

---

### Training Manual

Using Decision Support Science's Predictive Analytics Engine to Maximize Consumer Sales and CRM



*Decision Support Sciences. Better Science. Better Solutions.*

---

All contents Copyright © 2016 by Decision Support Sciences

PositionSolve™, ProductSolve™, SatisfactionSolve™, SatisfactionSolve /DP™, BigDataSolve™, BigDataSolve / DP™, SegmentSolve™ and InnovationSolve™ are trademarks of Decision Support Sciences (DSS).

No portion of this work may be reproduced in any form without the written consent of Decision Support Sciences.

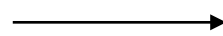
For permissions or other questions, contact DSS at [info@decisionsupportsciences.com](mailto:info@decisionsupportsciences.com).



# Introduction: What is Data Mining / Predictive Analytics?

*Data Mining is the Process of Finding Patterns in Enterprise Data in Order to Target Individuals with Some Type of Marketing Offer:*

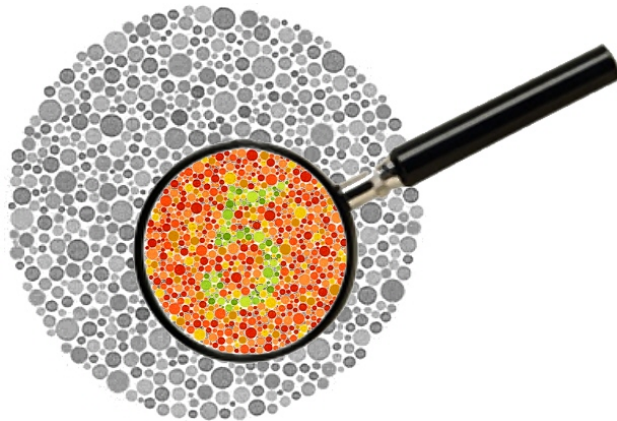
*Patterns in the Data*



*Offer*



*New Customers and/or More Unit Sales Per Customer*



*Data Mining / Predictive Analytics*

*Campaign Management*

## Introduction: What is BigDataSolve™?

BigDataSolve™ is a data mining automation tool that rapidly and automatically builds, runs, scans and validates over a million models per week on a modern laptop.

BigDataSolve™ can produce and scan thousands of models in the time it takes to put together one model with a conventional predictive analytics package. Rather than manually setting up each analysis run, the user can efficiently evaluate only the solutions BigDataSolve™ determines are the best ones.

Using the technology of distributed computing, the decision platform version of BigDataSolve™ (BigDataSolve /DP™) can be run on many computers at once connected by a local area network (LAN) or the internet. This leverages existing efficiently and economically.

---

## Table of Contents

### **I. Analysis Setup**

#### **1. Using Rules Files**

#### **2. Using the Rules Wizard**

### **II. Data Mining**

### **III. Appendix: Dialog Boxes**

---

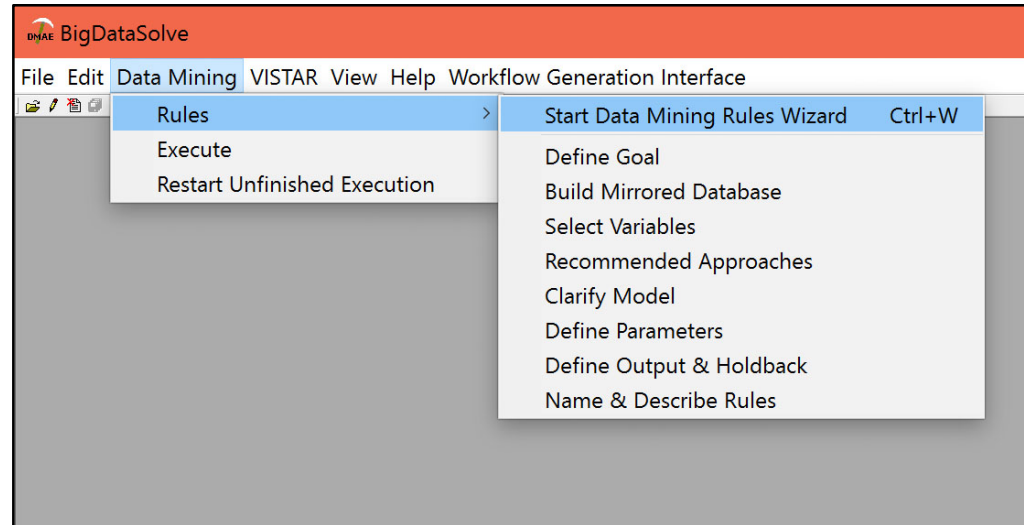
## **I. Analysis Setup**

**1. Using Rules Files**

**2. Using the Rules Wizard**



## Using Rules Files



### ■ Rules Files

- A rules file stores all of the analysis settings that you select for a run. If substantially the same databases, variables, or algorithms will be used multiple times, one rules file allows the strategist to not “start from scratch” every time.
- When high performance models are discovered, a rules file constitutes a methods repository and hence becomes part of the knowledge capital of your organization.
- NOTE: A rules file does not store the analysis reports or results...but it allows the run to be re-executed. The results are stored in the results directories(s).

