CHAPTER THREE:
DATA PREPARATION

CONTEXT AND PERSPECTIVE

Jerry is the marketing manager for a small Internet design and advertising firm. Jerry's boss asks him to develop a data set containing information about Internet users. The company will use this data to determine what kinds of people are using the Internet and how the firm may be able to market their services to this group of users.

To accomplish his assignment, Jerry creates an online survey and places links to the survey on several popular Web sites. Within two weeks, Jerry has collected enough data to begin analysis, but he finds that his data needs to be denormalized. He also notes that some observations in the set are missing values or they appear to contain invalid values. Jerry realizes that some additional work on the data needs to take place before analysis begins.

LEARNING OBJECTIVES

After completing the reading and exercises in this chapter, you should be able to:

- Explain the concept and purpose of data scrubbing
- List possible solutions for handling missing data
- Explain the role and perform basic methods for data reduction
- Define and handle inconsistent data
- Discuss the important and process of attribute reduction

APPLYING THE CRISP DATA MINING MODEL

Recall from Chapter 1 that the CRISP Data Mining methodology requires three phases *before* any actual data mining models are constructed. In the Context and Perspective paragraphs above, Jerry has a number of tasks before him, each of which fall into one of the first three phases of CRISP. First, Jerry must ensure that he has developed a clear **Organizational Understanding**. What is the purpose of this project for his employer? Why is he surveying Internet users? Which data points are important to collect, which would be nice to have, and which would be irrelevant or even distracting to the project? Once the data are collected, who will have access to the data set and through what mechanisms? How will the business ensure privacy is protected? All of these questions, and perhaps others, should be answered before Jerry even creates the survey mentioned in the second paragraph above.

Once answered, Jerry can then begin to craft his survey. This is where **Data Understanding** enters the process. What database system will he use? What survey software? Will he use a publicly available tool like SurveyMonkey™, a commercial product, or something homegrown? If he uses publicly available tool, how will he access and extract data for mining? Can he trust this third-party to secure his data and if so, why? How will the underlying database be designed? What mechanisms will be put in place to ensure consistency and integrity in the data? These are all questions of data understanding. An easy example of ensuring consistency might be if a person's home city were to be collected as part of the data. If the online survey just provides an open text box for entry, respondents could put just about anything as their home city. They might put New York, NY, N.Y., Nwe York, or any number of other possible combinations, including typos. This could be avoided by forcing users to select their home city from a dropdown menu, but considering the number cities there are in most countries, that list could be unacceptably long! So the choice of how to handle this potential data consistency problem isn't necessarily an obvious or easy one, and this is just one of many data points to be collected. While 'home state' or 'country' may be reasonable to constrain to a dropdown, 'city' may have to be entered freehand into a textbox, with some sort of data correction process to be applied later.

The 'later' would come once the survey has been developed and deployed, and data have been collected. With the data in place, the third CRISP-DM phase, **Data Preparation**, can begin. If you haven't installed OpenOffice and RapidMiner yet, and you want to work along with the examples given in the rest of the book, now would be a good time to go ahead and install these applications. Remember that both are freely available for download and installation via the Internet, and the links to both applications are given in Chapter 1. We'll begin by doing some data preparation in OpenOffice Base (the database application), OpenOffice Calc (the spreadsheet application), and then move on to other data preparation tools in RapidMiner. You should understand that the examples of data preparation in this book are only a subset of possible data preparation approaches.

COLLATION

Suppose that the database underlying Jerry's Internet survey is designed as depicted in the screenshot from OpenOffice Base in Figure 3-1.

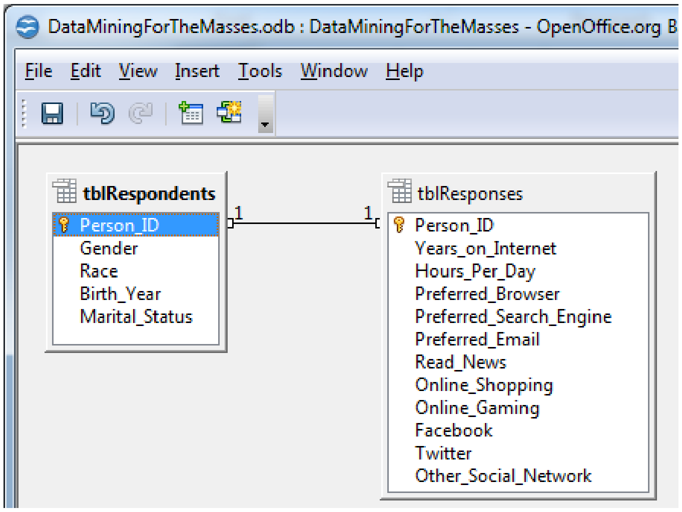
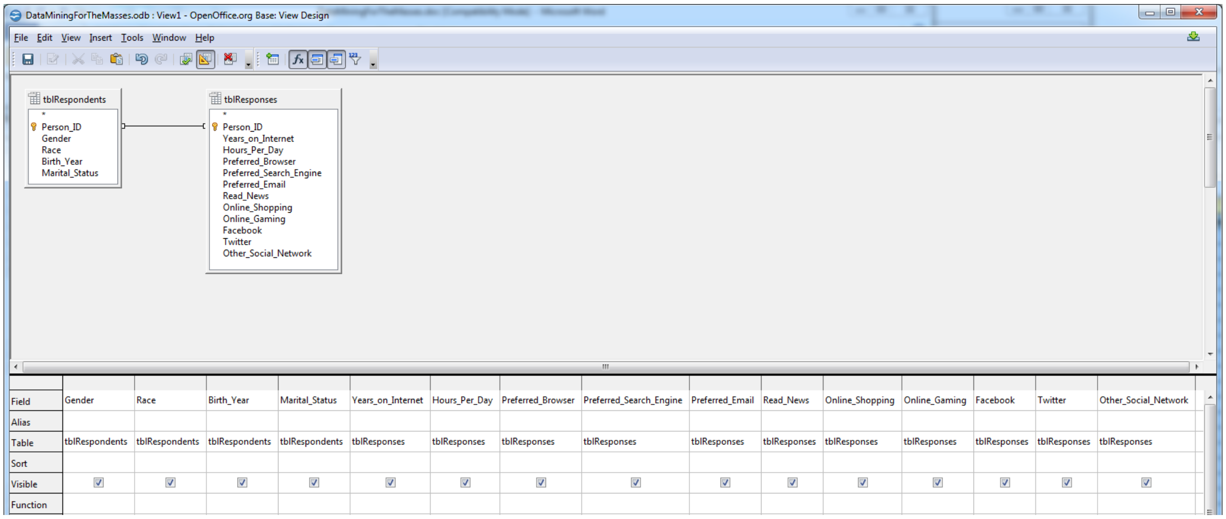


Figure 3-1: A simple relational (one-to-one) database for Internet survey data.

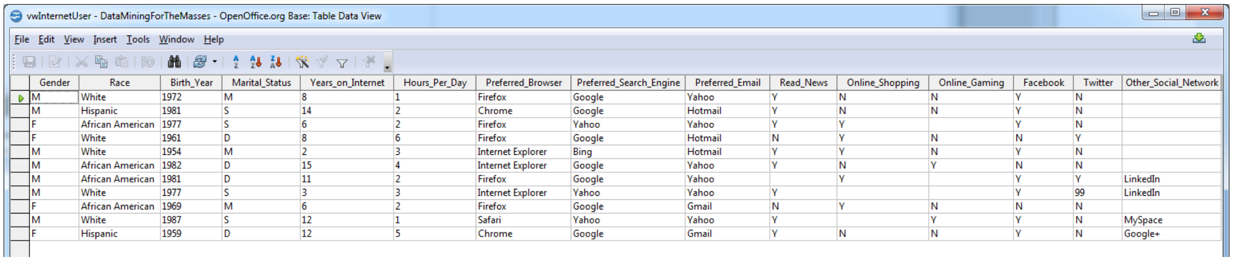
This design would enable Jerry to collect data about people in one table, and data about their Internet behaviors in another. RapidMiner would be able to connect to either of these tables in order to mine the responses, but what if Jerry were interested in mining data from both tables at once?

One simple way to collate data in multiple tables into a single location for data mining is to create a database view. A **view** is a type of pseudo-table, created by writing a SQL statement which is named and stored in the database. Figure 3-2 shows the creation of a view in OpenOffice Base, while Figure 3-3 shows the view in datasheet view.



Field	Gender	Race	Birth_Year	Marital_Status	Years_on_Internet	Hours_Per_Day	Preferred_Browser	Preferred_Search_Engine	Preferred_Email	Read_News	Online_Shopping	Online_Gaming	Facebook	Twitter	Other_Social_Network
Alias															
Table	tblRespondents	tblRespondents	tblRespondents	tblRespondents	tblResponses	tblResponses	tblResponses	tblResponses	tblResponses	tblResponses	tblResponses	tblResponses	tblResponses	tblResponses	tblResponses
Sort															
Visible	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Function															

Figure 3-2: Creation of a view in OpenOffice Base.


vwInternetUser - DataMiningForTheMasses - OpenOffice.org Base: Table Data View
File Edit View Insert Tools Window Help

Gender	Race	Birth_Year	Marital_Status	Years_on_Internet	Hours_Per_Day	Preferred_Browser	Preferred_Search_Engine	Preferred_Email	Read_News	Online_Shopping	Online_Gaming	Facebook	Twitter	Other_Social_Network
M	White	1972	M	8	1	Firefox	Google	Yahoo	Y	N	N	Y	N	
M	Hispanic	1981	S	14	2	Chrome	Google	Hotmail	Y	N	N	Y	N	
F	African American	1977	S	6	2	Firefox	Yahoo	Yahoo	Y	Y		Y	N	
F	White	1961	D	8	6	Firefox	Google	Hotmail	N	Y	N	N	Y	
M	White	1954	M	2	3	Internet Explorer	Bing	Hotmail	Y	Y	N	Y	N	
M	African American	1982	D	15	4	Internet Explorer	Google	Yahoo	Y	N	Y	N	N	
M	African American	1981	D	11	2	Firefox	Google	Yahoo		Y		Y	Y	LinkedIn
M	White	1977	S	3	3	Internet Explorer	Yahoo	Yahoo	Y			Y	99	LinkedIn
F	African American	1969	M	6	2	Firefox	Google	Gmail	N	Y	N	N	N	
M	White	1987	S	12	1	Safari	Yahoo	Yahoo	Y		Y	Y	N	MySpace
F	Hispanic	1959	D	12	5	Chrome	Google	Gmail	Y	N	N	Y	N	Google+

Figure 3-3: Results of the view from Figure 3-2 in datasheet view.

The creation of views is one way that data from a relational database can be collated and organized in preparation for data mining activities. In this example, although the personal information in the 'Respondents' table is only stored once in the database, it is displayed for each record in the 'Responses' table, creating a data set that is more easily mined because it is both richer in information and consistent in its formatting.

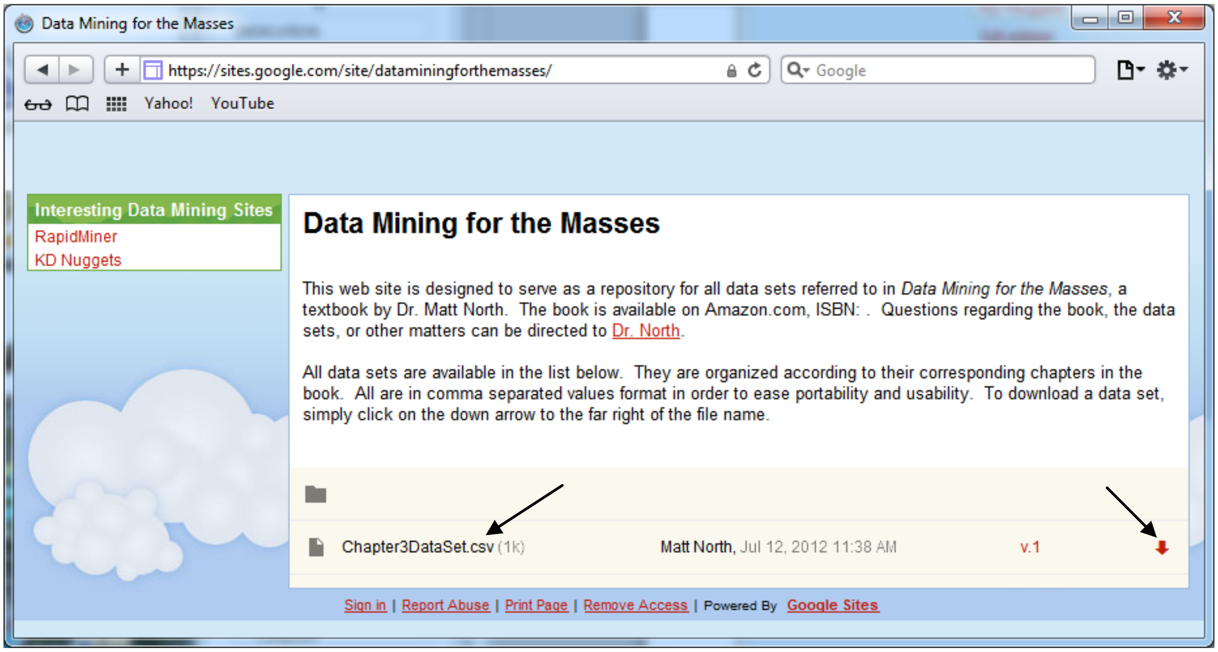
DATA SCRUBBING

In spite of our very best efforts to maintain quality and integrity during data collection, it is inevitable that some anomalies will be introduced into our data at some point. The process of data scrubbing allows us to handle these anomalies in ways that make sense for us. In the remainder of this chapter, we will examine data scrubbing in four different ways: handling missing data, reducing data (observations), handling inconsistent data, and reducing attributes.

HANDS ON EXERCISE

Starting now, and throughout the next chapters of this book, there will be opportunities for you to put your hands on your computer and follow along. In order to do this, you will need to be sure to install OpenOffice and RapidMiner, as was discussed in the section *A Note about Tools* in Chapter 1. You will also need to have an Internet connection to access this book's companion web site, where copies of all data sets used in the chapter exercises are available. The companion web site is located at:

https://sites.google.com/site/dataminingforthemasses/


Data Mining for the Masses
https://sites.google.com/site/dataminingforthemasses/
Google
Yahoo!
YouTube
Interesting Data Mining Sites
RapidMiner
KD Nuggets

Data Mining for the Masses

This web site is designed to serve as a repository for all data sets referred to in *Data Mining for the Masses*, a textbook by Dr. Matt North. The book is available on Amazon.com, ISBN: . Questions regarding the book, the data sets, or other matters can be directed to Dr. North.

All data sets are available in the list below. They are organized according to their corresponding chapters in the book. All are in comma separated values format in order to ease portability and usability. To download a data set, simply click on the down arrow to the far right of the file name.

Chapter3DataSet.csv (1k) Matt North, Jul 12, 2012 11:38 AM v.1

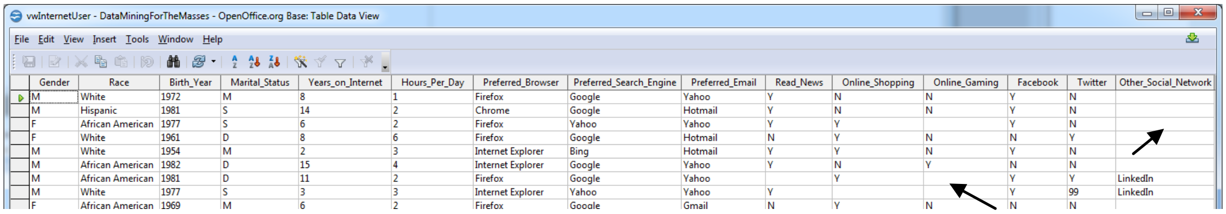
Sign in | Report Abuse | Print Page | Remove Access | Powered By Google Sites

Figure 3-4. *Data Mining for the Masses* companion web site.

