excel

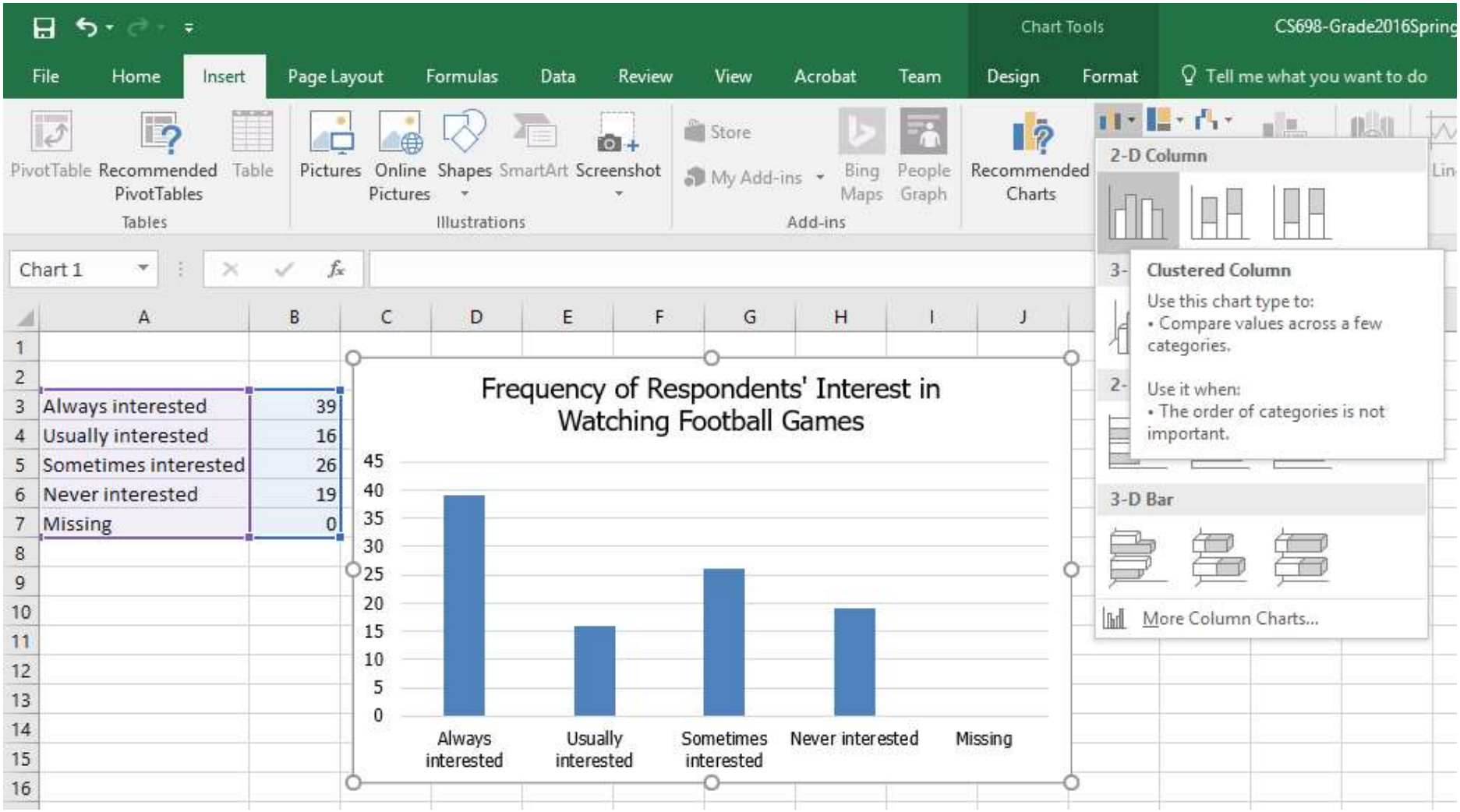
Student ID	Student	E-mail	Major	Attend1	Attend2	Attend3	Attend4	Attend5
				0				
				100				
				100				
				100				
				100				
				100				
				100				
				100				
				100				
				100				
				100				
				100				
				0				
				100				
				0				
				100				
				0				
				0				
				100				
				100				
				0				
				0				
				100				
				68.181818				
				47.673129				

=AVERAGE(E4:E25)

=STDEV(E4:E25)