### ■ To Begin using BigDataSolve™, Simply Start or Open an Existing File:

- To start a rules file from scratch, start the Rules Wizard. This is done by selecting the Select Start Rules Wizard from the Rules submenu of the Data Mining menu, as shown at the right.
- To modify the existing file or set the preferences for the new rules file, use the Rules Wizard.
- To execute a new or existing rules file, select Data Mining, Execute.

## Using the Rules Wizard: Step 1



### ■ Using Rules Files

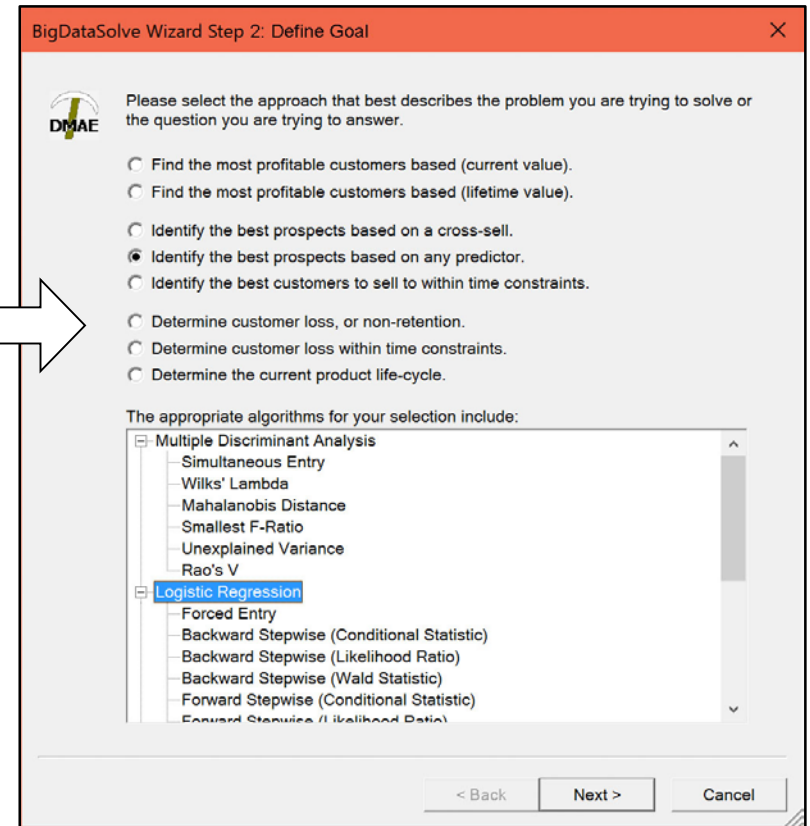
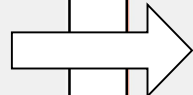
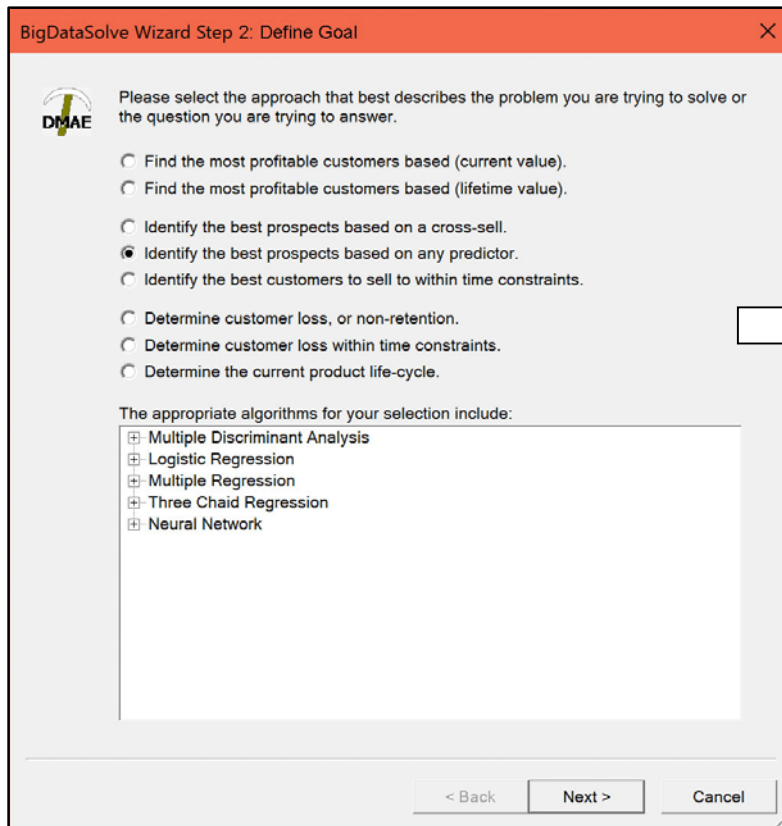
- To start the Rules Wizard, select the pencil icon button from the toolbar or select the Start Rules Wizard from the Rules submenu of the Data Mining menu. The first screen will allow you to indicate whether you want to start from scratch, modify an existing file, or open an existing rules file.
  - **Start from Scratch:** All settings will be blank or set to their default values.
  - **Modify the Existing Rules:** Settings will be set exactly as during the prior saved run. If you modify these the modified rules file can be saved under a different filename at the end of the wizard.
  - **Open a rules file to work from:** an Open File dialog box will appear when you click on the Next button. Browse through existing rules files to select the file desired, and click OK.
- The options to create a new rules file or open an existing one are also available manually from the File menu (the New Rules and Open Rules menu items).