You can download the Chapter 3 data set, which is an export of the view created in OpenOffice Base, from the web site by locating it in the list of files and then clicking the down arrow to the far right of the file name, as indicated by the black arrows in Figure 3-4 You may want to consider creating a folder labeled 'data mining' or something similar where you can keep copies of your data—more files will be required and created as we continue through the rest of the book, especially when we get into building data mining models in RapidMiner. Having a central place to keep everything together will simplify things, and upon your first launch of the RapidMiner software, you'll be prompted to create a repository, so it's a good idea to have a space ready. Once you've downloaded the Chapter 3 data set, you're ready to begin learning how to handle and prepare data for mining in RapidMiner.

PREPARING RAPIDMINER, IMPORTING DATA, AND HANDLING MISSING DATA

Our first task in data preparation is to handle missing data, however, because this will be our first time using RapidMiner, the first few steps will involve getting RapidMiner set up. We'll then move straight into handling missing data. **Missing data** are data that do not exist in a data set. As you can see in Figure 3-5, missing data is not the same as zero or some other value. It is blank, and the value is unknown. Missing data are also sometimes known in the database world as **null**. Depending on your objective in data mining, you may choose to leave missing data as they are, or you may wish to replace missing data with some other value.


vwInternetUser - DataMiningForTheMasses - OpenOffice.org Base: Table Data View
File Edit View Insert Tools Window Help

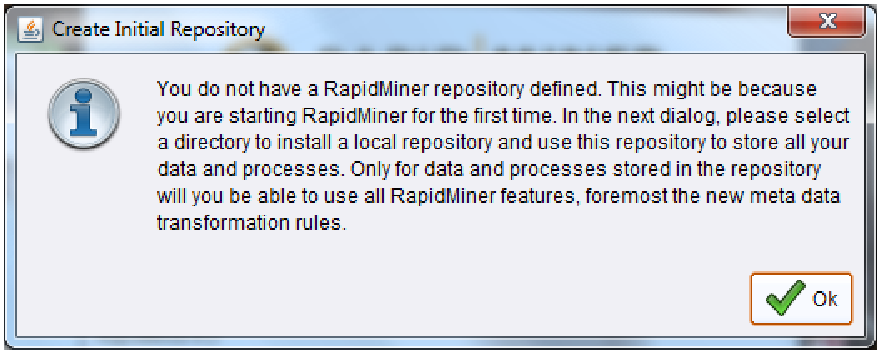
Gender	Race	Birth_Year	Marital_Status	Years_on_Internet	Hours_Per_Day	Preferred_Browser	Preferred_Search_Engine	Preferred_Email	Read_News	Online_Shopping	Online_Gaming	Facebook	Twitter	Other_Social_Network
M	White	1972	M	8	1	Firefox	Google	Yahoo	Y	N	N	Y	N	
M	Hispanic	1981	S	14	2	Chrome	Google	Hotmail	Y	N	N	Y	N	
F	African American	1977	S	6	2	Firefox	Yahoo	Yahoo	Y	Y		Y	N	
F	White	1961	D	8	6	Firefox	Google	Hotmail	N	Y	N	N	Y	
M	White	1954	M	2	3	Internet Explorer	Bing	Hotmail	Y	Y	N	Y	N	
M	African American	1982	D	15	4	Internet Explorer	Google	Yahoo	Y	N	Y	N	N	
M	African American	1981	D	11	2	Firefox	Google	Yahoo		Y		Y	Y	LinkedIn
M	White	1977	S	3	3	Internet Explorer	Yahoo	Yahoo	Y			Y	99	LinkedIn
F	African American	1969	M	6	2	Firefox	Google	Gmail	N	Y	N	N	N	

Figure 3-5: Some missing data within the survey data set.

The creation of views is one way that data from a relational database can be collated and organized in preparation for data mining activities. In this example, our database view has missing data in a number of its attributes. Black arrows indicate a couple of these attributes in Figure 3-5 above. In some instances, missing data are not a problem, they are expected. For example, in the Other Social Network attribute, it is entirely possible that the survey respondent did not indicate that they use social networking sites other than the ones proscribed in the survey. Thus, missing data are probably accurate and acceptable. On the other hand, in the Online Gaming attribute, there are answers of either 'Y' or 'N', indicating that the respondent either does, or does not participate in online gaming. But what do the missing, or null values in this attribute indicate? It is unknown to us. For the purposes of data mining, there are a number of options available for handling missing data.

To learn about handling missing data in RapidMiner, follow the steps below to connect to your data set and begin modifying it:

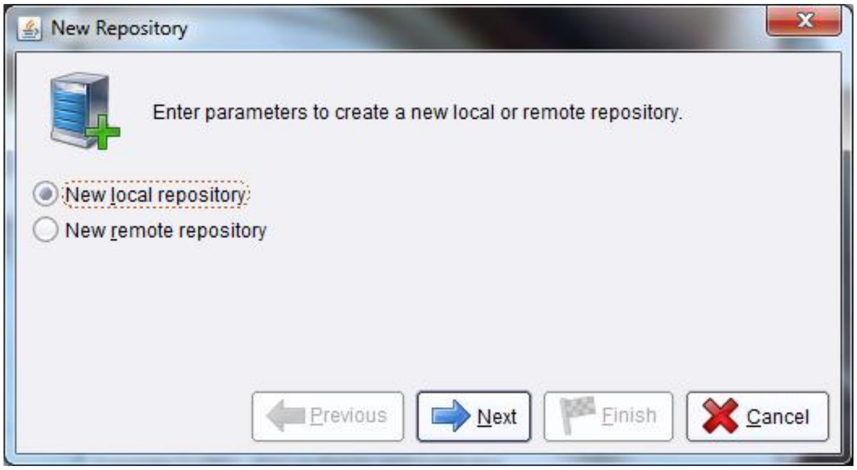
1) Launch the RapidMiner application. This can be done by double clicking your desktop icon or by finding it in your application menu. The first time RapidMiner is launched, you will get the message depicted in Figure 3-6. Click OK to set up a repository.



You do not have a RapidMiner repository defined. This might be because you are starting RapidMiner for the first time. In the next dialog, please select a directory to install a local repository and use this repository to store all your data and processes. Only for data and processes stored in the repository will you be able to use all RapidMiner features, foremost the new meta data transformation rules.

Figure 3-6. The prompt to create an initial data repository for RapidMiner to use.

2) For most purposes (and for all examples in this book), a local repository will be sufficient. Click OK to accept the default option as depicted in Figure 3-7.

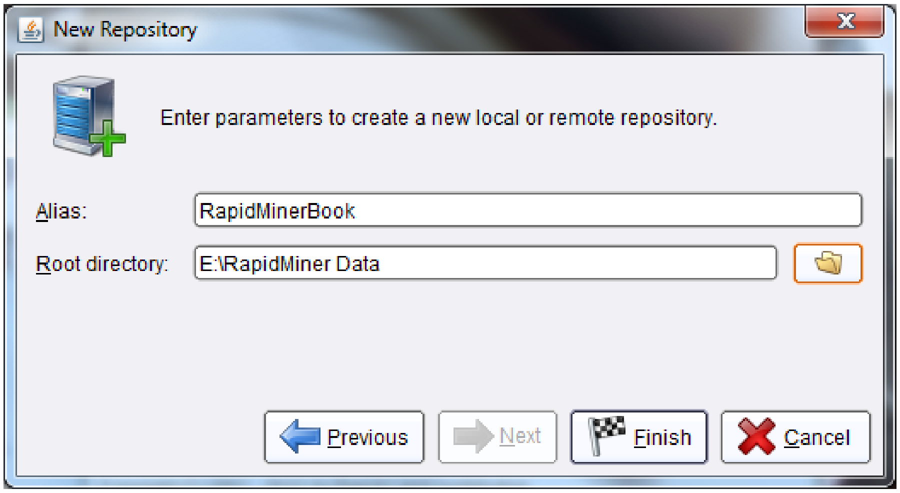


Enter parameters to create a new local or remote repository.

● New local repository
○ New remote repository

Figure 3-7. Setting up a local data repository.

3) In the example given in Figure 3-8, we have named our repository ‘RapidMinerBook, and pointed it to our data folder, RapidMiner Data, which is found on our E: drive. Use the folder icon to browse and find the folder or directory you created for storing your RapidMiner data sets. Then click Finish.



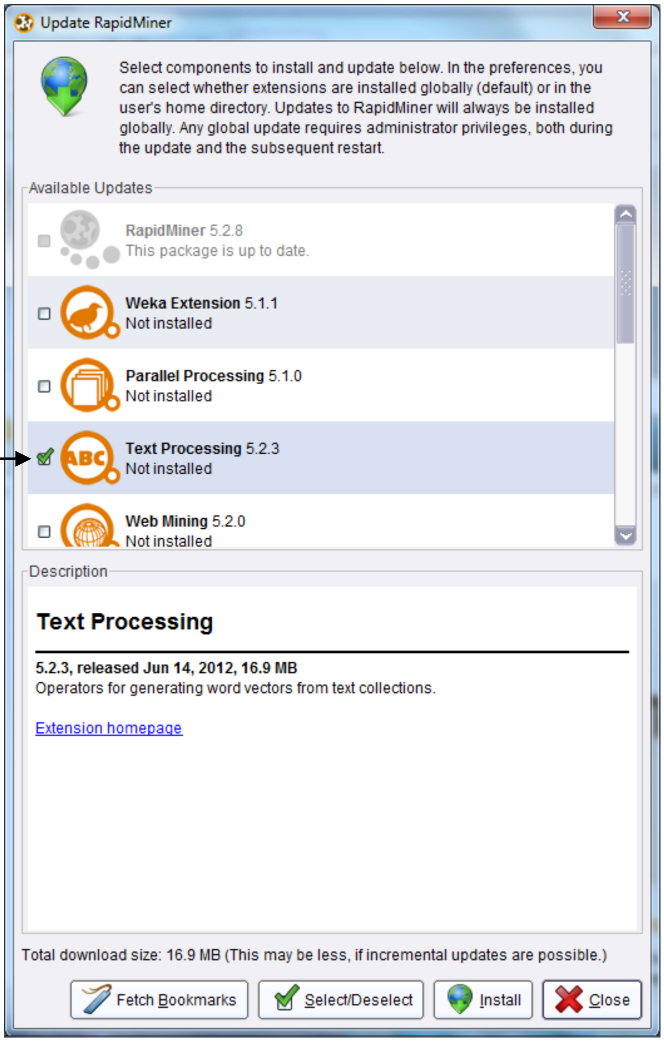
Enter parameters to create a new local or remote repository.

Alias: RapidMinerBook

Root directory: E:\RapidMiner Data

Figure 3-8. Setting the repository name and directory.

4) You may get a notice that updates are available. If this is the case, go ahead and accept the option to update, where you will be presented with a window similar to Figure 3-9. Take advantage of the opportunity to add in the Text Mining module (indicated by the black arrow), since Chapter 12 will deal with Text Mining. Double click the check box to add a green check mark indicating that you wish to install or update the module, then click Install.



Update RapidMiner

Select components to install and update below. In the preferences, you can select whether extensions are installed globally (default) or in the user's home directory. Updates to RapidMiner will always be installed globally. Any global update requires administrator privileges, both during the update and the subsequent restart.

Available Updates

□ RapidMiner 5.2.8
This package is up to date.

□ Weka Extension 5.1.1
Not installed

□ Parallel Processing 5.1.0
Not installed

☑ Text Processing 5.2.3
Not installed

□ Web Mining 5.2.0
Not installed

Description

Text Processing

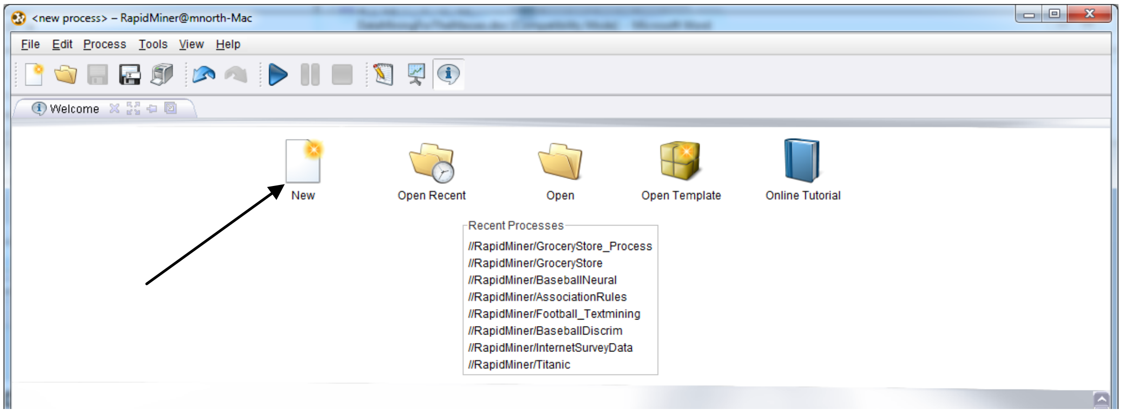
5.2.3, released Jun 14, 2012, 16.9 MB
Operators for generating word vectors from text collections.

Extension homepage

Total download size: 16.9 MB (This may be less, if incremental updates are possible.)

Figure 3-9. Installing updates and adding the Text Mining module.

5) Once the updates and installations are complete, RapidMiner will open and your window should look like Figure 3-10:



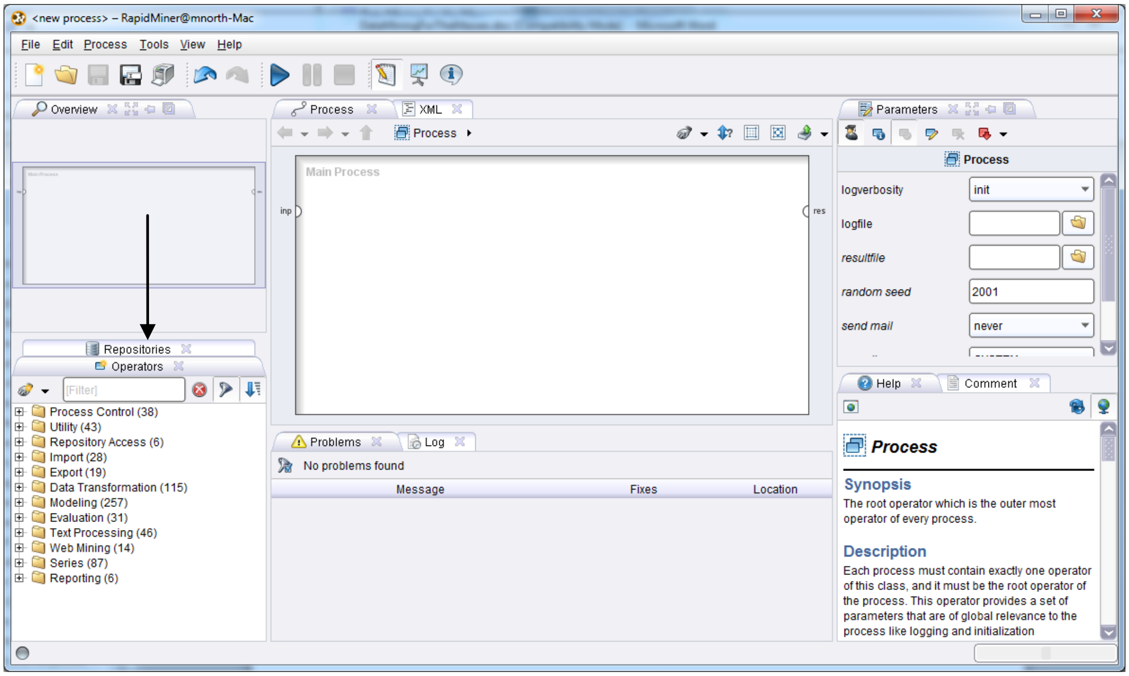
<new process> – RapidMiner@mnorth-Mac

Recent Processes

//RapidMiner/GroceryStore_Process
//RapidMiner/GroceryStore
//RapidMiner/BaseballNeural
//RapidMiner/AssociationRules
//RapidMiner/Football_Textmining
//RapidMiner/BaseballDiscrim
//RapidMiner/InternetSurveyData
//RapidMiner/Titanic

Figure 3-10. The RapidMiner start screen.