Data exploration tools

Frequency histogram - MS Excel



Data exploration tools

Slicing/Dicing and Drilling Down into Data Sets/Results Spreadsheets

A “deep dive” into details and aspects of a data set

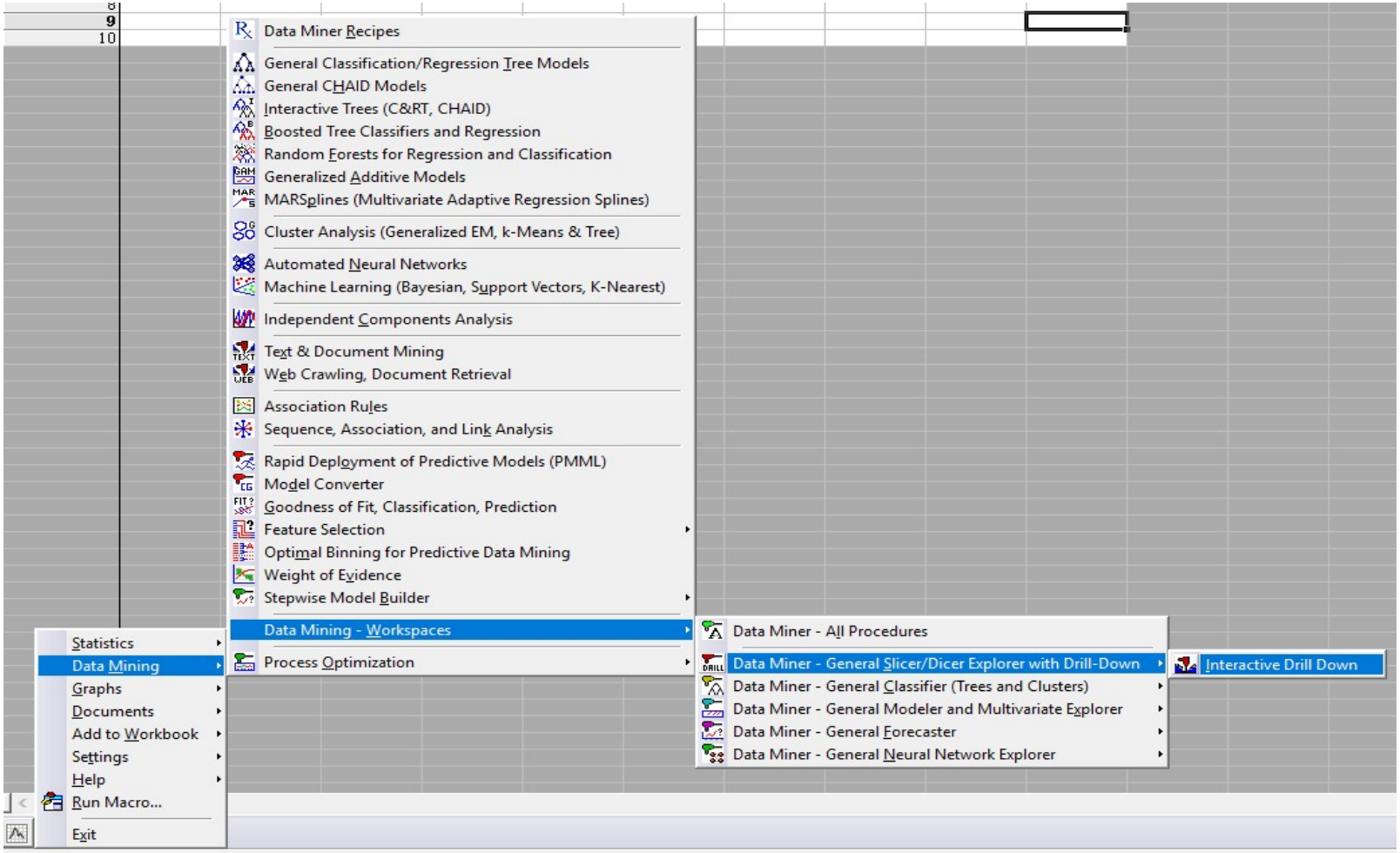
Allow you to specify which variables to analyze

STATISTICA Data Miner

Data exploration tools

Slicing/Dicing and Drilling Down into Data Sets/Results Spreadsheets

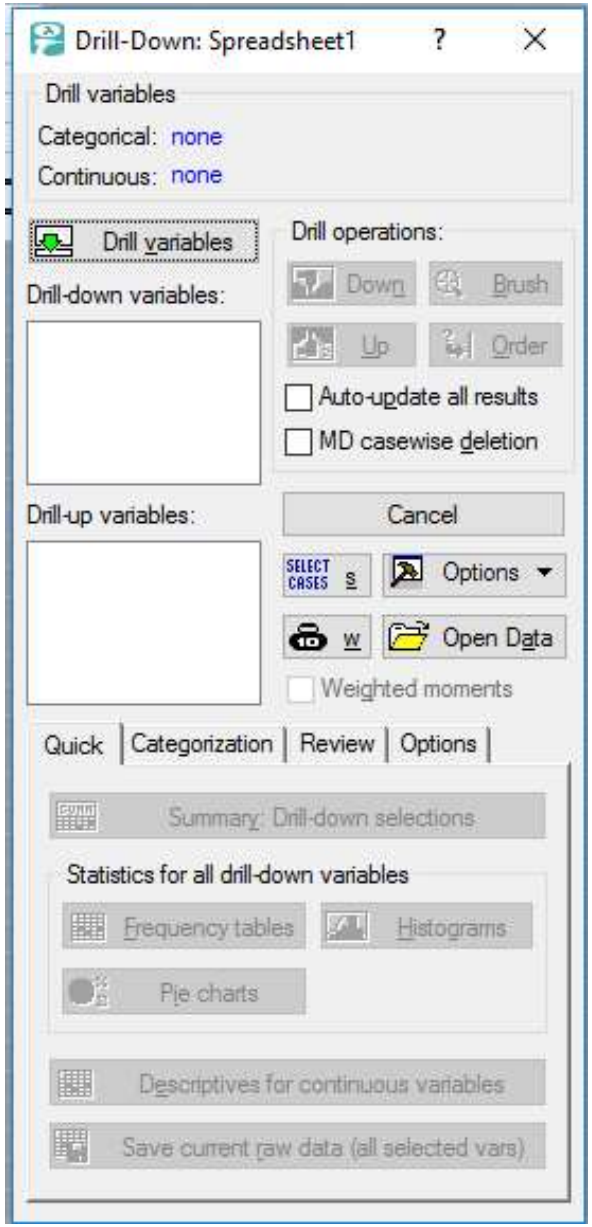
The menu pathway in STATISTICA for accessing the Interactive Drill Down tool



Data exploration tools

Slicing/Dicing and Drilling Down into Data Sets/Results Spreadsheets

If you select the Interactive Drill Down option, an interactive dialog box will appear, allowing you to specify which variables to analyze