## Using the Rules Wizard: Step 2

### ■ Define Goal

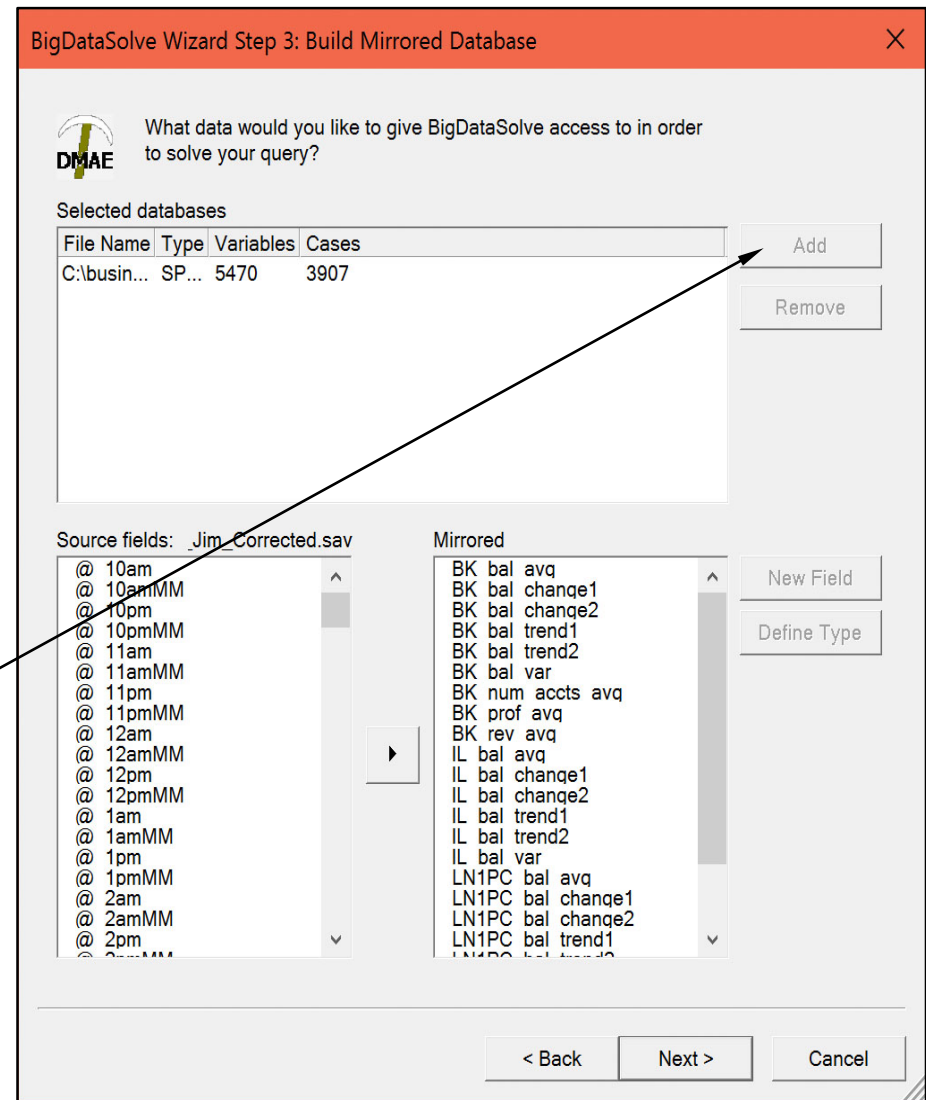
- BigDataSolve™ develops optimum predictive analytics solutions. In this dialog, typical categories of business problems are listed along with the preferred techniques to solve them.
- When you select a problem you wish to address, the algorithms that apply to it will appear in the lower text box. To open or close branches, click on the + or - boxes to the left of the tree.