6) Next we will need to start a new data mining project in RapidMiner. To do this we click on the 'New' icon as indicated by the black arrow in Figure 3-10. The resulting window should look like Figure 3-11.



No problems found

Process

logverbosity: init
logfile:
resultfile:
random seed: 2001
send mail: never

Process

Synopsis

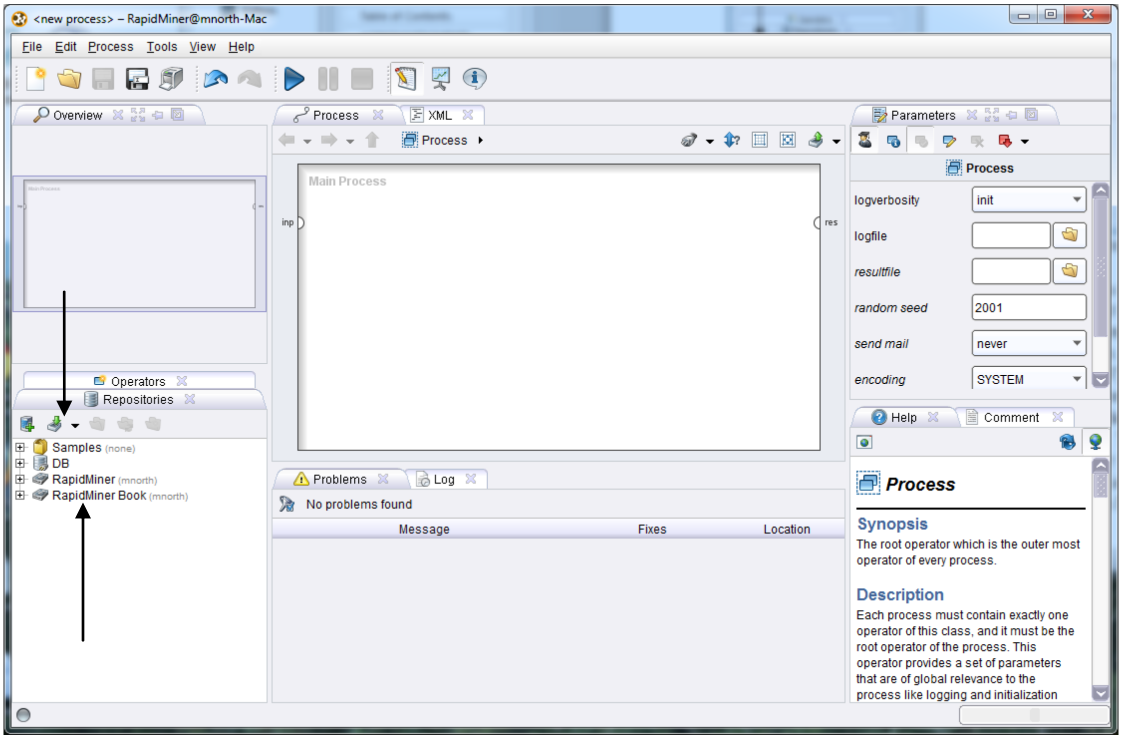
The root operator which is the outer most operator of every process.

Description

Each process must contain exactly one operator of this class, and it must be the root operator of the process. This operator provides a set of parameters that are of global relevance to the process like logging and initialization

Figure 3-11. Getting started with a new project in RapidMiner.

7) Within RapidMiner there are two main areas that hold useful tools: **Repositories** and **Operators**. These are accessed by the tabs indicated by the black arrow in Figure 3-11. The Repositories area is the place where you will connect to each data set you wish to mine. The Operators area is where all data mining tools are located. These are used to build models and otherwise manipulate data sets. Click on Repositories. You will find that the initial repository we created upon our first launch of the RapidMiner software is present in the list.



Process

logverbosity: init

logfile:

resultfile:

random seed: 2001

send mail: never

encoding: SYSTEM

No problems found

Process

Synopsis

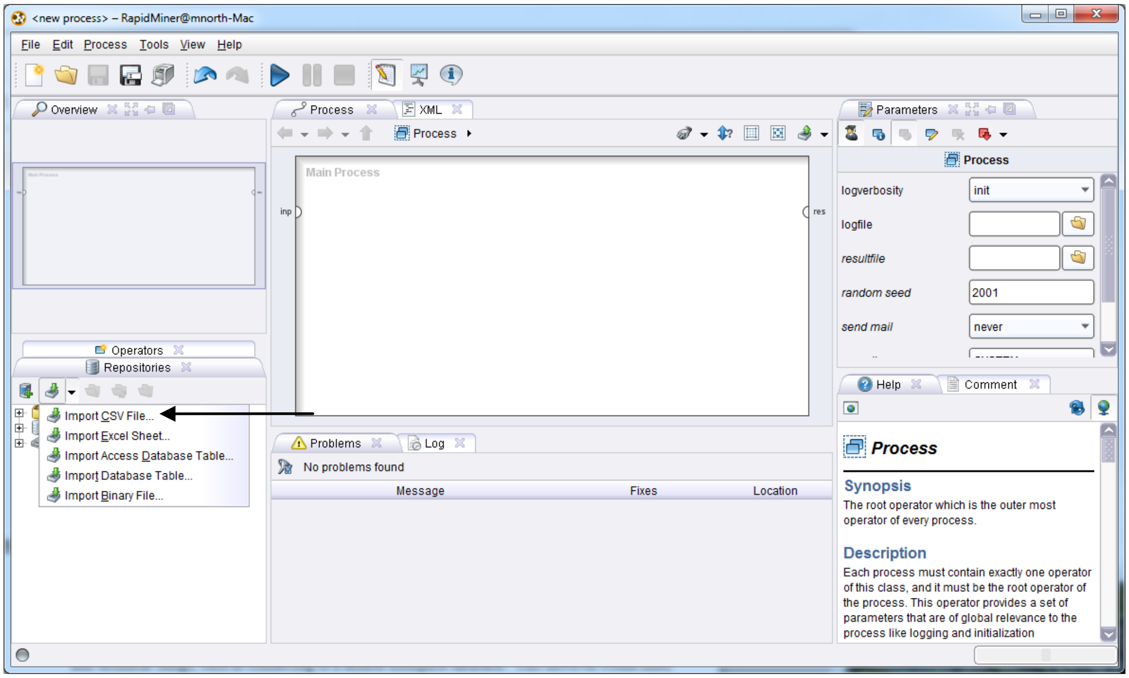
The root operator which is the outer most operator of every process.

Description

Each process must contain exactly one operator of this class, and it must be the root operator of the process. This operator provides a set of parameters that are of global relevance to the process like logging and initialization

Figure 3-12. Adding a data set to a repository in RapidMiner.

8) Because the focus of this book is to introduce data mining to the broadest possible audience, we will not use all of the tools available in RapidMiner. At this point, we could do a number of complicated and technical things, such as connecting to a remote enterprise database. This however would likely be overwhelming and inaccessible to many readers. For the purposes of this text, we will therefore only be connecting to **comma separate values (CSV)** files. You should know that most data mining projects incorporate extremely large data sets encompassing dozens of attributes and thousands or even millions of observations. We will use smaller data sets in this text, but the foundational concepts illustrated are the same for large or small data. The Chapter 3 data set downloaded from the companion web site is very small, comprised of only 15 attributes and 11 observations. Our next step is to connect to this data set. Click on the Import icon, which is the second icon from the left in the Repositories area, as indicated by the black arrow in Figure 3-12.



Process

logverbosity: init

logfile:

resultfile:

random seed: 2001

send mail: never

No problems found

Message	Fixes	Location

Process

Synopsis

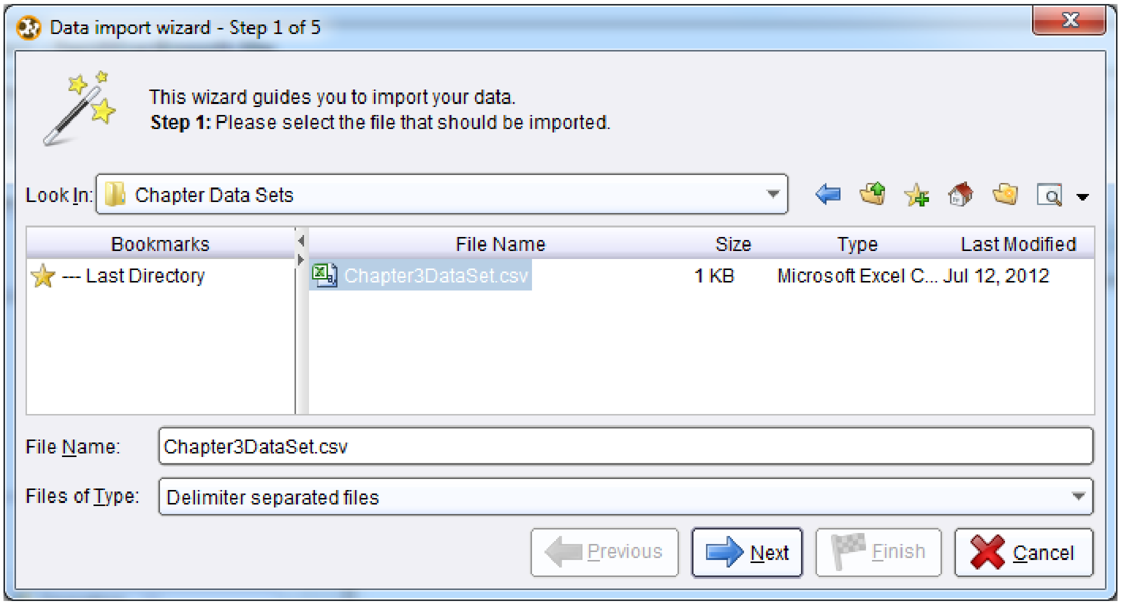
The root operator which is the outer most operator of every process.

Description

Each process must contain exactly one operator of this class, and it must be the root operator of the process. This operator provides a set of parameters that are of global relevance to the process like logging and initialization

Figure 3-13. Importing a CSV file.

9) You will see by the black arrow in Figure 3-13 that you can import from a number of different data sources. Note that by importing, you are bringing your data into a RapidMiner file, rather than working with data that are already stored elsewhere. If your data set is extremely large, it may take some time to import the data, and you should be mindful of disk space that is available to you. As data sets grow, you may be better off using the first (leftmost) icon to set up a remote repository in order to work with data already stored in other areas. As previously explained, all examples in this text will be conducted by importing CSV files that are small enough to work with quickly and easily. Click on the Import CSV File option.



Data import wizard - Step 1 of 5

This wizard guides you to import your data.
Step 1: Please select the file that should be imported.

Look In: Chapter Data Sets

Bookmarks
--- Last Directory

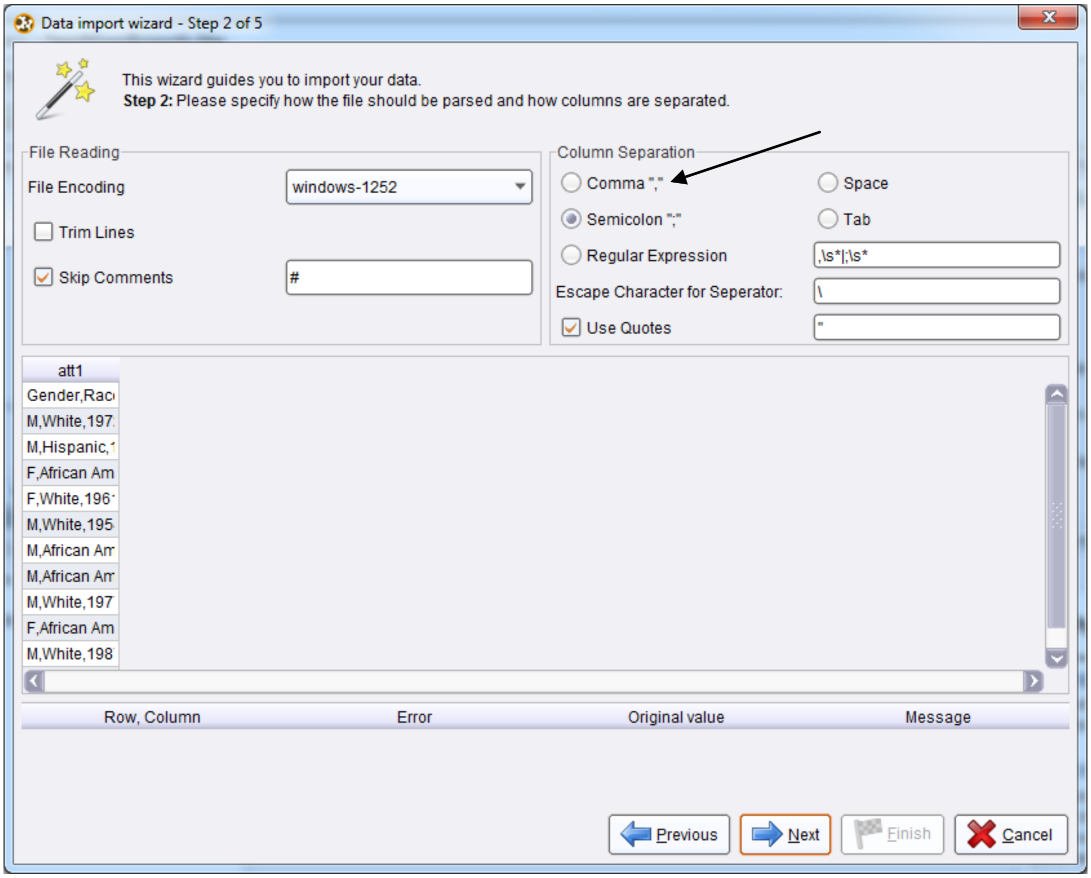
File Name	Size	Type	Last Modified
Chapter3DataSet.csv	1 KB	Microsoft Excel C...	Jul 12, 2012

File Name: Chapter3DataSet.csv

Files of Type: Delimiter separated files

Figure 3-14. Locating the data set to import.

10) When the data import wizard opens, navigate to the folder where your data set is stored and select the file. In this example, only one file is visible: the Chapter 3 data set downloaded from the companion web site. Click Next.



Data import wizard - Step 2 of 5

This wizard guides you to import your data.
Step 2: Please specify how the file should be parsed and how columns are separated.