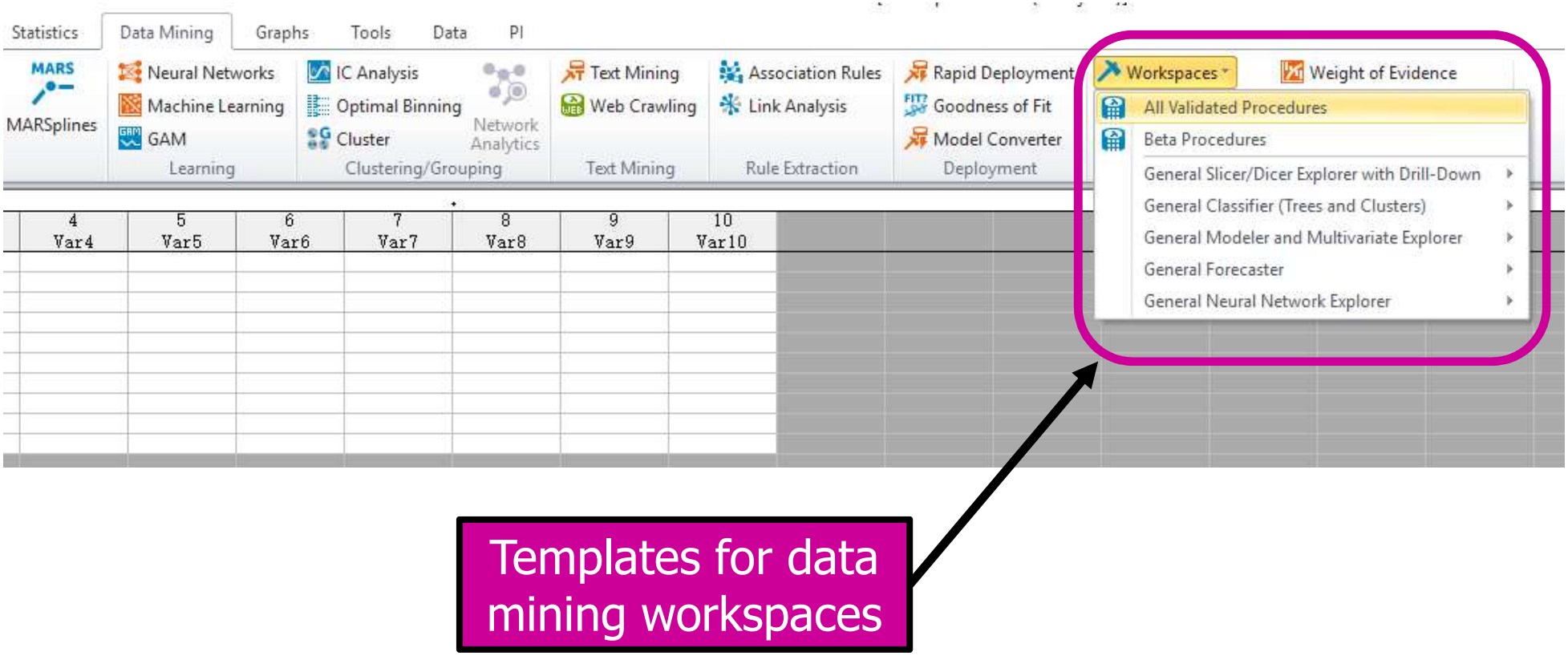
Modeling management tools

As you get used to making data mining projects, you may want to start from a *blank* Data Miner Workspace, adding each thing needed as you create the project

But a good way to start is to use predefined *templates*. These templates already have DM Nodes placed in the workspace; thus, you only have to input the data set and any other nodes to use these templates as a fast method for initial exploration of a data set