## Using the Rules Wizard: Step 3


### ■ Build a Mirrored Database (A Modeling Mart)

- A database is required to provide the input for the data mining. The input file must be an IBM-SPSS 7.5 or higher .sav file. IBM-SPSS need not be installed on the computer before using BigDataSolve™. **Note: SPSS should NOT be running when BigDataSolve™ is open; if SPSS is running, the BigDataSolve™ server will not connect to the root server. (The *BigDataSolve™ root server* application is the single monitor program that creates the models.)**
- Select the Add button to open a browse dialog box to choose a file. While the file is being loaded, a progress indicator will tell you the database name and fields that are being imported.
- Select the variables to use from the Source Fields box, and move them to the Mirrored Fields box using the arrow button in the middle.



## Using the Rules Wizard: Step 4

BigDataSolve Wizard Step 4: Select Variables

 Which data fields would you like BigDataSolve to schedule for prediction, and which fields should be used to predict those variables?

Mirrored

Field to Predict  
TotalAssetDollars

Fields to be used in prediction

- BK bal avq
- BK bal change1
- BK bal change2
- BK bal trend1
- BK bal trend2
- BK bal var
- BK num accts avq
- BK prof avq
- BK rev avq
- IL bal avq
- IL bal change1
- IL bal change2
- IL bal trend1
- IL bal trend2
- IL bal var
- LN1PC bal avq
- LN1PC bal change1
- LN1PC bal change2
- LN1PC bal trend1
- LN1PC bal trend2
- LN1PC bal var
- LN1PC num accts avq

Continuous Dependent Variable  
Classification is correct if  
5 or 10 %  
of actual value, whichever is greater

< Back    Next >    Cancel

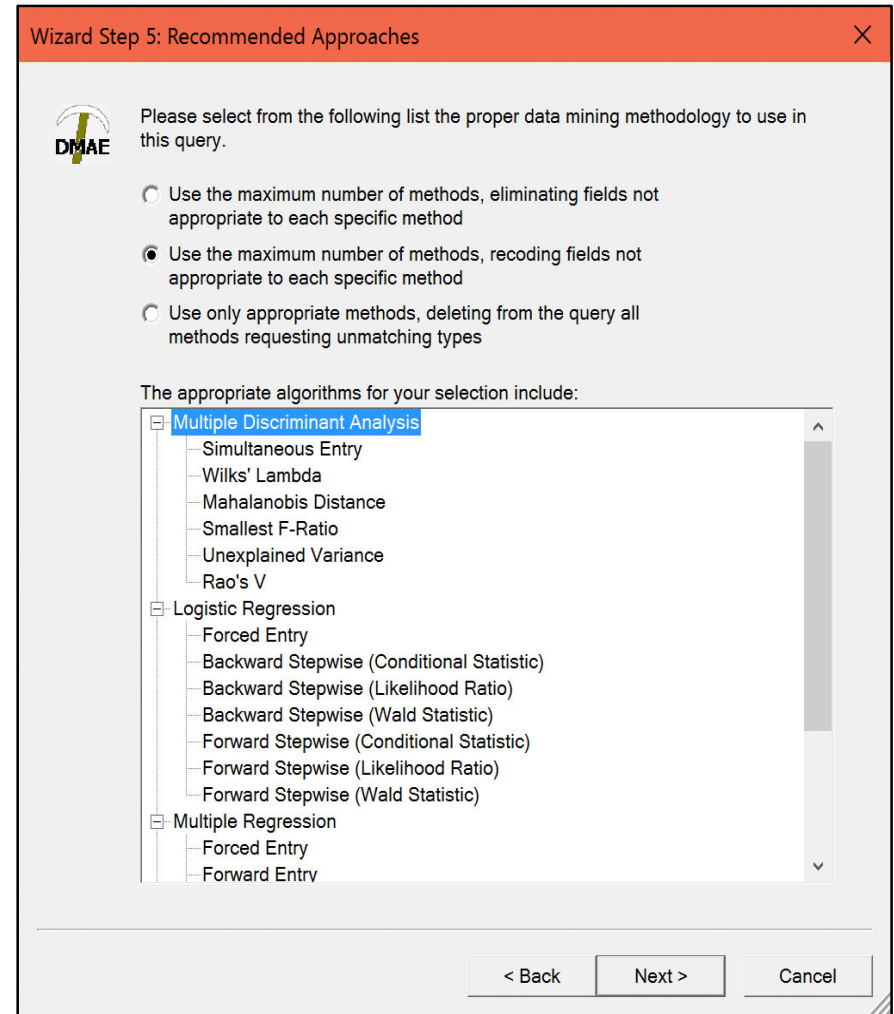
### ■ Select Variables

- Select one variable to predict, and at least one variable to be used in prediction. The field to predict may be either a categorical (discrete) variable with 2 or more non-missing levels, or a scalar variable. Highlight the variable(s) in the Mirrored Fields box, and use the arrows to move them to the box of selected variables.
- All of the available fields are originally listed in the Mirrored fields list box. In this example, we have selected the field TotalAssetDollars to be the variable to predict and ALL the rest of the fields to be used in the prediction.