File Reading

File Encoding: windows-1252

☐ Trim Lines

☑ Skip Comments #

Column Separation

○ Comma ","　○ Space

● Semicolon ";"　○ Tab

○ Regular Expression ,\s*|;\s*

Escape Character for Seperator: \

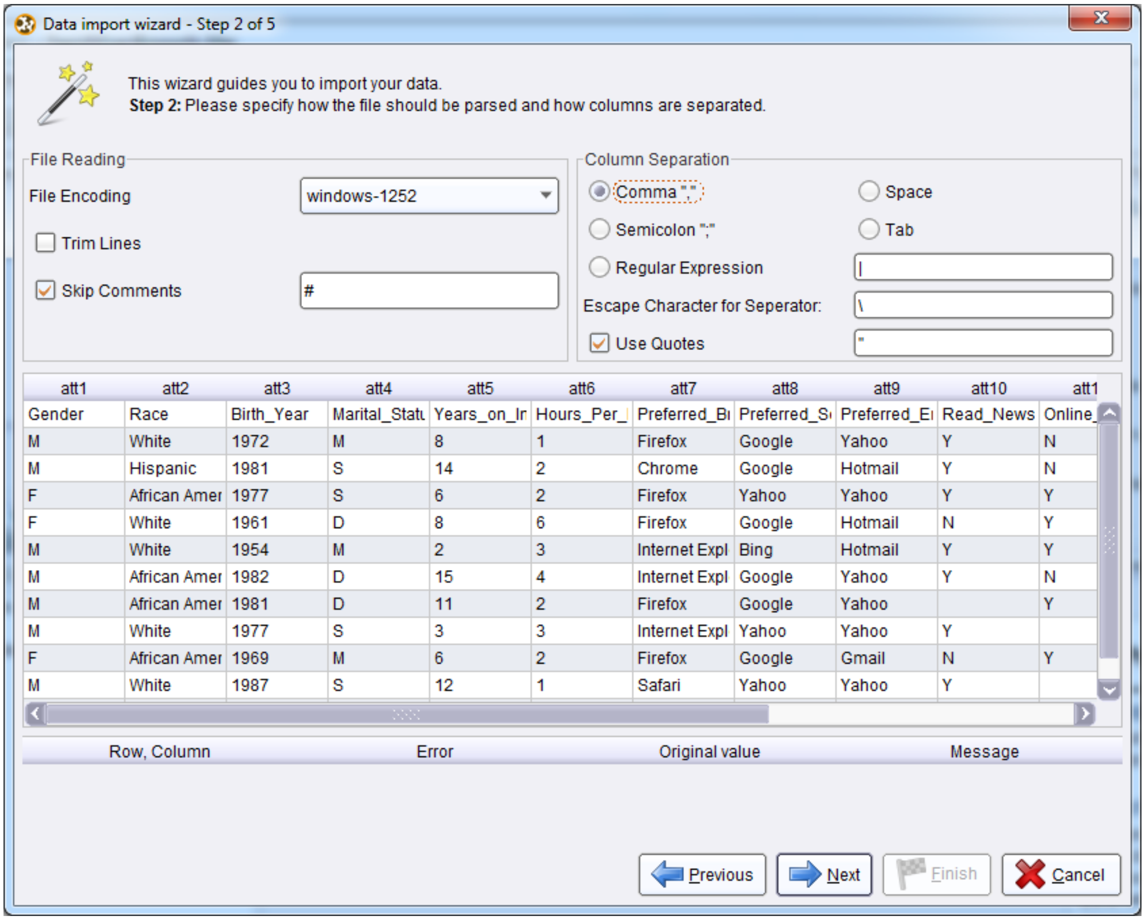
☑ Use Quotes "

att1
Gender,Rac
M,White,197
M,Hispanic,'
F,African Am
F,White,196
M,White,195
M,African Ar
M,African Ar
M,White,197
F,African Am
M,White,198

Row, Column	Error	Original value	Message

Figure 3-15. Configuring attribute separation.

11) By default, RapidMiner looks for semicolons as attribute separators in our data. We must change the column separation delimiter to be Comma, in order to be able to see each attribute separated correctly. **Note:** If your data naturally contain commas, then you should be careful as you are collecting or collating your data to use a delimiter that does not naturally occur in the data. A semicolon or a pipe (|) symbol can often help you avoid unintended column separation.


Data import wizard - Step 2 of 5

This wizard guides you to import your data.
Step 2: Please specify how the file should be parsed and how columns are separated.

File Reading

File Encoding: windows-1252

☐ Trim Lines

☑ Skip Comments: #

Column Separation

● Comma "," ○ Space

○ Semicolon ";" ○ Tab

○ Regular Expression: |

Escape Character for Seperator: \

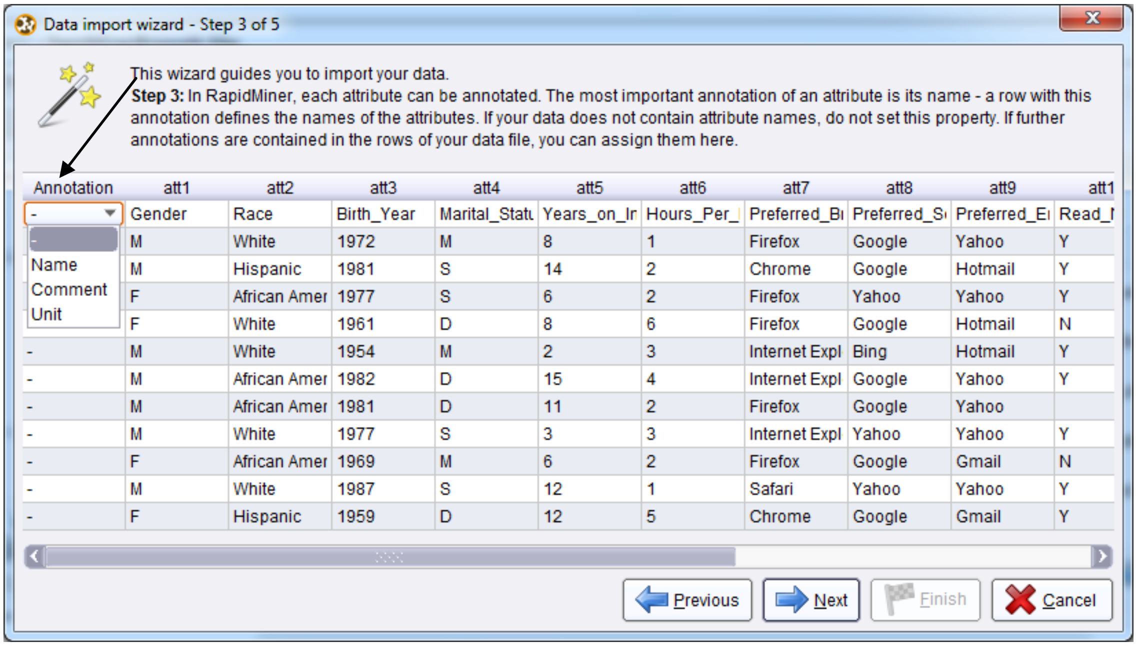
☑ Use Quotes: "

att1	att2	att3	att4	att5	att6	att7	att8	att9	att10	att1
Gender	Race	Birth_Year	Marital_Statu	Years_on_In	Hours_Per_	Preferred_Br	Preferred_S	Preferred_Em	Read_News	Online_
M	White	1972	M	8	1	Firefox	Google	Yahoo	Y	N
M	Hispanic	1981	S	14	2	Chrome	Google	Hotmail	Y	N
F	African Amer	1977	S	6	2	Firefox	Yahoo	Yahoo	Y	Y
F	White	1961	D	8	6	Firefox	Google	Hotmail	N	Y
M	White	1954	M	2	3	Internet Expl	Bing	Hotmail	Y	Y
M	African Amer	1982	D	15	4	Internet Expl	Google	Yahoo	Y	N
M	African Amer	1981	D	11	2	Firefox	Google	Yahoo		Y
M	White	1977	S	3	3	Internet Expl	Yahoo	Yahoo	Y	
F	African Amer	1969	M	6	2	Firefox	Google	Gmail	N	Y
M	White	1987	S	12	1	Safari	Yahoo	Yahoo	Y	

Row, Column	Error	Original value	Message

Figure 3-16. A preview of attributes separated into columns with the Comma option selected.

12) Once the preview shows columns for each attribute, click Next. Note that RapidMiner has treated our attribute names as if they are our first row of data, or in other words, our first observation. To fix this, click the Annotation dropdown box next to this row and set it to **Name**, as indicated in Figure 3-17. With the attribute names designated correctly, click Next.

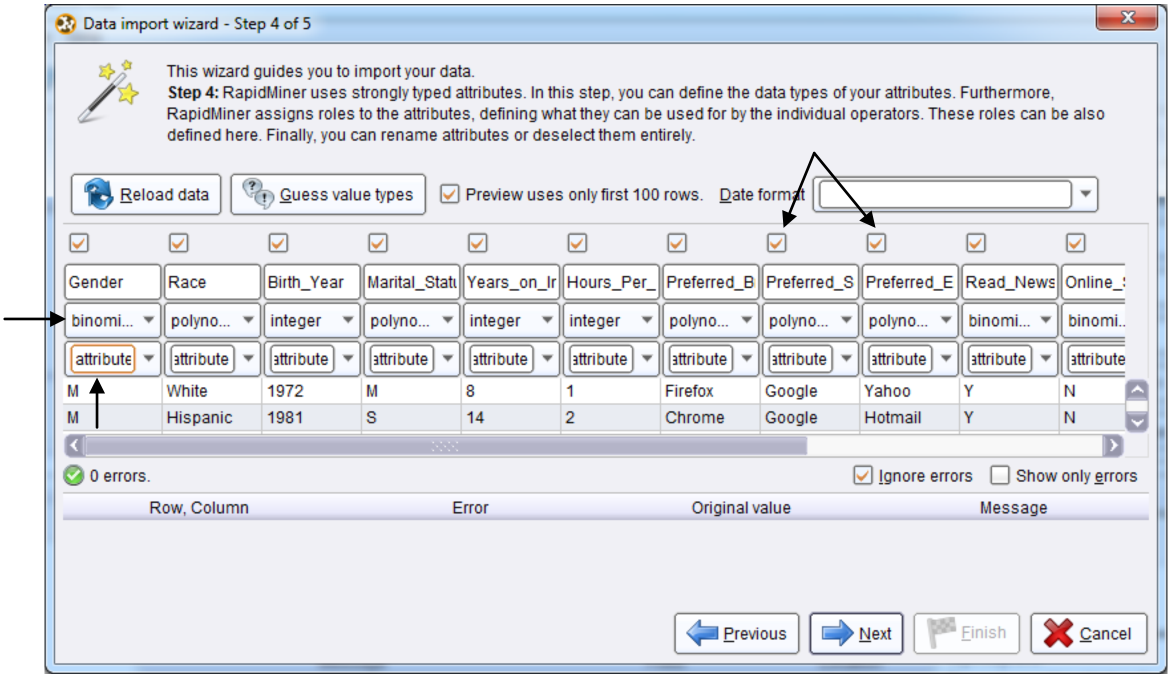


This wizard guides you to import your data.
Step 3: In RapidMiner, each attribute can be annotated. The most important annotation of an attribute is its name - a row with this annotation defines the names of the attributes. If your data does not contain attribute names, do not set this property. If further annotations are contained in the rows of your data file, you can assign them here.

Annotation	att1	att2	att3	att4	att5	att6	att7	att8	att9	att1
-	Gender	Race	Birth_Year	Marital_Statu	Years_on_Ir	Hours_Per_	Preferred_Bı	Preferred_Sı	Preferred_Eı	Read_I
	M	White	1972	M	8	1	Firefox	Google	Yahoo	Y
	M	Hispanic	1981	S	14	2	Chrome	Google	Hotmail	Y
	F	African Amer	1977	S	6	2	Firefox	Yahoo	Yahoo	Y
	F	White	1961	D	8	6	Firefox	Google	Hotmail	N
-	M	White	1954	M	2	3	Internet Expl	Bing	Hotmail	Y
-	M	African Amer	1982	D	15	4	Internet Expl	Google	Yahoo	Y
-	M	African Amer	1981	D	11	2	Firefox	Google	Yahoo	
-	M	White	1977	S	3	3	Internet Expl	Yahoo	Yahoo	Y
-	F	African Amer	1969	M	6	2	Firefox	Google	Gmail	N
-	M	White	1987	S	12	1	Safari	Yahoo	Yahoo	Y
-	F	Hispanic	1959	D	12	5	Chrome	Google	Gmail	Y

Figure 3-17. Setting the attribute names.

13) In step 4 of the data import wizard, RapidMiner will take its best guess at a **data type** for each attribute. The data type is the kind of data an attribute holds, such as numeric, text or date. These can be changed in this screen, but for our purposes in Chapter 3, we will accept the defaults. Just below each attribute's data type, RapidMiner also indicates a **Role** for each attribute to play. By default, all columns are imported simply with the role of 'attribute', however we can change these here if we know that one attribute is going to play a specific role in a data mining model that we will create. Since roles can be set within RapidMiner's main process window when building data mining models, we will accept the default of 'attribute' whenever we import data sets in exercises in this text. Also, you may note that the check boxes above each attribute in this window allow you to *not* import some of the attributes if you don't want to. This is accomplished by simply clearing the checkbox. Again, attributes can be excluded from models later, so for the purposes of this text, we will always include all attributes when importing data. All of these functions are indicated by the black arrows in Figure 3-18. Go ahead and accept these defaults as they stand and click Next.



Data import wizard - Step 4 of 5

This wizard guides you to import your data.
Step 4: RapidMiner uses strongly typed attributes. In this step, you can define the data types of your attributes. Furthermore, RapidMiner assigns roles to the attributes, defining what they can be used for by the individual operators. These roles can be also defined here. Finally, you can rename attributes or deselect them entirely.

Reload data | Guess value types | ☑ Preview uses only first 100 rows. | Date format

☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Gender	Race	Birth_Year	Marital_Statu	Years_on_Ir	Hours_Per_	Preferred_B	Preferred_S	Preferred_E	Read_News	Online_
binomi...	polyno...	integer	polyno...	integer	integer	polyno...	polyno...	polyno...	binomi...	binomi..
attribute	attribute	attribute	attribute	attribute	attribute	attribute	attribute	attribute	attribute	attribute
M	White	1972	M	8	1	Firefox	Google	Yahoo	Y	N
M	Hispanic	1981	S	14	2	Chrome	Google	Hotmail	Y	N

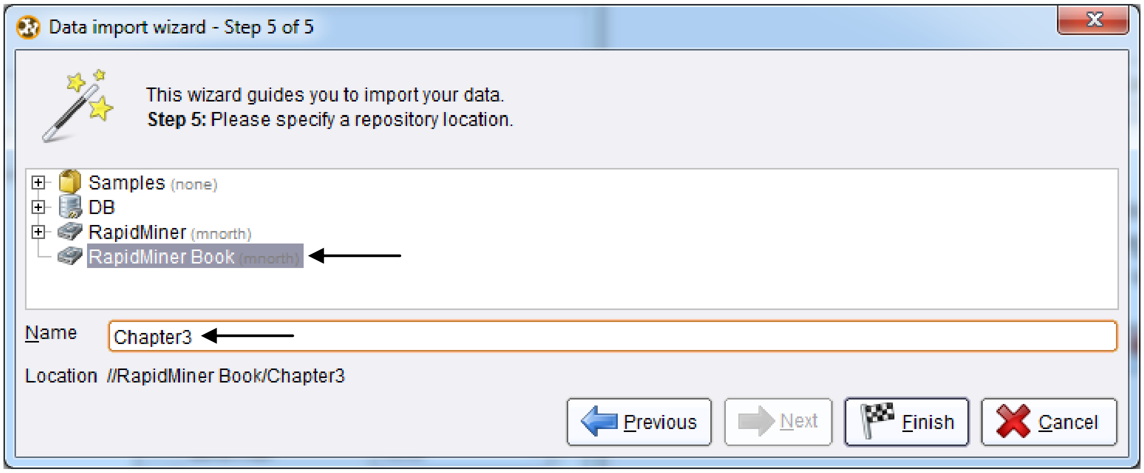
0 errors. ☑ Ignore errors ☐ Show only errors

Row, Column	Error	Original value	Message

Previous | Next | Finish | Cancel

Figure 3-18. Setting data types, roles and import attributes.