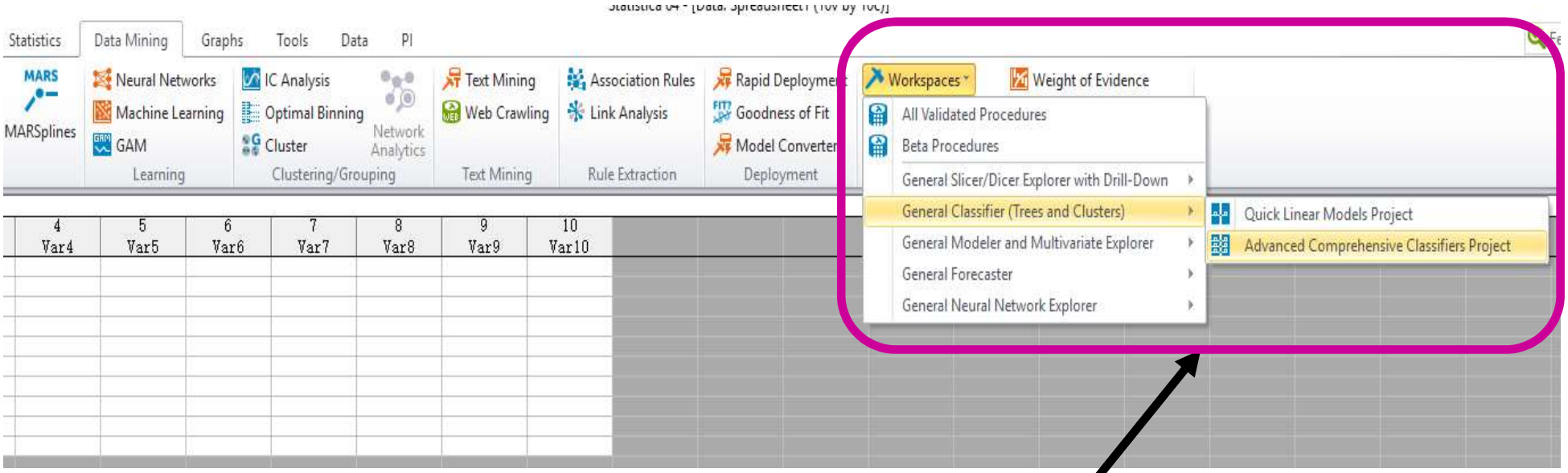
Modeling management tools

Data Mining → Workspaces



Modeling management tools

Data Mining → Workspaces



Let's release on Advanced Comprehensive Classifiers Project template

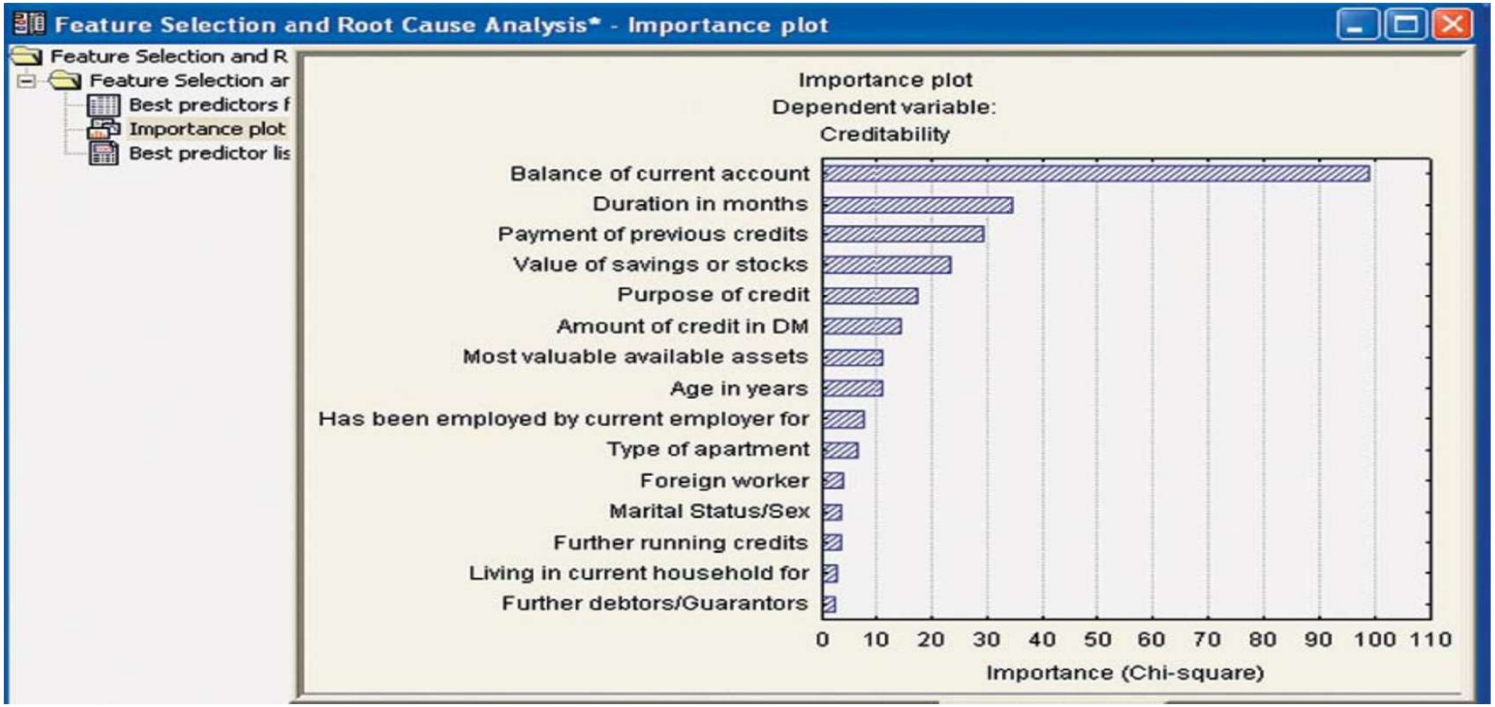
Modeling management tools

Feature selection

- ✓ Can save a lot of time by reducing the number of variables, i.e., the "dimensionality"
- ✓ Thus increase probability that the model will be more robust
- ✓ Will be discussed in later lecture

Importance plots of variables

- ✓ Credit Scoring data set, similar to those used by bankers and credit card companies to determine whether to give credit to an applicant



In-place data processing

In-place Data Processing (IDP)

The *conventional* way to access data in database tables is to extract that information using an Open Database Connectivity (ODBC) driver

The problems:

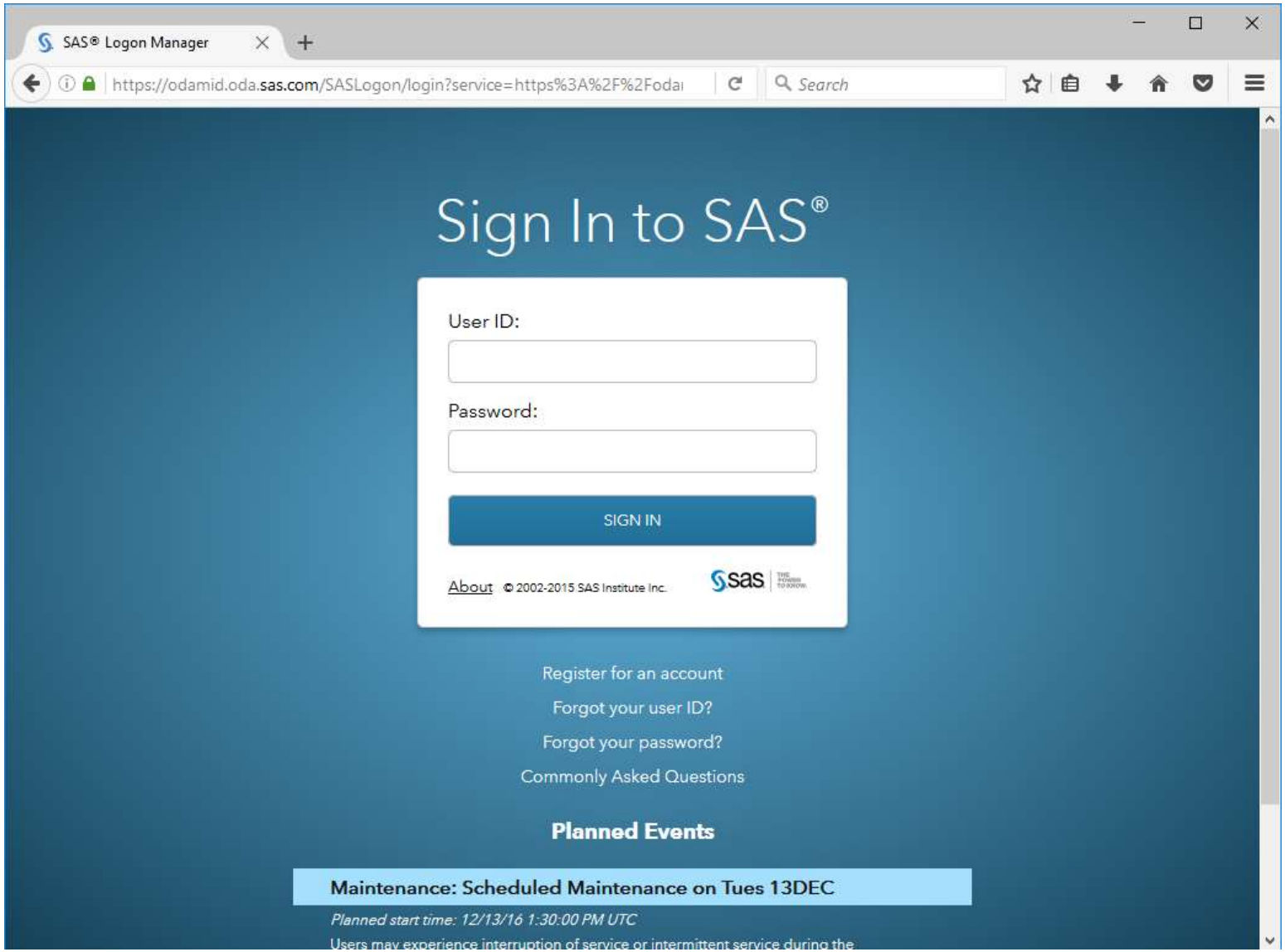
- ✓ The *space* required to hold the extracted data in the form of flat files
- ✓ The need to *duplicate* data on an analytical computing system
- ✓ The need to *integrate* multiple extracts to form the analytic record for data mining processing
- ✓ The *time* required for download, scheduling of downloads
- ✓ The *difficulty* in working with very large data sets
- ✓ The need for the *ODBC driver software* to be available and properly configured for the two systems participating in the download operation