## Using the Rules Wizard: Step 5

### ■ Identify Recommended Approaches

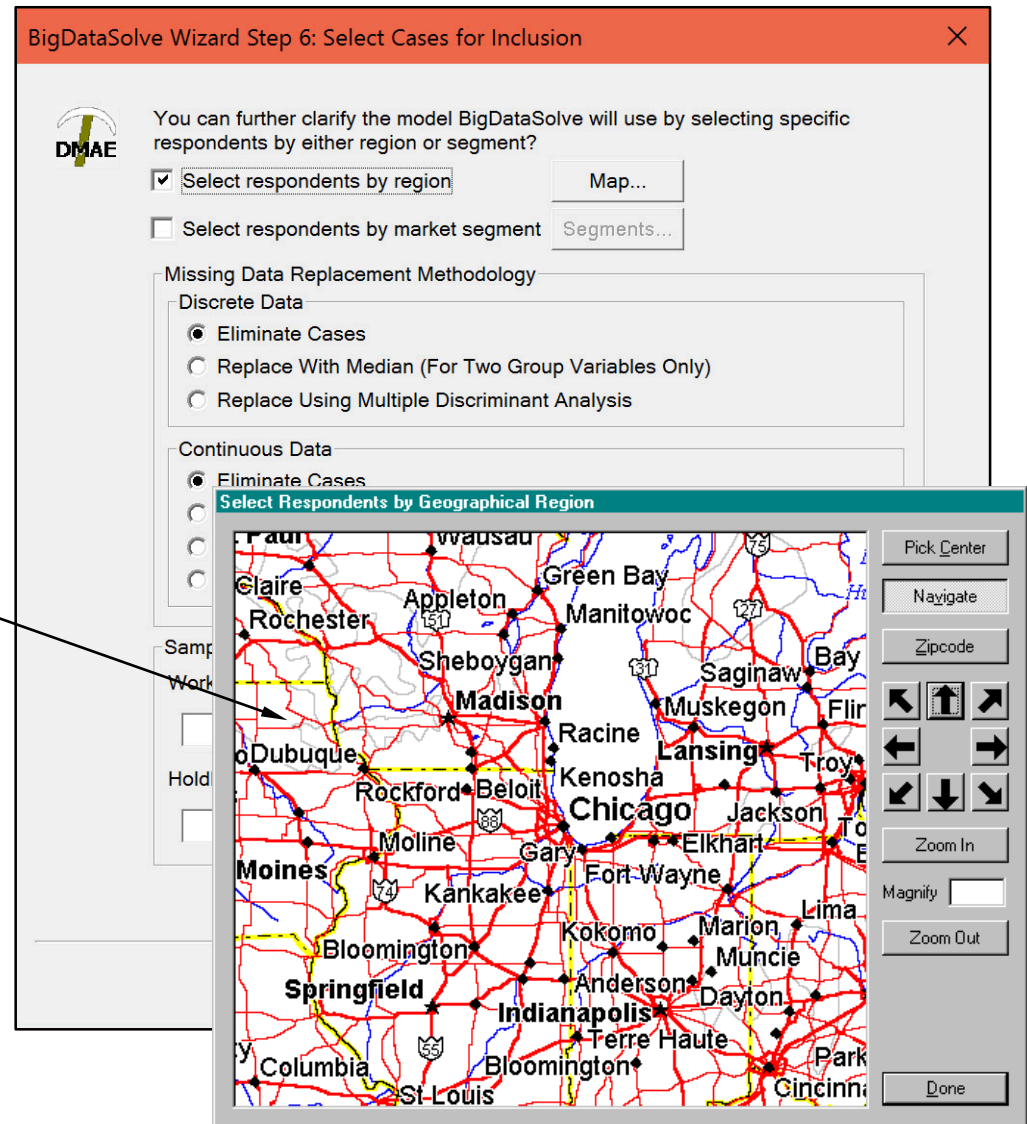
- The data variables that were selected for the analysis may not be appropriate for every analysis method that is used. There are three options for dealing with data fields that are incompatible with an analysis method.
  - **Use the maximum number of methods, eliminating fields not appropriate to each specific method:** This option leaves the incompatible data field out of the particular method for which it is inappropriate.
  - **Use the maximum number of methods, recoding fields not appropriate to each specific method:** The second option attempts to use all of the selected data fields. Data that is not compatible with a specific method will be recoded (i.e. a continuous variable will be encoded as a dummy variable).
  - **Use only appropriate methods, deleting from the query all methods requesting unmatched types:** This option eliminates the analysis method, rather than the data field, if there is incompatible data.