14) The final step is to choose a repository to store the data set in, and to give the data set a name within RapidMiner. In Figure 3-19, we have chosen to store the data set in the RapidMiner Book repository, and given it the name Chapter3. Once we click Finish, this data set will become available to us for any type of data mining process we would like to build upon it.



Data import wizard - Step 5 of 5

This wizard guides you to import your data.
Step 5: Please specify a repository location.

- Samples (none)
- DB
- RapidMiner (mnorth)
- RapidMiner Book (mnorth)

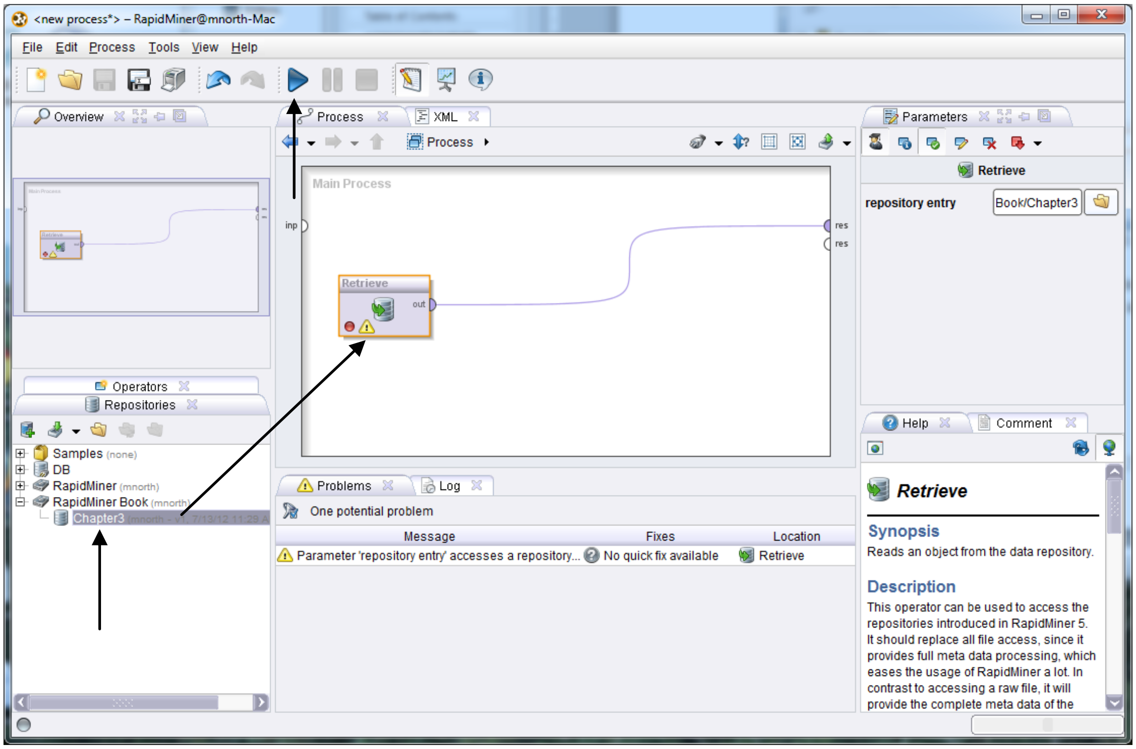
Name Chapter3

Location //RapidMiner Book/Chapter3

Previous | Next | Finish | Cancel

Figure 3-19. Selecting the repository and setting a data set name for our imported CSV file.

15) We can now see that the data set is available for use in RapidMiner. To begin using it in a RapidMiner data mining process, simply drag the data set and drop it in the Main Process window, as has been done in Figure 3-20.



Retrieve

repository entry: Book/Chapter3

One potential problem

Message	Fixes	Location
Parameter 'repository entry' accesses a repository...	No quick fix available	Retrieve

Retrieve

Synopsis

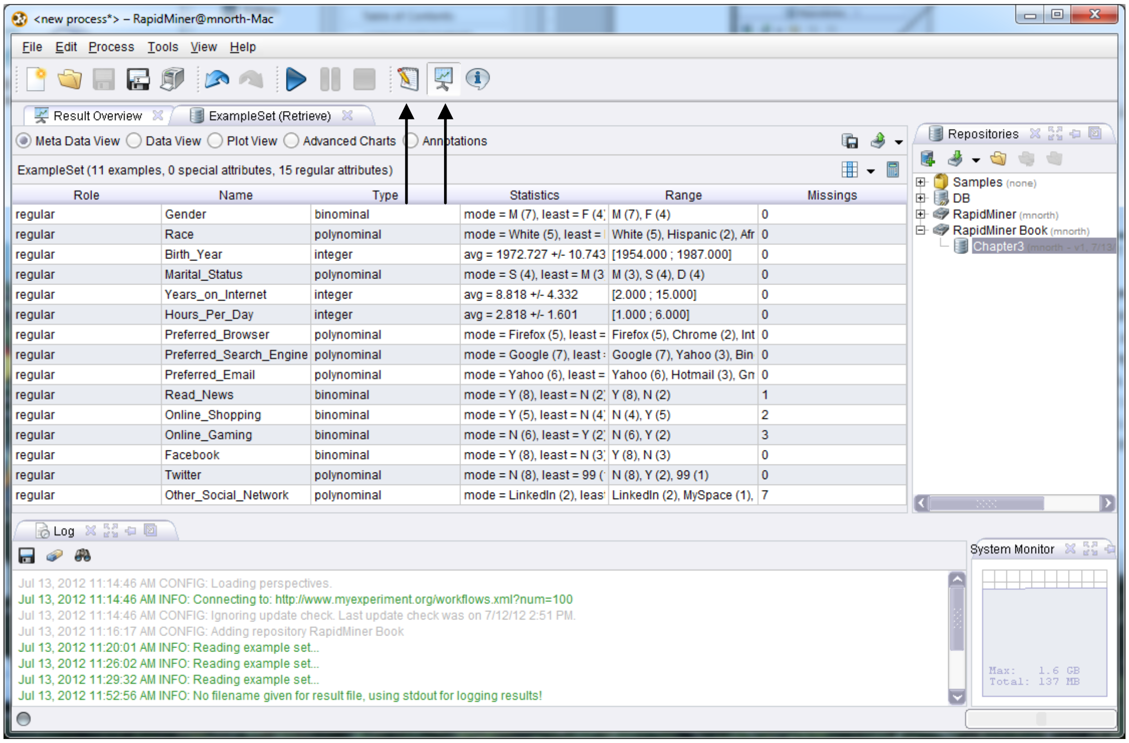
Reads an object from the data repository.

Description

This operator can be used to access the repositories introduced in RapidMiner 5. It should replace all file access, since it provides full meta data processing, which eases the usage of RapidMiner a lot. In contrast to accessing a raw file, it will provide the complete meta data of the

Figure 3-20. Adding a data set to a process in RapidMiner.

16) Each rectangle in a process in RapidMiner is an **operator**. The Retrieve operator simply gets a data set and makes it available for use. The small half-circles on the sides of the operator, and of the Main Process window, are called **ports**. In Figure 3-20, an output (*out*) port from our data set's Retrieve operator is connected to a result set (*res*) port via a **spline**. The splines, combined with the operators connected by them, constitute a data mining **stream**. To run a data mining stream and see the results, click the blue, triangular Play button in the toolbar at the top of the RapidMiner window. This will change your view from **Design Perspective**, which is the view pictured in Figure 3-20 where you can change your data mining stream, to **Results Perspective**, which shows your stream's results, as pictured in Figure 3-21. When you hit the Play button, you may be prompted to save your process, and you are encouraged to do so. RapidMiner may also ask you if you wish to overwrite a saved process each time it is run, and you can select your preference on this prompt as well.



ExampleSet (11 examples, 0 special attributes, 15 regular attributes)

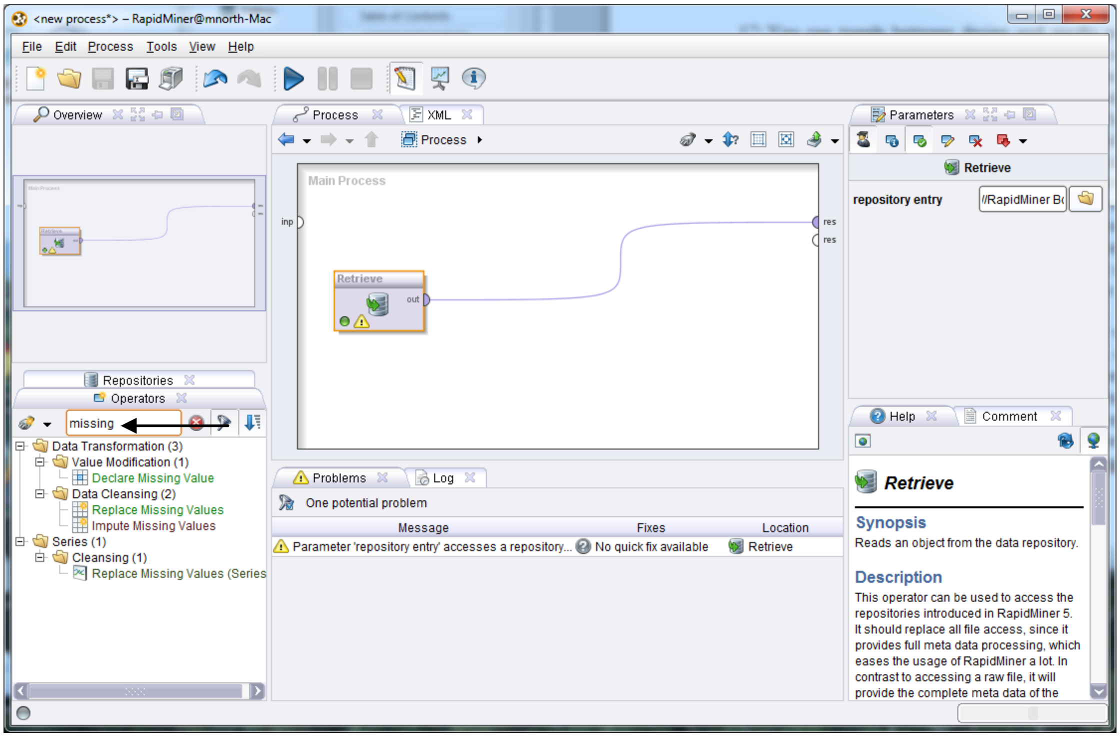
Role	Name	Type	Statistics	Range	Missings
regular	Gender	binominal	mode = M (7), least = F (4	M (7), F (4)	0
regular	Race	polynominal	mode = White (5), least =	White (5), Hispanic (2), Afr	0
regular	Birth_Year	integer	avg = 1972.727 +/- 10.743	[1954.000 ; 1987.000]	0
regular	Marital_Status	polynominal	mode = S (4), least = M (3	M (3), S (4), D (4)	0
regular	Years_on_Internet	integer	avg = 8.818 +/- 4.332	[2.000 ; 15.000]	0
regular	Hours_Per_Day	integer	avg = 2.818 +/- 1.601	[1.000 ; 6.000]	0
regular	Preferred_Browser	polynominal	mode = Firefox (5), least =	Firefox (5), Chrome (2), Int	0
regular	Preferred_Search_Engine	polynominal	mode = Google (7), least	Google (7), Yahoo (3), Bin	0
regular	Preferred_Email	polynominal	mode = Yahoo (6), least =	Yahoo (6), Hotmail (3), Gn	0
regular	Read_News	binominal	mode = Y (8), least = N (2	Y (8), N (2)	1
regular	Online_Shopping	binominal	mode = Y (5), least = N (4	N (4), Y (5)	2
regular	Online_Gaming	binominal	mode = N (6), least = Y (2	N (6), Y (2)	3
regular	Facebook	binominal	mode = Y (8), least = N (3	Y (8), N (3)	0
regular	Twitter	polynominal	mode = N (8), least = 99 (	N (8), Y (2), 99 (1)	0
regular	Other_Social_Network	polynominal	mode = LinkedIn (2), leas	LinkedIn (2), MySpace (1),	7

```
Jul 13, 2012 11:14:46 AM CONFIG: Loading perspectives.
Jul 13, 2012 11:14:46 AM INFO: Connecting to: http://www.myexperiment.org/workflows.xml?num=100
Jul 13, 2012 11:14:46 AM CONFIG: Ignoring update check. Last update check was on 7/12/12 2:51 PM.
Jul 13, 2012 11:16:17 AM CONFIG: Adding repository RapidMiner Book
Jul 13, 2012 11:20:01 AM INFO: Reading example set...
Jul 13, 2012 11:26:02 AM INFO: Reading example set...
Jul 13, 2012 11:29:32 AM INFO: Reading example set...
Jul 13, 2012 11:52:56 AM INFO: No filename given for result file, using stdout for logging results!
```

Figure 3-21. Results perspective for the Chapter3 data set.

17) You can toggle between design and results perspectives using the two icons indicated by the black arrows in Figure 3-21. As you can see, there is a rich set of information in results perspective. In the **meta data** view, basic descriptive statistics are given. It is here that we can also get a sense for the number of observations that have missing values in each attribute of the data set. The columns in meta data view can be stretched to make their contents more readable. This is accomplished by hovering your mouse over the faint vertical gray bars between each column, then clicking and dragging to make them wider. The information presented here can be very helpful in deciding where missing data are located, and what to do about it. Take for example the Online_Gaming attribute. The results perspective shows us that we have six 'N' responses in that attribute, two 'Y' responses, and three missing. We could use the **mode**, or most common response to replace the missing values. This of course assumes that the most common response is accurate for all observations, and this may not be accurate. As data miners, we must be responsible for thinking about each change we make in our data, and whether or not we threaten the integrity of our data by making that change. In some instances the consequences could be drastic. Consider, for instance, if the mode for an attribute of Felony_Conviction were 'Y'. Would we really want to convert all missing values in this attribute to 'Y' simply because that is the mode in our data set? Probably not; the

implications about the persons represented in each observation of our data set would be unfair and misrepresentative. Thus, we will change the missing values in the current example to illustrate how to handle missing values in RapidMiner, recognizing that what we are about to do won't always be the right way to handle missing data. In order to have RapidMiner handle the change from missing to 'N' for the three observations in our Online_Gaming variable, click the design perspective icon.



repository entry: //RapidMiner B(

missing

One potential problem

Message	Fixes	Location
Parameter 'repository entry' accesses a repository...	No quick fix available	Retrieve

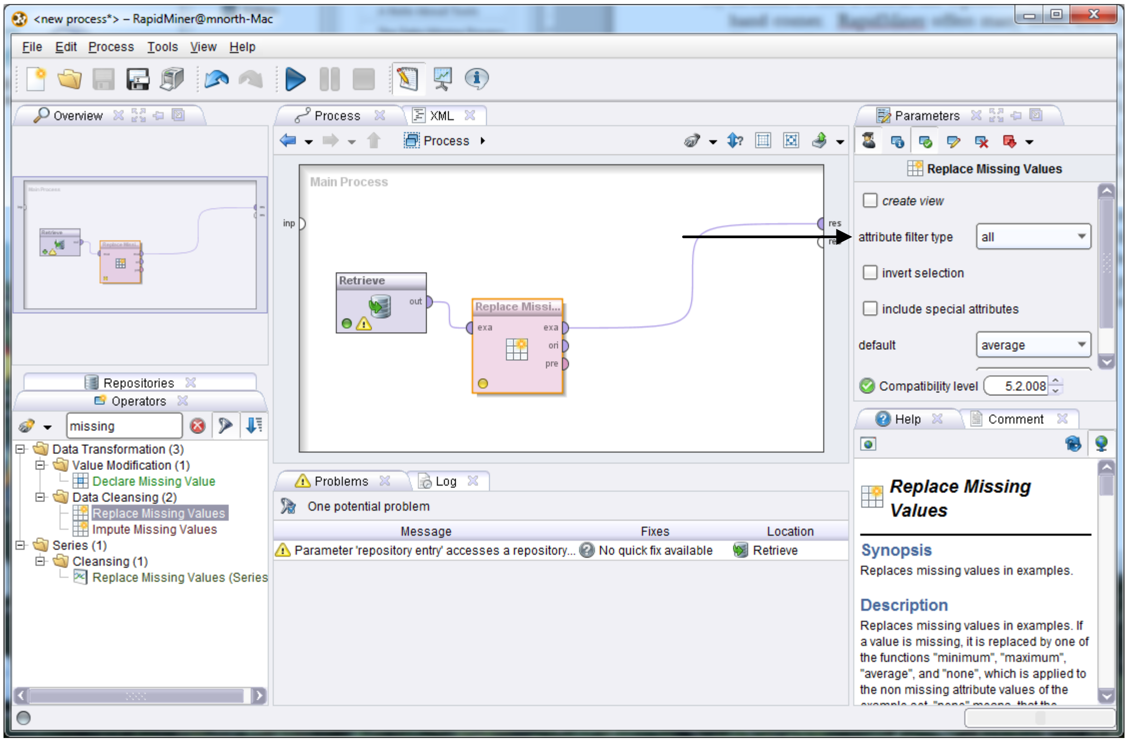
Reads an object from the data repository.