Several approaches available to permit analytical processing of data without extraction to external flat files, and access data directly in tables in a database

- ✓ SAS-Enterprise Miner
- ✓ STATISTICA Data Miner
- ✓ SPSS Clementine: provides links to data mining tools for various database management system vendors

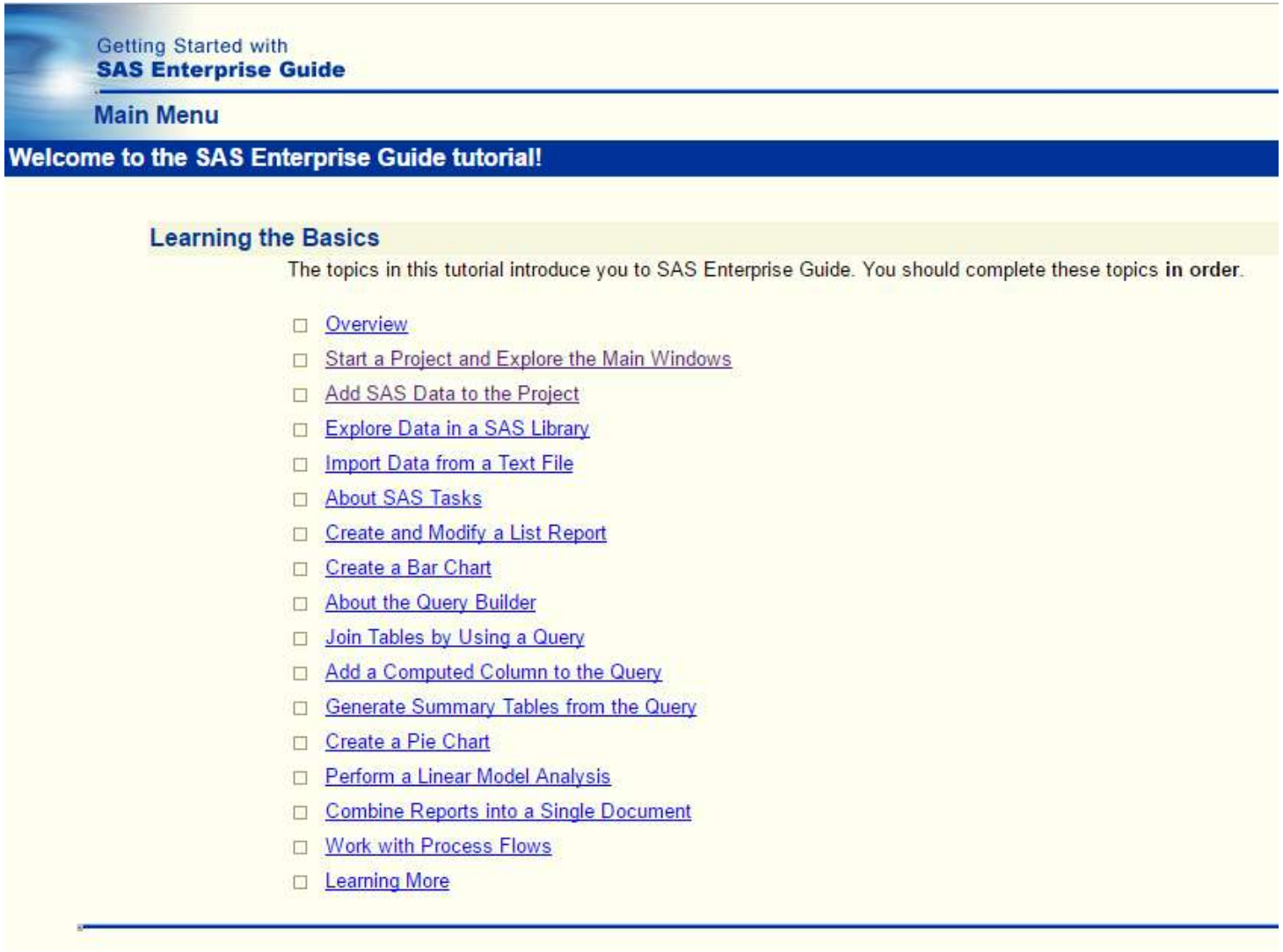
SAS enterprise

http://support.sas.com/ctx3/sodareg/index.html