## Using the Rules Wizard: Step 6

### ■ Select Cases for Inclusion by Region

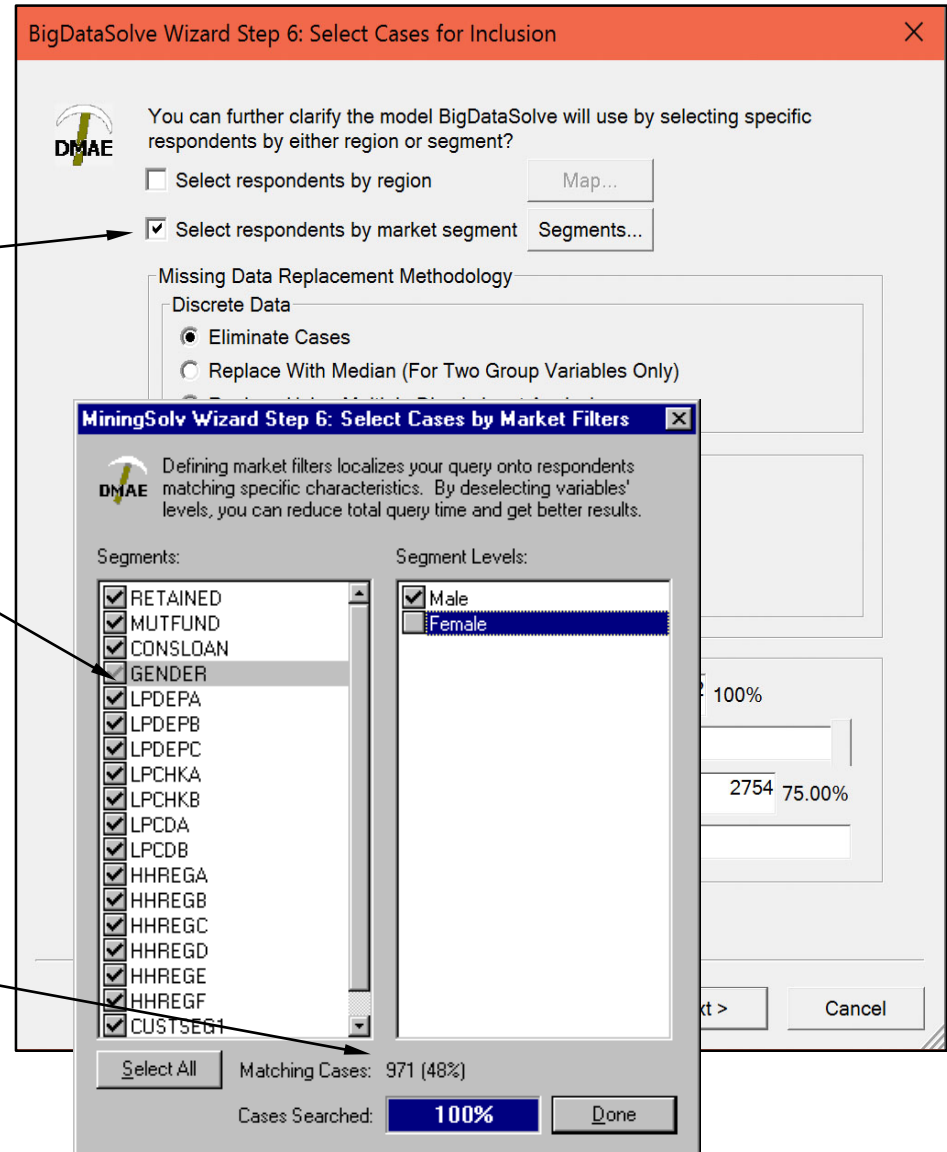
- A specific segment of respondents can be extracted from the data for an analysis run. These respondents can be selected either by region or by market segment.
  - The region option can only be used if location information has been included in the selected databases. If the region option is checked, a map dialog will appear. Use the zoom and arrow controls to select the desired region.



## Using the Rules Wizard: Step 6 (cont'd)

### ■ Select Cases for Inclusion by Segment

- To select by segment rather than by region click on the “Select respondents by region”
- Market segment: Specific segments of the market can be included or excluded from the analysis. Use the mouse to click on the checkmark to the left of the segment or level. The levels for the highlighted segment appear in the Segment Levels box on the right. Segments and levels with a checkmark will be included in the analysis. Select the Done button when you are finished selecting segments. The total number of cases to be used is displayed at the bottom of the dialog at “Matching Cases”.



## Using the Rules Wizard: Step 6 (cont'd)

### ■ Select Cases for Inclusion, cont.

- In data mining, models must be validated. The most reliable and real-world method of validation is to test the model on cases not used to build the model. This allows the strategist to compare what the model predicts to what is already known about the disposition of that variable. This is called hold-back sample validation.
- The Sample Size reflects the number of valid cases, taking into account any segments that were selected. A portion of the valid cases can be randomly selected by adjusting the working file size slider.
- The Rules Wizard allows you to specify the holdback sample. The holdback sample is the portion of the data that is **excluded** from the current analysis run. Note: the holdback sample is NOT the sample of respondents being **included** in the analysis.
- The percentage of the sample and number of respondents in the holdback sample may be specified above the slider.