This operator can be used to access the repositories introduced in RapidMiner 5. It should replace all file access, since it provides full meta data processing, which eases the usage of RapidMiner a lot. In contrast to accessing a raw file, it will provide the complete meta data of the

Figure 3-22. Finding an operator to handle missing values.

18) In order to find a tool in the Operators area, you can navigate through the folder tree in the lower left hand corner. RapidMiner offers many tools, and sometimes, finding the one you want can be tricky. There is a handy search box, indicated by the black arrow in Figure 3-22, that allows you to type in key words to find tools that might do what you need. Type the word 'missing' into this box, and you will see that RapidMiner automatically searches for tools with this word in their name. We want to replace missing values, and we can see that within the Data Transformation tool area, inside a sub-area called Value Modification, there is an operator called Replace Missing Values. Let's add this operator to our stream. Click and hold on the operator name, and drag it up to your spline. When you point your mouse cursor on the spline, the spline will turn slightly bold, indicating that when you let go of your mouse button, the operator will be connected into the stream. If you let go and the Replace Missing Values operator fails to connect into your stream, you can reconfigure

your splines manually. Simply click on the *out* port in your Retrieve operator, and then click on the *exa* port on the Replace Missing Values operator. *Exa* stands for example set, and remember that 'examples' is the word RapidMiner uses for observations in a data set. Be sure the *exa* port from the Replace Missing Values operator is connected to your result set (*res*) port so that when you run your process, you will have output. Your model should now look similar to Figure 3-23.



Replace Missing Values

☐ create view

attribute filter type: all

☐ invert selection

☐ include special attributes

default: average

Compatibility level 5.2.008

One potential problem

Message	Fixes	Location
Parameter 'repository entry' accesses a repository...	No quick fix available	Retrieve

Replace Missing Values

Synopsis

Replaces missing values in examples.

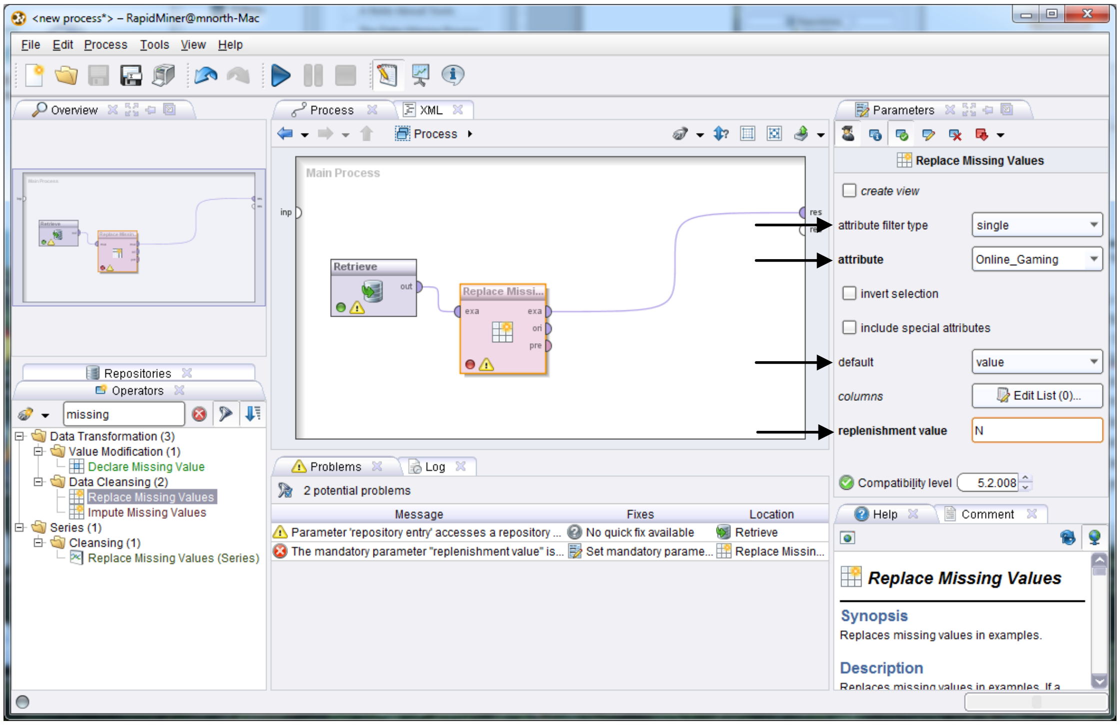
Description

Replaces missing values in examples. If a value is missing, it is replaced by one of the functions "minimum", "maximum", "average", and "none", which is applied to the non missing attribute values of the example set. "none" means, that the

Figure 3-23. Adding a missing value operator to the stream.

19) When an operator is selected in RapidMiner, it has an orange rectangle around it. This will also enable you to modify that operator's **parameters**, or properties. The Parameters pane is located on the right side of the RapidMiner window, as indicated by the black arrow in Figure 3-23. For this exercise, we have decided to change all missing values in the Online_Gaming attribute to be 'N', since this is the most common response in that attribute. To do this, change the 'attribute filter type' to 'single', and you will see that a dropdown box appears, allowing you to choose the Online_Gaming attribute as the target for modification. Next, expand the 'default' dropdown box, and select 'value', which will cause a 'replenishment value' box to appear. Type the replacement value 'N' in this box. Note that you may need to expand your RapidMiner window, or use the vertical scroll bar on the left of the Parameters pane in order to see all options, as the options change based on what you have selected. When you are finished, your parameters should look like the

ones in Figure 3-24. Parameter settings that were changed are highlighted with black arrows.



missing

Replace Missing Values

□ create view

attribute filter type: single

attribute: Online_Gaming

□ invert selection

□ include special attributes

default: value

columns: Edit List (0)...

replenishment value: N

Compatibility level: 5.2.008

2 potential problems

Message	Fixes	Location
Parameter 'repository entry' accesses a repository ...	No quick fix available	Retrieve
The mandatory parameter "replenishment value" is...	Set mandatory parame...	Replace Missin...

Replace Missing Values

Synopsis

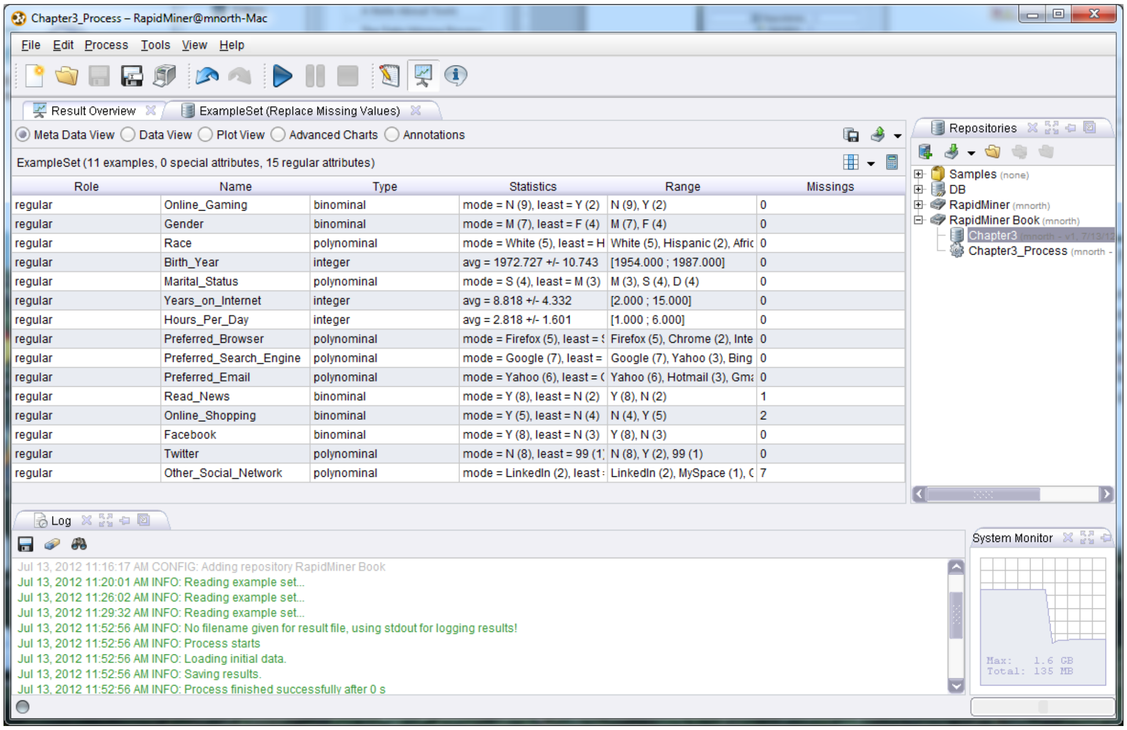
Replaces missing values in examples.

Description

Replaces missing values in examples. If a

Figure 3-24. Missing value parameters.

20) You should understand that there are many other options available to you in the parameters pane. We will not explore all of them here, but feel free to experiment with them. For example, instead of changing a single attribute at a time, you could change a subset of the attributes in your data set. You will learn much about the flexibility and power of RapidMiner by trying out different tools and features. When you have your parameter set, click the play button. This will run your process and switch you to results perspective once again. Your results should look like Figure 3-25.


Chapter3_Process – RapidMiner@mnorth-Mac
File Edit Process Tools View Help
Result Overview
ExampleSet (Replace Missing Values)
● Meta Data View ○ Data View ○ Plot View ○ Advanced Charts ○ Annotations
Repositories
Samples (none)
DB
RapidMiner (mnorth)
RapidMiner Book (mnorth)
Chapter3
Chapter3_Process
Log
System Monitor
Max: 1.6 GB
Total: 135 MB

ExampleSet (11 examples, 0 special attributes, 15 regular attributes)

Role	Name	Type	Statistics	Range	Missings
regular	Online_Gaming	binominal	mode = N (9), least = Y (2)	N (9), Y (2)	0
regular	Gender	binominal	mode = M (7), least = F (4)	M (7), F (4)	0
regular	Race	polynominal	mode = White (5), least = H	White (5), Hispanic (2), Afric	0
regular	Birth_Year	integer	avg = 1972.727 +/- 10.743	[1954.000 ; 1987.000]	0
regular	Marital_Status	polynominal	mode = S (4), least = M (3)	M (3), S (4), D (4)	0
regular	Years_on_Internet	integer	avg = 8.818 +/- 4.332	[2.000 ; 15.000]	0
regular	Hours_Per_Day	integer	avg = 2.818 +/- 1.601	[1.000 ; 6.000]	0
regular	Preferred_Browser	polynominal	mode = Firefox (5), least = (	Firefox (5), Chrome (2), Inte	0
regular	Preferred_Search_Engine	polynominal	mode = Google (7), least =	Google (7), Yahoo (3), Bing	0
regular	Preferred_Email	polynominal	mode = Yahoo (6), least = (	Yahoo (6), Hotmail (3), Gmi	0
regular	Read_News	binominal	mode = Y (8), least = N (2)	Y (8), N (2)	1
regular	Online_Shopping	binominal	mode = Y (5), least = N (4)	N (4), Y (5)	2
regular	Facebook	binominal	mode = Y (8), least = N (3)	Y (8), N (3)	0
regular	Twitter	polynominal	mode = N (8), least = 99 (1)	N (8), Y (2), 99 (1)	0
regular	Other_Social_Network	polynominal	mode = LinkedIn (2), least :	LinkedIn (2), MySpace (1), (	7

```
Jul 13, 2012 11:16:17 AM CONFIG: Adding repository RapidMiner Book
Jul 13, 2012 11:20:01 AM INFO: Reading example set...
Jul 13, 2012 11:26:02 AM INFO: Reading example set...
Jul 13, 2012 11:29:32 AM INFO: Reading example set...
Jul 13, 2012 11:52:56 AM INFO: No filename given for result file, using stdout for logging results!
Jul 13, 2012 11:52:56 AM INFO: Process starts
Jul 13, 2012 11:52:56 AM INFO: Loading initial data.
Jul 13, 2012 11:52:56 AM INFO: Saving results.
Jul 13, 2012 11:52:56 AM INFO: Process finished successfully after 0 s
```

Figure 3-25. Results of changing missing data.

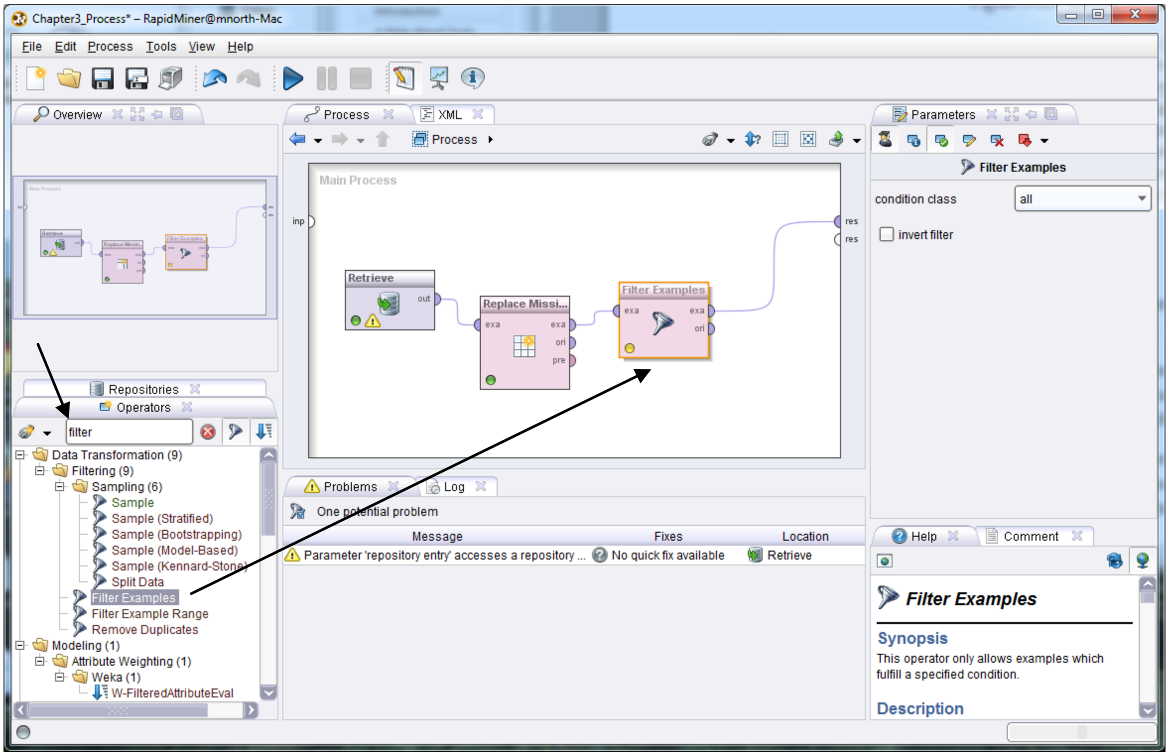
21) You can see now that the Online_Gaming attribute has been moved to the top of our list, and that there are zero missing values. Click on the Data View radio button, above and to the left hand side of the attribute list to see your data in a spreadsheet-type view. You will see that the Online_Gaming variable is now populated with only 'Y' and 'N' values. We have successfully replaced all missing values in that attribute. While in Data View, take note of how missing values are annotated in other variables, Online_Shopping for example. A question mark (?) denotes a missing value in an observation. Suppose that for this variable, we do not wish to replace the null values with the mode, but rather, that we wish to remove those observations from our data set prior to mining it. This is accomplished through data reduction.