SAS enterprise

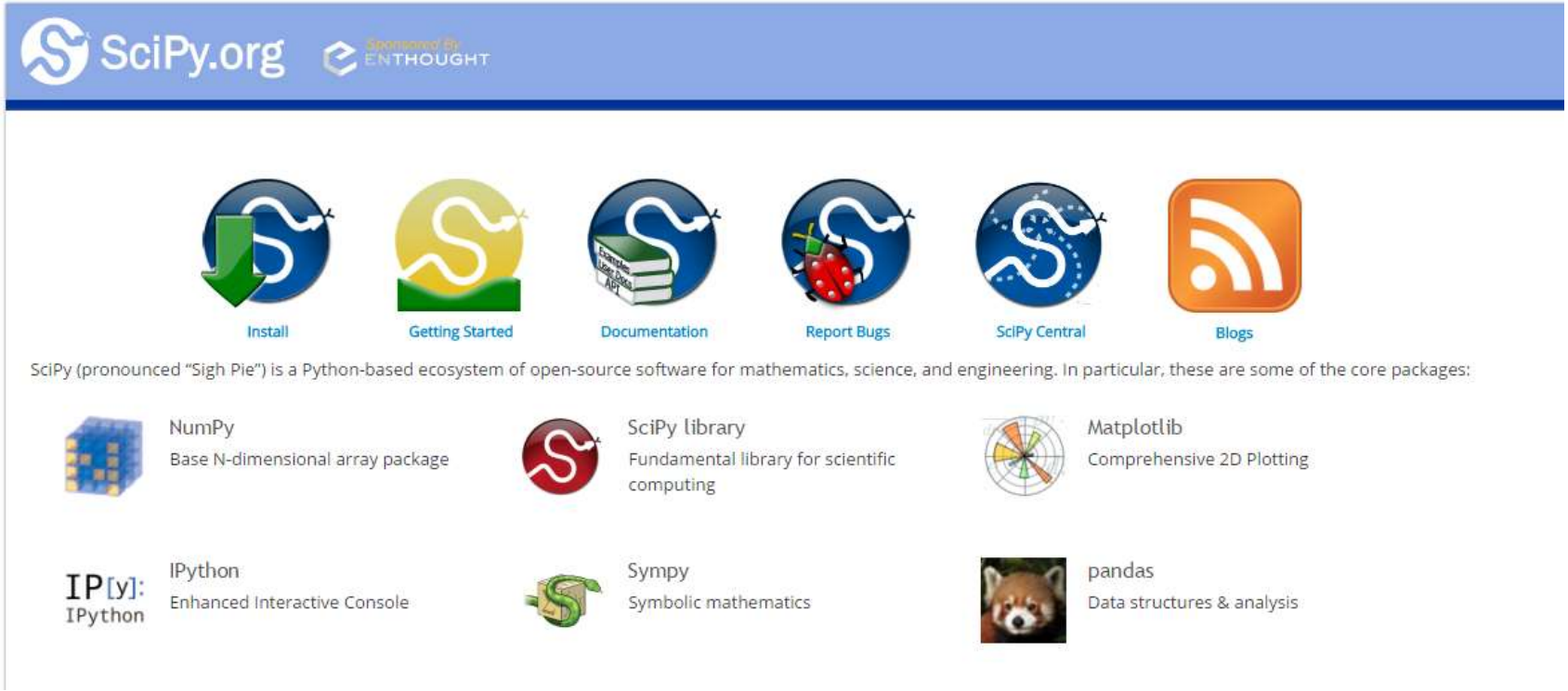
http://support.sas.com/documentation/onlinedoc/guide/tut71/en/menu.htm



Python (libraries)

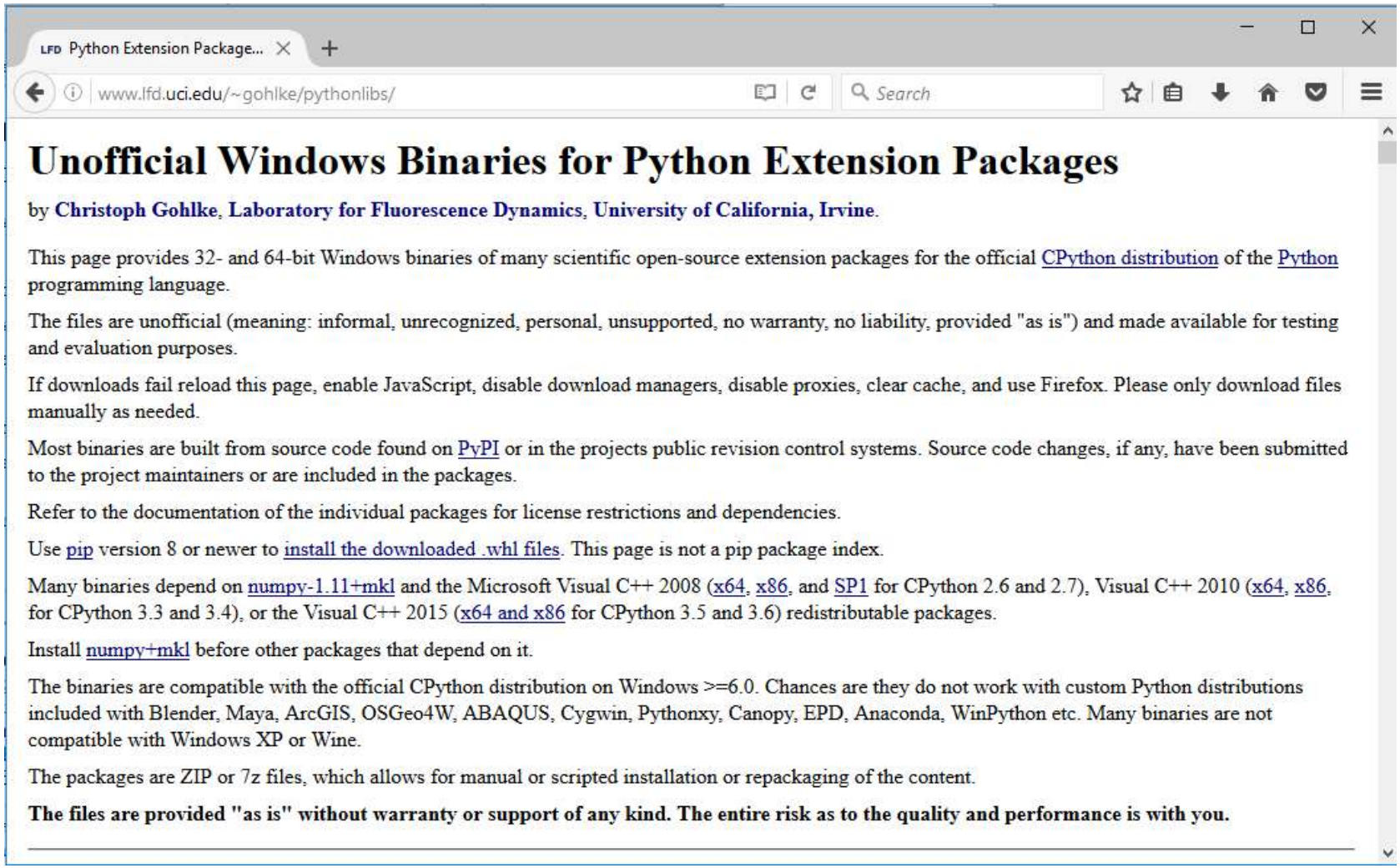
https://www.python.org/

Scientific Computing Tools for Python
https://www.scipy.org/



Python (libraries)

http://www.lfd.uci.edu/~gohlke/pythonlibs/
Unofficial Windows Binaries for Python Extension Packages