BigDataSolve Wizard Step 6: Select Cases for Inclusion

You can further clarify the model BigDataSolve will use by selecting specific respondents by either region or segment?

Select respondents by region Map...

Select respondents by market segment Segments...

Missing Data Replacement Methodology

Discrete Data

Eliminate Cases

Replace With Median (For Two Group Variables Only)

Replace Using Multiple Discriminant Analysis

Continuous Data

Eliminate Cases

Replace with Mean

Replace Using Multiple Regression

Replace Using Maximum Likelihood Estimation

Sample Size (3672 Valid Cases)

Working File Size:  100%

Holdback Sample (75% recommended):  75.00%

< Back Next > Cancel

## Using the Rules Wizard: Step 7

BigDataSolve Wizard Step 7: Define Parameters

**DMAE** Please customize the ways in which you want BigDataSolve to work with the data in order to get better output. By selecting different algorithms from the drop-down box, you can see what the values in the sliders mean for each particular approach.

Data mining method:  
Multiple Discriminant Analysis

Customize each algorithm individually  Disable this algorithm  
 Use the same settings for all algorithms

Methods: DIRECT plus 4/5 of the others

Criteria: PIN varied from 0.01 to 0.09 in 23 steps

Number of Nodes Per Layer

Iterations: Not Applicable to this Algorithm

Transformations: powers of 0.50000 to 1.50000 by 0.14286

Transformation Types: Power plus 1/2 of the others

Multiple Discriminant Analysis will generate 1488 syntax scripts.  
 All algorithms together will generate 5391 syntax scripts.

Advanced  Cutoff at: 0

< Back Next > Cancel

### ■ Define Parameters

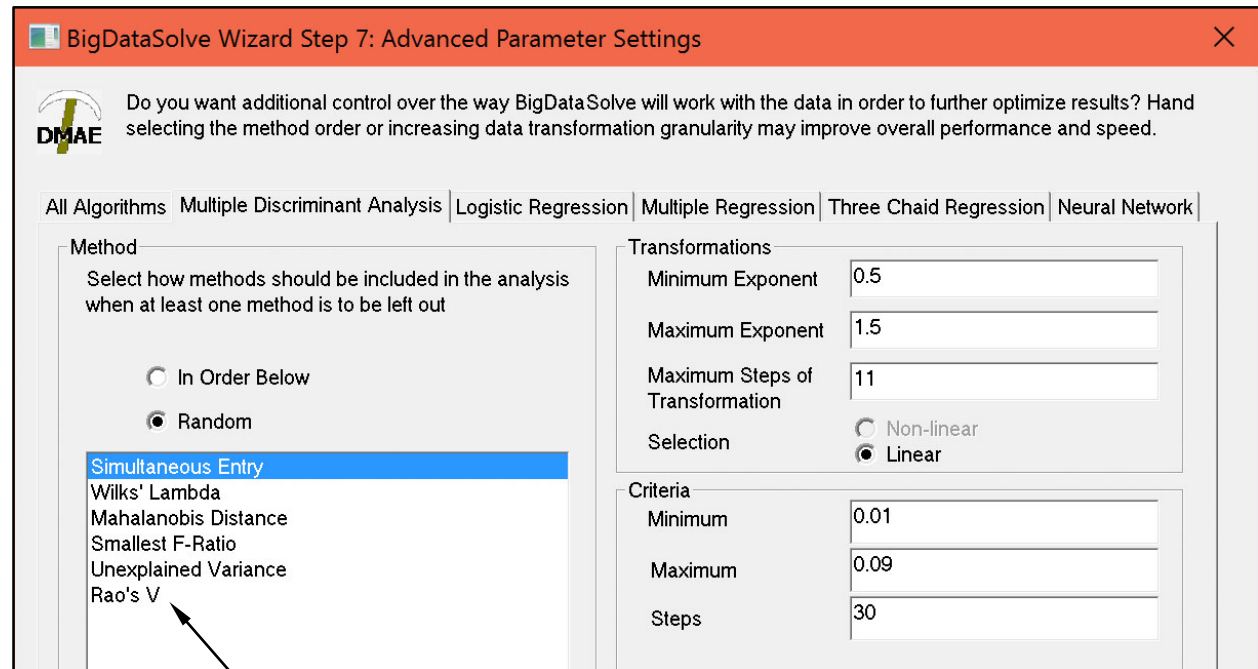
- This dialog displays the analysis specifications for each data mining method.
- The parameters can be set the same for all algorithms, or uniquely for each algorithm. Use the radio buttons under the Data mining method box to specify your choice.
  - If 'Use the same settings for all algorithms' is selected, *All algorithms* must be selected under *Data mining method* in order to move the sliders.
  - To customize each algorithm individually, set the parameters for one method, then select the next method to customize.
- Move the sliders to change the level for that criteria. If an algorithm is being customized individually, the current level of the parameter is displayed on the right side of the box.
- The combined effect of the changing the sliders on the number of model runs is displayed below the last slider; it changes dynamically as the level of any specific parameter is modified.