DATA REDUCTION

Go ahead and switch back to design perspective. The next set of steps will teach you to reduce the number of observations in your data set through the process of filtering.

1) In the search box within the Operators tab, type in the word 'filter'. This will help you locate the 'Filter Examples' operator, which is what we will use in this example. Drag the

Filter Examples operator over and connect it into your stream, right after the Replace Missing Values operator. Your window will look like Figure 3-26.



Filter Examples

condition class: all

□ invert filter

One potential problem

Message	Fixes	Location
Parameter 'repository entry' accesses a repository ...	No quick fix available	Retrieve

Filter Examples

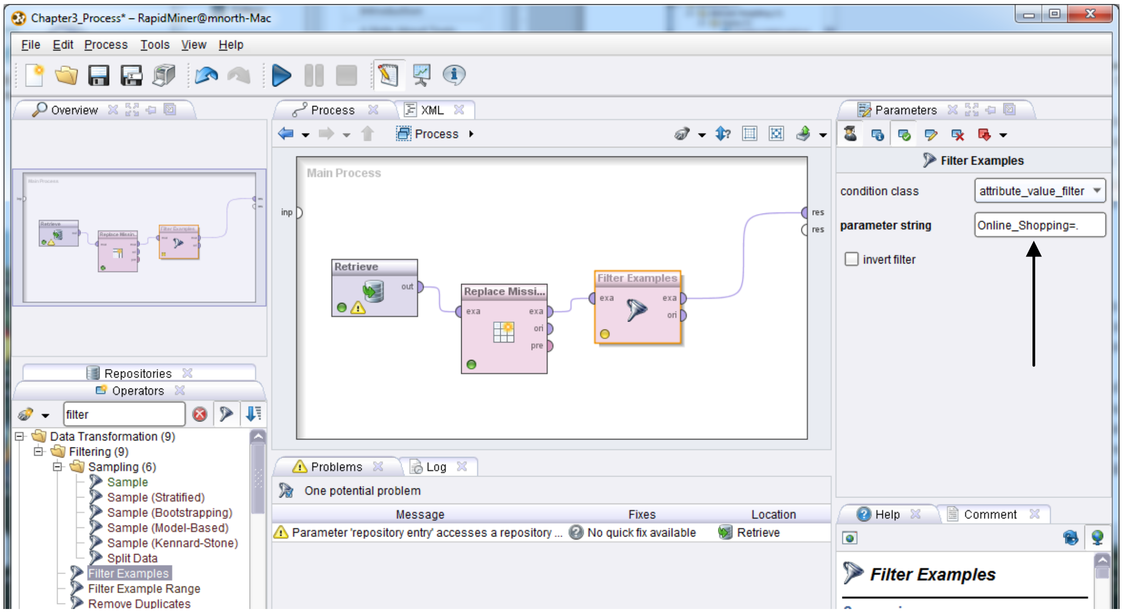
Synopsis

This operator only allows examples which fulfill a specified condition.

Description

Figure 3-26. Adding a filter to the stream.

2) In the condition class, choose 'attribute_value_filter', and for the parameter_string, type the following: **Online_Shopping=.** Be sure to include the period. This parameter string refers to our attribute, Online_Shopping, and it tells RapidMiner to filter out all observations where the value in that attribute is missing. This is a bit confusing, because in Data View in results perspective, missings are denoted by a question mark (?), but when entering the parameter string, missings are denoted by a period (.). Once you've typed these parameter values in, your screen will look like Figure 3-27.



Filter Examples

condition class: attribute_value_filter

parameter string: Online_Shopping=.

□ invert filter

One potential problem

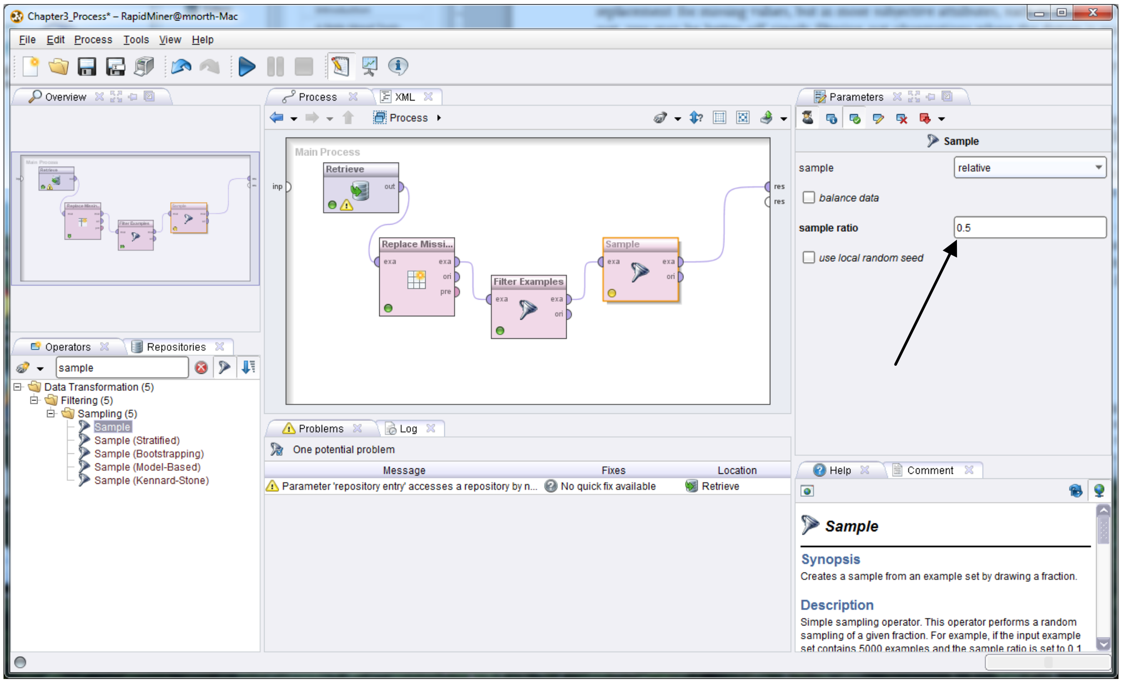
Message	Fixes	Location
Parameter 'repository entry' accesses a repository ...	No quick fix available	Retrieve

Figure 3-27. Adding observation filter parameters.

Go ahead and run your model by clicking the play button. In results perspective, you will now see that your data set has been reduced from eleven observations (or examples) to nine. This is because the two observations where the Online_Shopping attribute had a missing value have been removed. You'll be able to see that they're gone by selecting the Data View radio button. They have not been deleted from the original source data, they are simply removed from the data set at the point in the stream where the filter operator is located and will no longer be considered in any downstream data mining operations. In instances where the missing value cannot be safely assumed or computed, removal of the entire observation is often the best course of action. When attributes are numeric in nature, such as with ages or number of visits to a certain place, an arithmetic measure of central tendency, such as **mean, median** or **mode** might be an acceptable replacement for missing values, but in more subjective attributes, such as whether one is an online shopper or not, you may be better off simply filtering out observations where the datum is missing. (One cool trick you can try in RapidMiner is to use the Invert Filter option in design perspective. In this example, if you check that check box in the parameters pane of the Filter Examples operator, you will *keep* the missing observations, and filter out the rest.)

Data mining can be confusing and overwhelming, especially when data sets get large. It doesn't have to be though, if we manage our data well. The previous example has shown how to filter out observations containing undesired data (or missing data) in an attribute, but we can also reduce data to test out a data mining model on a smaller subset of our data. This can greatly reduce processing time while testing a model to see if it will work to answer our questions. Follow the steps below to take a **sample** of our data set in RapidMiner.

1) Using the search techniques previously demonstrated, use the Operators search feature to find an operator called 'Sample' and add this to your stream. In the parameters pane, set the sample to be to be a 'relative' sample, and then indicate you want to retain 50% of your observations in the resulting data set by typing **.5** into the sample ratio field. Your window should look like Figure 3-28.



sample: sample

Sample

sample: relative

☐ balance data

sample ratio: 0.5

☐ use local random seed

One potential problem

Message	Fixes	Location
Parameter 'repository entry' accesses a repository by n...	No quick fix available	Retrieve

Sample

Synopsis

Creates a sample from an example set by drawing a fraction.

Description

Simple sampling operator. This operator performs a random sampling of a given fraction. For example, if the input example set contains 5000 examples and the sample ratio is set to 0.1

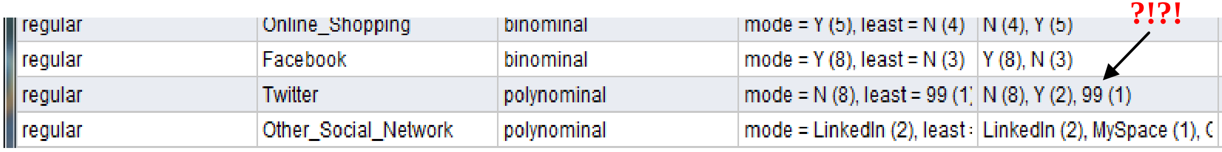
Figure 3-28. Taking a 50% random sample of the data set.

2) When you run your model now, you will find that your results only contain four or five observations, randomly selected from the nine that were remaining after our filter operator removed records that had missing Online_Shopping values.

Thus you can see that there are many ways, and various reasons to reduce data by decreasing the number of observations in your data set. We'll now move on to handling inconsistent data, but before doing so, it is going to be important to reset our data back to its original form. While filtering, we removed an observation that we will need in order to illustrate what inconsistent data is, and to demonstrate how to handle it in RapidMiner. This is a good time to learn how to remove operators from your stream. Switch back to design perspective and click on your Sampling operator. Next, right click and choose Delete, or simply press the Delete key on your keyboard. Delete the Filter Examples operator at this time as well. Note that your spline that was connected to the *res* port is also deleted. This is not a problem, you can reconnect the *exa* port from the Replace Missing Values operator to the *res* port, or you will find that the spline will reappear when you complete the steps under Handling Inconsistent Data.

HANDLING INCONSISTENT DATA

Inconsistent data is different from missing data. Inconsistent data occurs when a value does exist, however that value is not valid or meaningful. Refer back to Figure 3-25, a close up version of that image is shown here as Figure 3-29.



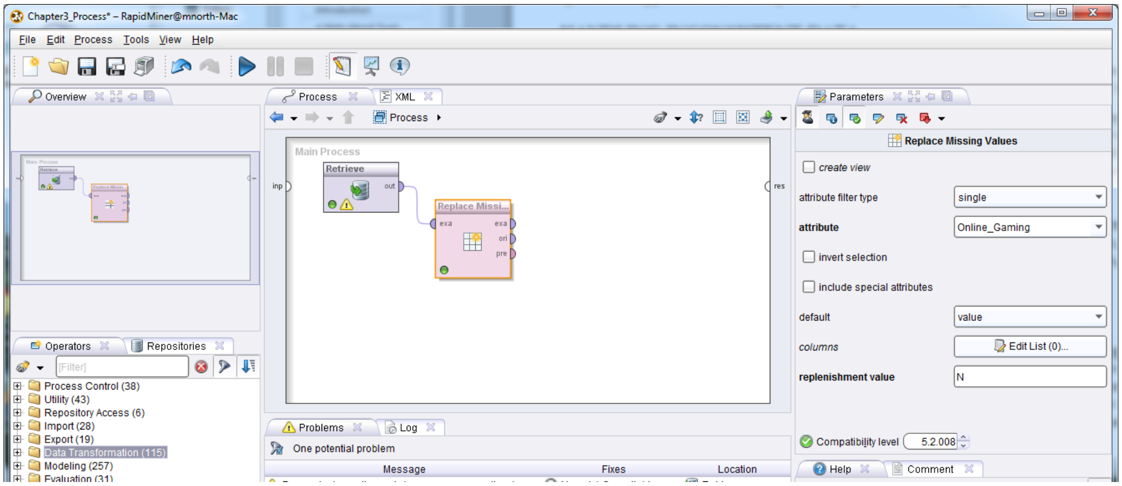
regular	Online_Shopping	binominal	mode = Y (5), least = N (4)	N (4), Y (5)
regular	Facebook	binominal	mode = Y (8), least = N (3)	Y (8), N (3)
regular	Twitter	polynominal	mode = N (8), least = 99 (1)	N (8), Y (2), 99 (1)
regular	Other_Social_Network	polynominal	mode = LinkedIn (2), least	LinkedIn (2), MySpace (1), (

?!?!

Figure 3-29. Inconsisten data in the Twitter attribute.

What is that 99 doing there? It seems that the only two valid values for the Twitter attribute should be 'Y' and 'N'. This is a value that is inconsistent and is therefore meaningless. As data miners, we can decide if we want to filter this observation out, as we did with the missing Online_Shopping records, or, we could use an operator designed to allow us to replace certain values with others.

1) Return to design perspective if you are not already there. Ensure that you have deleted your sampling and filter operators from your stream, so that your window looks like Figure 3-30.



Replace Missing Values

☐ create view

attribute filter type: single

attribute: Online_Gaming

☐ invert selection

☐ include special attributes

default: value

columns: Edit List (0)...

replenishment value: N

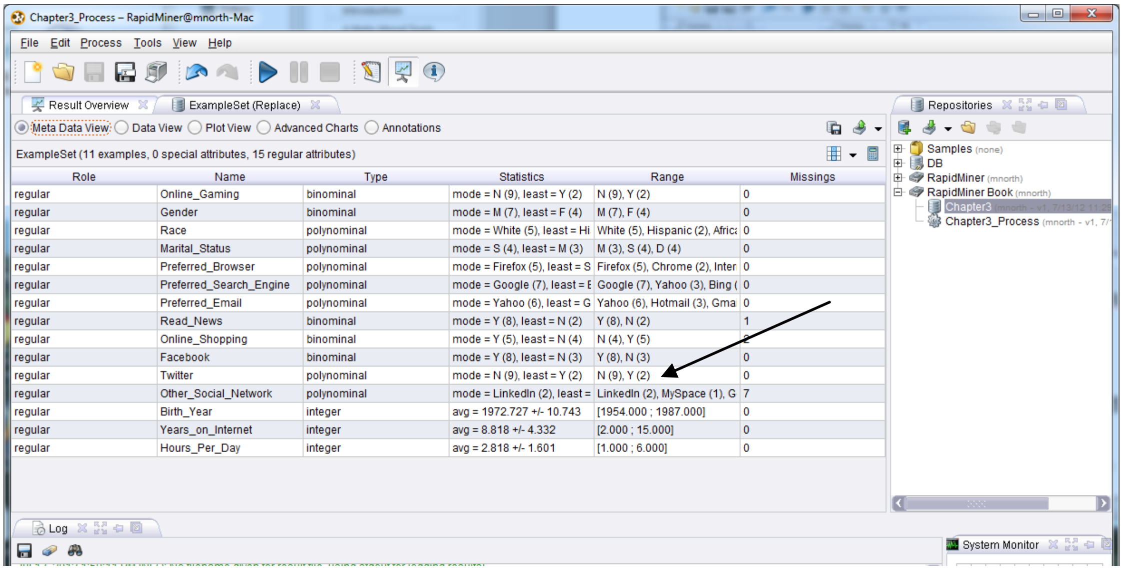
Compatibility level 5.2.008

One potential problem

Figure 3-30. Returning to a full data set in RapidMiner.