R for statistical computing

https://www.r-project.org/

[Home]

Download

CRAN

R Project

About R
Logo
Contributors
What's New?
Reporting Bugs
Development Site
Conferences
Search

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To **download R**, please choose your preferred CRAN mirror.

If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

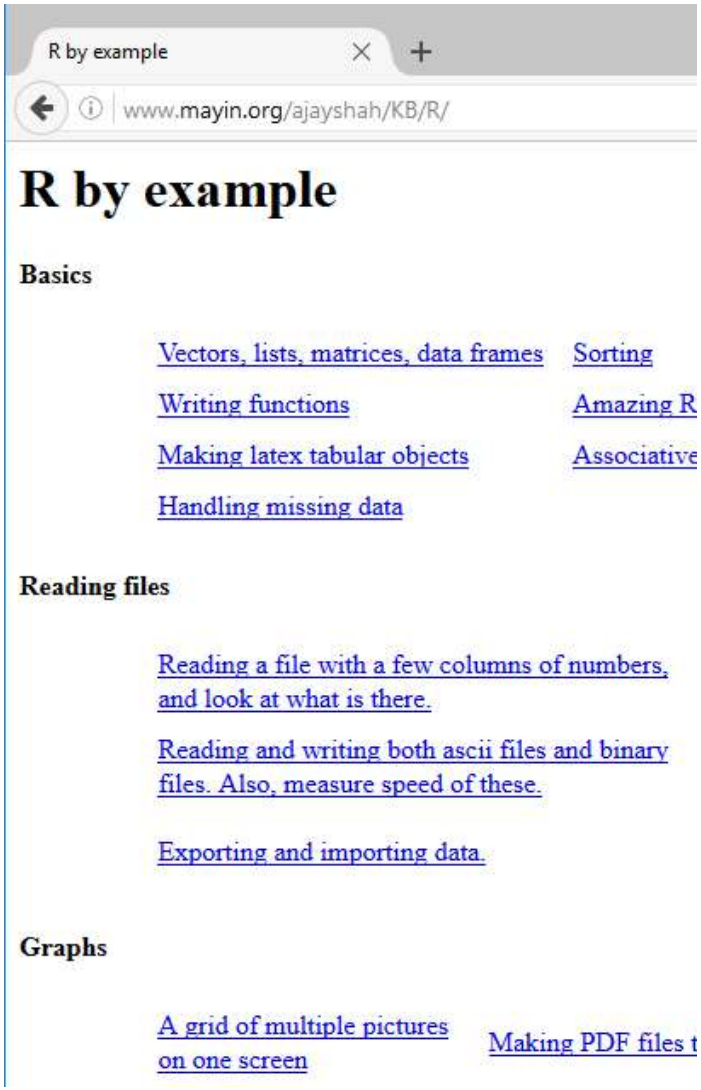
News

- The R Foundation welcomes five new ordinary members: Jennifer Bryan, Dianne Cook, Julie Josse, Tomas Kalibera, and Balasubramanian Narasimhan.
- **R version 3.3.2 (Sincere Pumpkin Patch)** has been released on Monday 2016-10-31.
- **The R Journal Volume 8/1** is available.

R for statistical computing

Learn by examples: http://www.mayin.org/ajayshah/KB/R/

- ✓Basics
- ✓Reading files
- ✓Graphs
- ✓Probability and statistics
- ✓Regression
- ✓Time-series analysis
- ✓Suggestions for learning R
- ✓…



Remind - Hadoop-related Apache Projects

- Ambari™: A web-based tool for provisioning, managing, and monitoring Hadoop clusters.It also provides a dashboard for viewing cluster health and ability to view MapReduce, Pig and Hive applications visually.
- Avro™: A data serialization system.
- Cassandra™: A scalable multi-master database with no single points of failure.
- Chukwa™: A data collection system for managing large distributed systems.
- HBase™: A scalable, distributed database that supports structured data storage for large tables.
- Hive™: A data warehouse infrastructure that provides data summarization and ad hoc querying.
- Mahout™: A Scalable machine learning and data mining library.
- Pig™: A high-level data-flow language and execution framework for parallel computation.
- Spark™: A fast and general compute engine for Hadoop data. Spark provides a simple and expressive programming model that supports a wide range of applications, including ETL, machine learning, stream processing, and graph computation.
- Tez™: A generalized data-flow programming framework, built on Hadoop YARN, which provides a powerful and flexible engine to execute an arbitrary DAG of tasks to process data for both batch and interactive use-cases.
- ZooKeeper™: A high-performance coordination service for distributed applications.