2) Note that we don't need to remove the Replace Missing Values operator, because it is not removing any observations in our data set. It only changes the values in the Online_Gaming attribute, which won't affect our next operator. Use the search feature in the Operators tab to find an operator called Replace. Drag this operator into your stream. If your splines had been disconnected during the deletion of the sampling and filtering operators, as is the case in Figure 3-30, you will see that your splines are automatically reconnected when you add the Replace operator to the stream.

3) In the parameters pane, change the attribute filter type to single, then indicate Twitter as the attribute to be modified. In truth, in this data set there is only one instance of the value 99 across all attributes and observations, so this change to a single attribute is not actually necessary in this example, but it is good to be thoughtful and intentional with every step in a data mining process. Most data sets will be far larger and more complex that the Chapter 3 data set we are currently working with. In the 'replace what' field, type the value 99, since this is the value we're looking to replace. Finally, in the 'replace by' field, we must decide what we want to have in the place of the 99. If we leave this field blank, then the observation will have a missing (?) when we run the model and switch to Data View in results perspective. We could also choose the mode of 'N', and given that 80% of the survey respondents indicated that they did not use Twitter, this would seem a safe course of action. You may choose the value you would like to use. For the book's example, we will enter 'N' and then run our model. You can see in Figure 3-31 that we now have nine values of 'N', and two of 'Y' for our Twitter attribute.


Chapter3_Process – RapidMiner@mnorth-Mac
File Edit Process Tools View Help
Result Overview
ExampleSet (Replace)
◉ Meta Data View ○ Data View ○ Plot View ○ Advanced Charts ○ Annotations
Repositories
Samples (none)
DB
RapidMiner (mnorth)
RapidMiner Book (mnorth)
Chapter3 (mnorth - v1, 7/13/12 11:28
Chapter3_Process (mnorth - v1, 7/
Log
System Monitor

ExampleSet (11 examples, 0 special attributes, 15 regular attributes)

Role	Name	Type	Statistics	Range	Missings
regular	Online_Gaming	binominal	mode = N (9), least = Y (2)	N (9), Y (2)	0
regular	Gender	binominal	mode = M (7), least = F (4)	M (7), F (4)	0
regular	Race	polynominal	mode = White (5), least = Hi	White (5), Hispanic (2), Afric	0
regular	Marital_Status	polynominal	mode = S (4), least = M (3)	M (3), S (4), D (4)	0
regular	Preferred_Browser	polynominal	mode = Firefox (5), least = S	Firefox (5), Chrome (2), Inter	0
regular	Preferred_Search_Engine	polynominal	mode = Google (7), least = E	Google (7), Yahoo (3), Bing (	0
regular	Preferred_Email	polynominal	mode = Yahoo (6), least = G	Yahoo (6), Hotmail (3), Gma	0
regular	Read_News	binominal	mode = Y (8), least = N (2)	Y (8), N (2)	1
regular	Online_Shopping	binominal	mode = Y (5), least = N (4)	N (4), Y (5)	2
regular	Facebook	binominal	mode = Y (8), least = N (3)	Y (8), N (3)	0
regular	Twitter	polynominal	mode = N (9), least = Y (2)	N (9), Y (2)	0
regular	Other_Social_Network	polynominal	mode = LinkedIn (2), least =	LinkedIn (2), MySpace (1), G	7
regular	Birth_Year	integer	avg = 1972.727 +/- 10.743	[1954.000 ; 1987.000]	0
regular	Years_on_Internet	integer	avg = 8.818 +/- 4.332	[2.000 ; 15.000]	0
regular	Hours_Per_Day	integer	avg = 2.818 +/- 1.601	[1.000 ; 6.000]	0

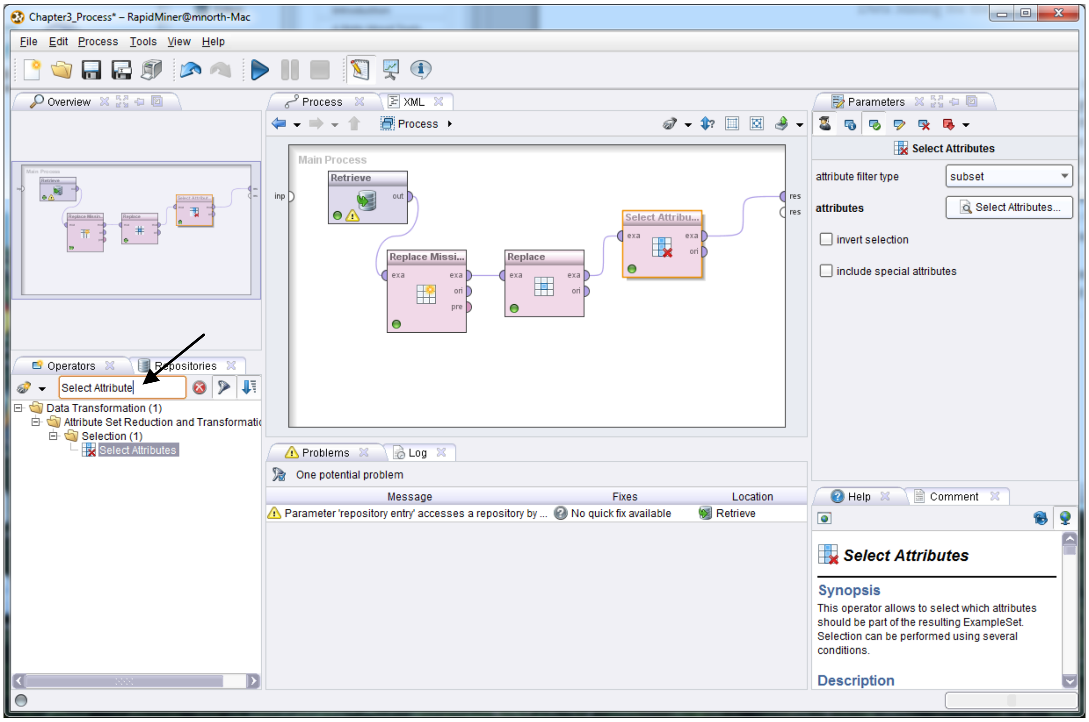
Figure 3-31. Replacement of inconsistent value with a consistent one.

Keep in mind that not all inconsistent data is going to be as easy to handle as replacing a single value. It would be entirely possible that in addition to the inconsistent value of 99, values of 87, 96, 101, or others could be present in a data set. If this were the case, it might take multiple replacements and/or missing data operators to prepare the data set for mining. In numeric data we might also come across data which are accurate, but which are also statistical outliers. These might also be considered to be inconsistent data, so an example in a later chapter will illustrate the handling of statistical outliers. Sometimes data scrubbing can become tedious, but it will ultimately affect the usefulness of data mining results, so these types of activities are important, and attention to detail is critical.

ATTRIBUTE REDUCTION

In many data sets, you will find that some attributes are simply irrelevant to answering a given question. In Chapter 4 we will discuss methods for evaluating correlation, or the strength of relationships between given attributes. In some instances, you will not know the extent to which a certain attribute will be useful without statistically assessing that attribute's correlation to the other data you will be evaluating. In our process stream in RapidMiner, we can remove attributes that are not very interesting in terms of answering a given question without completely deleting them from the data set. Remember, simply because certain variables in a data set aren't interesting for answering a certain question doesn't mean those variables won't ever be interesting. This is why we recommended bringing in all attributes when importing the Chapter 3 data set earlier in this chapter—uninteresting or irrelevant attributes are easy to exclude within your stream by following these steps:

1) Return to design perspective. In the operator search field, type Select Attribute. The Select Attributes operator will appear. Drag it onto the end of your stream so that it fits between the Replace operator and the result set port. Your window should look like Figure 3-32.



Select Attribute

attribute filter type: subset

attributes: Select Attributes...

□ invert selection

□ include special attributes

One potential problem

Message	Fixes	Location
Parameter 'repository entry' accesses a repository by ...	No quick fix available	Retrieve

Select Attributes

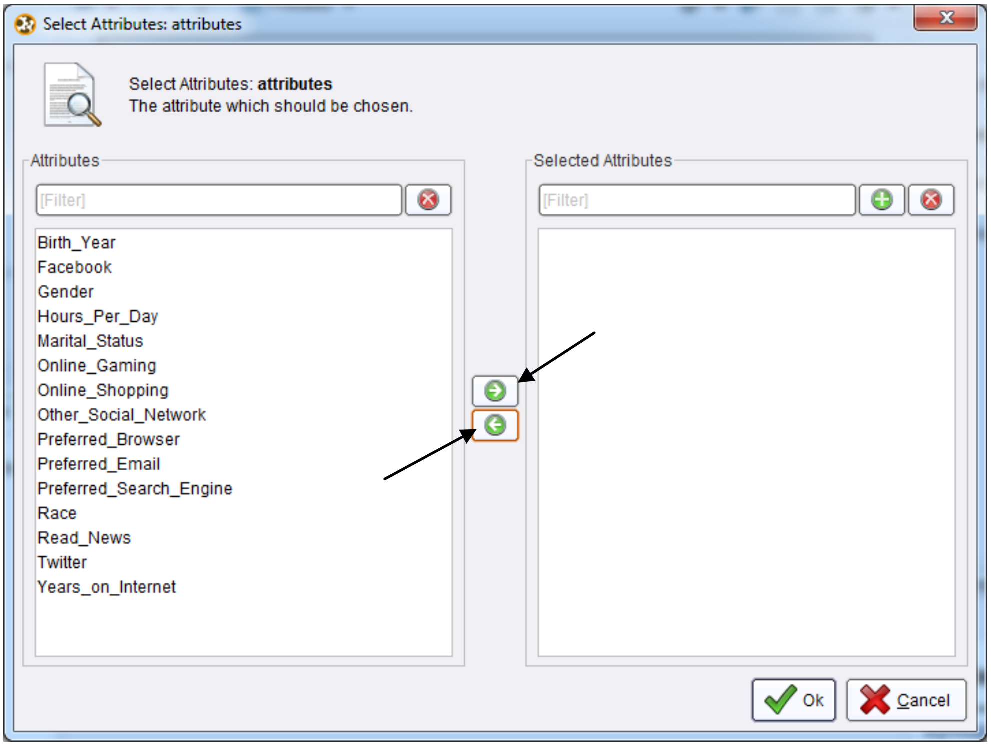
Synopsis

This operator allows to select which attributes should be part of the resulting ExampleSet. Selection can be performed using several conditions.

Description

Figure 3-32. Selecting a subset of a data set's attributes.

2) In the Parameters pane, set the attribute filter type to 'subset', then click the Select Attributes button; a window similar to Figure 3-33 will appear.



Select Attributes: **attributes**
The attribute which should be chosen.

Birth_Year
Facebook
Gender
Hours_Per_Day
Marital_Status
Online_Gaming
Online_Shopping
Other_Social_Network
Preferred_Browser
Preferred_Email
Preferred_Search_Engine
Race
Read_News
Twitter
Years_on_Internet

Figure 3-33. The attribute subset selection window.

3) Using the green right and left arrows, you can select which attributes you would like to keep. Suppose we were going to study the demographics of Internet users. In this instance, we might select Birth_Year, Gender, Marital_Status, Race, and perhaps Years_on_Internet, and move them to the right under Selected Attributes using the right green arrow. You can select more than one attribute at a time by holding down your control or shift keys (on a Windows computer) while clicking on the attributes you want to select or deselect. We could then click OK, and these would be the only attributes we would see in results perspective when we run our model. All subsequent downstream data mining operations added to our model will act only upon this subset of our attributes.

CHAPTER SUMMARY

This chapter has introduced you to a number of concepts related to data preparation. Recall that Data Preparation is the third step in the CRISP-DM process. Once you have established Organizational Understanding as it relates to your data mining plans, and developed Data Understanding in terms of what data you need, what data you have, where it is located, and so forth; you can begin to prepare your data for mining. This has been the focus of this chapter.

The chapter used a small and very simple data set to help you learn to set up the RapidMiner data mining environment. You have learned about viewing data sets in OpenOffice Base, and learned some ways that data sets in relational databases can be collated. You have also learned about comma separated values (CSV) files.

We have then stepped through adding CSV files to a RapidMiner data repository in order to handle missing data, reduce data through observation filtering, handle inconsistencies in data, and reduce the number of attributes in a model. All of these methods will be used in future chapters to prepare data for modeling.

Data mining is most successful when conducted upon a foundation of well-prepared data. Recall the quotation from Chapter 1from *Alice's Adventures in Wonderland*—which way you go does not matter very much if you don't know, or don't care, where you are going. Likewise, the value of where you arrive when you complete a data mining exercise will largely depend upon how well you prepared to get there. Sometimes we hear the phrase "It's better than nothing". Well, in data mining, results gleaned from poorly prepared data might be "Worse than nothing", because they may be misleading. Decisions based upon them could lead an organization down a detrimental and costly path. Learn to value the process of data preparation, and you will learn to be a better data miner.

REVIEW QUESTIONS

1) What are the four main processes of data preparation discussed in this chapter? What do they accomplish and why are they important?

2) What are some ways to collate data from a relational database?

3) For what kinds of problems might a data set need to be scrubbed?

4) Why is it often better to perform reductions using operators rather than excluding attributes or observations as data are imported?

5) What is a data repository in RapidMiner and how is one created?

6) How might inconsistent data cause later trouble in data mining activities?

EXERCISE

1) Locate a data set of any number of attributes and observations. You may have access to data sets through personal data collection or through your employment, although if you use an employer's data, make sure to do so only by permission! You can also search the Internet for data set libraries. A simple search on the term 'data sets' in your favorite search engine will yield a number of web sites that offer libraries of data sets that you can use for academic and learning purposes. Download a data set that looks interesting to you and complete the following:

2) Format the data set into a CSV file. It may come in this format, or you may need to open the data in OpenOffice Calc or some similar software, and then use the File > Save As feature to save your data as a CSV file.

3) Import your data into your RapidMiner repository. Save it in the repository as Chapter3_Exercise.

4) Create a new, blank process stream in RapidMiner and drag your data set into the process window.

5) Run your process and examine your data set in both meta data view and Data View. Note if any attributes have missing or inconsistent data.

6) If you found any missing or inconsistent data, use operators to handle these. Perhaps try browsing through the folder tree in the Operators tab and experiment with some operators that were not covered in this chapter.

7) Try filtering out some observations based on some attibute's value, and filter out some attributes.

8) Document where you found your data set, how you prepared it for import into RapidMiner, and what data preparation activities you applied to it.