Key Components of Mahout

Collaborative Filtering

- User-Based Collaborative Filtering - single machine
- Item-Based Collaborative Filtering - **single machine** / MapReduce
- Matrix Factorization with Alternating Least Squares - **single machine** / MapReduce
- Matrix Factorization with Alternating Least Squares on Implicit Feedback- **single machine** / MapReduce
- Weighted Matrix Factorization, SVD++, Parallel SGD - **single machine**

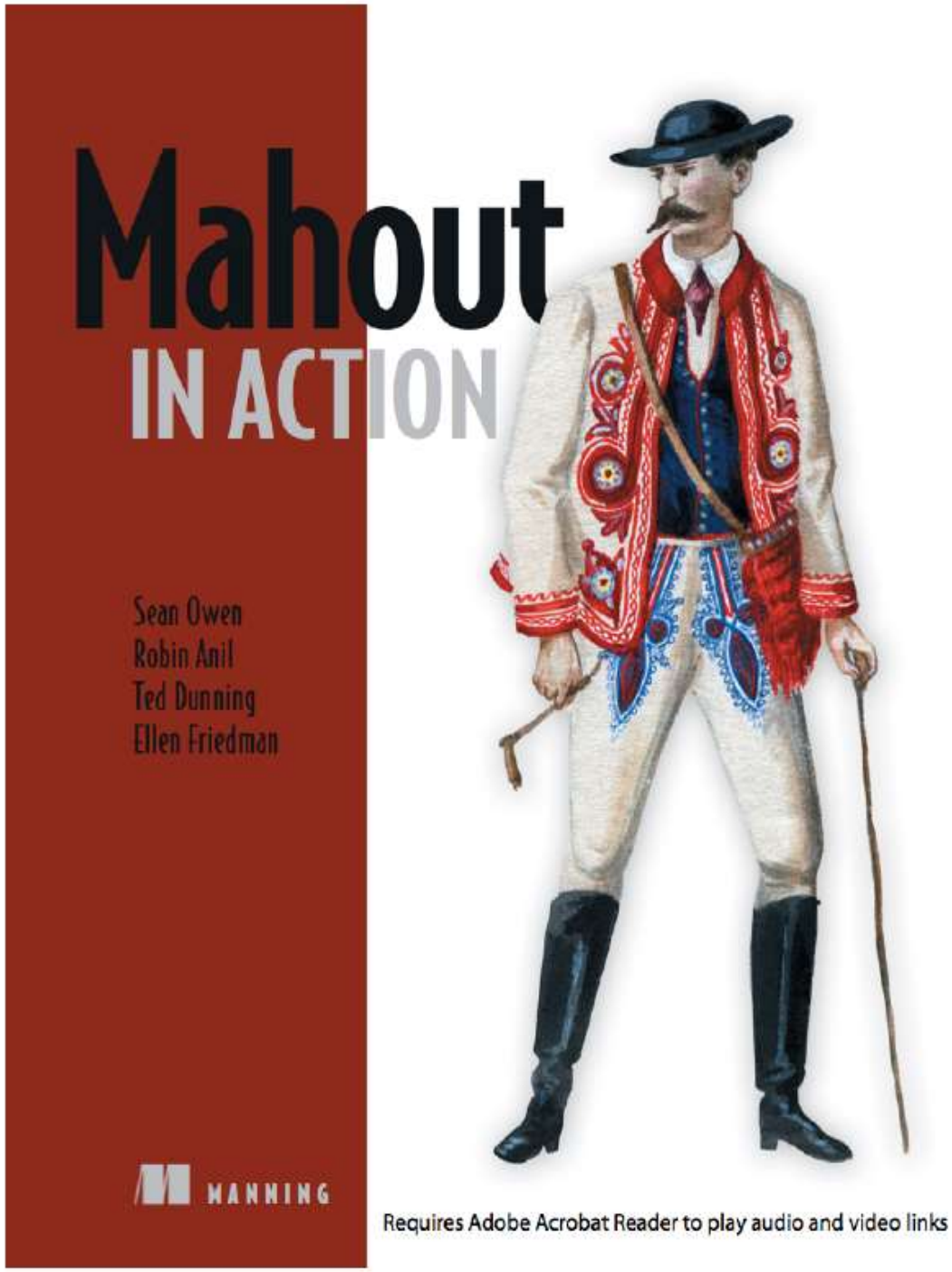
Classification

- Logistic Regression - trained via SGD - single machine
- Naive Bayes/ Complementary Naive Bayes - MapReduce
- Random Forest - **MapReduce**
- Hidden Markov Models - single machine
- Multilayer Perceptron - **single machine**

Clustering

- Canopy Clustering - **single machine** / MapReduce (deprecated, will be removed once Streaming k-Means is stable enough)
- k-Means Clustering - single machine / MapReduce
- Fuzzy k-Means - single machine / MapReduce
- Streaming k-Means - single machine / MapReduce
- Spectral Clustering - MapReduce

Mahout reference book



Requires Adobe Acrobat Reader to play audio and video links

download mahout

Latest release version 0.9 has

- User and Item based recommenders
- Matrix factorization based recommenders
- K-Means, Fuzzy K-Means clustering
- Latent Dirichlet Allocation
- Singular Value Decomposition
- Logistic regression classifier
- (Complementary) Naive Bayes classifier
- Random forest classifier
- High performance java collections
- A vibrant community

25 April 2014 - Goodbye MapReduce

The Mahout community decided to move its codebase onto modern data processing systems that offer a richer programming model and more efficient execution than Hadoop MapReduce. **Mahout will therefore reject new MapReduce algorithm implementations from now on**. We will however keep our widely used MapReduce algorithms in the codebase and maintain them.

We are building our future implementations on top of a DSL for linear algebraic operations which has been developed over the last months. Programs written in this DSL are automatically optimized and executed in parallel on Apache Spark.

Methodology

Project/Problem-driven

1. Clearly formulate your problem
2. What are the inputs?
   - Understand your dataset
   - Cleaning, sampling, reducing, etc. if necessary
3. What are the expected outputs?
   - Be precise about the expected outcome

Problem

---

4. Overview/survey
   - There are some existed
5. Try the tools (by examples)
6. Choose the one for your problem
7. You may use multiple appropriate tools along the way
   - Solve sub-problems using different tool
   - You certainly can write your own code
8. Repeat

Solution

A Review of Software Packages for Data Mining

Dominique HAUGHTON, Joel DEICHMANN, Abdolreza ESHGHI,
Selin SAYEK, Nicholas TEEBAGY, and Heikki TOPI

For Today

The 4th assignment

- Read Chapter 6 of the textbook "*Handbook of Statistical Analysis and Data Mining Applications*"
- Homework 4 is assigned
- Review of data mining tools
  - "A Review of Software Packages for Data Mining"
- Please submit a PDF file through moodle

Thanks ! ☺

Questions ?