## Using the Rules Wizard: Step 7

### ■ Define Parameters

- To further customize the analysis, select the “Advanced” button at the bottom of the prior dialog to determine the order of the selected methods or the transformations settings.

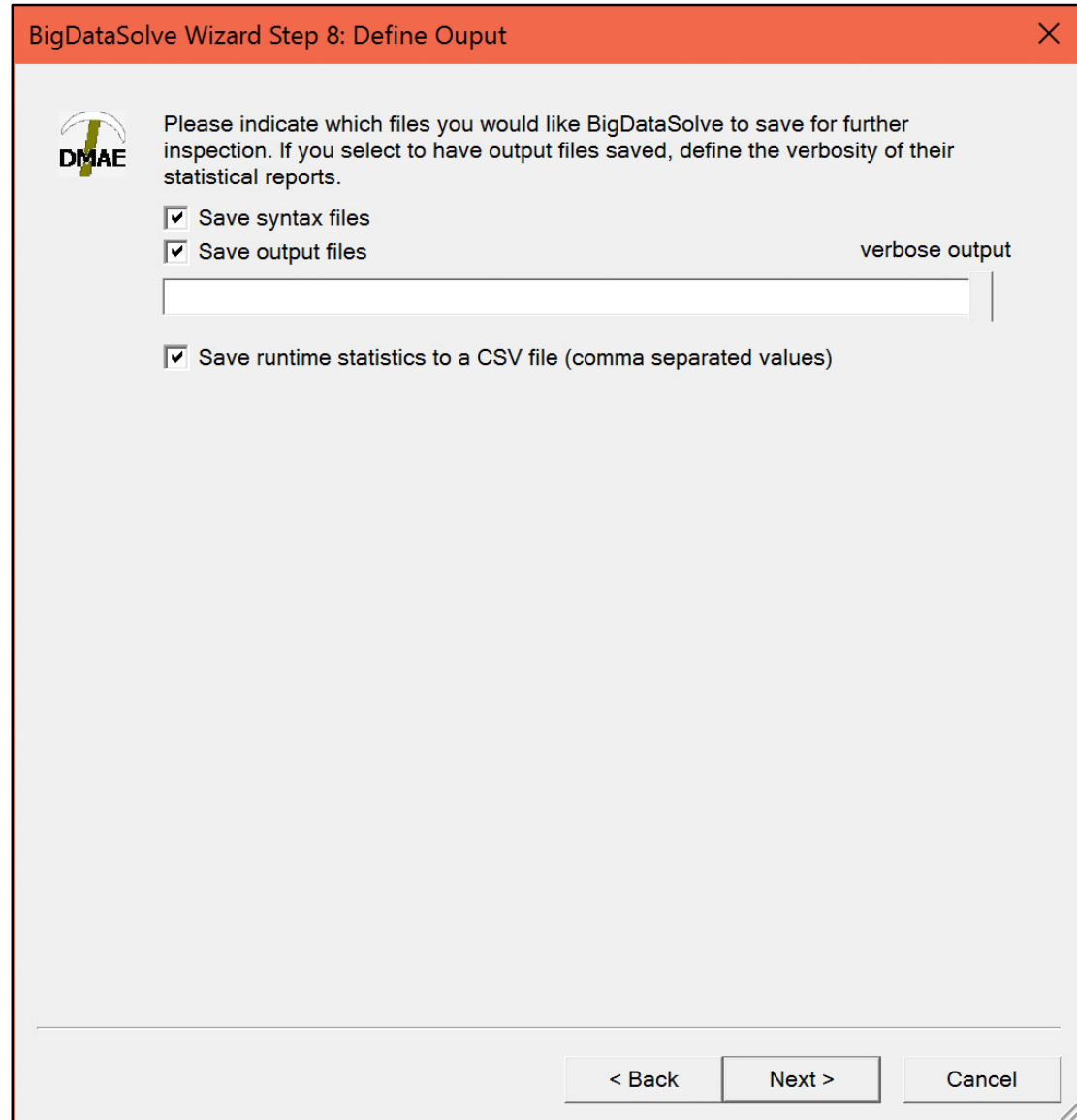


- For example, if Rao’s V is set as the first method in the advanced dialog for Multiple Discriminant Analysis AND “In Order Below” is selected, and only 1/6 methods is selected on the Define Parameters dialog (as the settings are shown), then scripts will only be generated for MDA using the Rao’s V method.
- The total number of syntax scripts (models) to be generated is shown at the bottom of the prior dialog. If a cutoff number is set, BigDataSolve™ will randomly choose from the possible runs the cutoff number to run in the analysis.

## Using the Rules Wizard: Step 8

### ■ Define Output Depth and Holdback Sample

- In this dialog, the strategist can select how BigDataSolve™ handles the output, and which output is saved for future use.
- Changing the style of output affects how much information is included in the output files. This does not affect the runtime statistics information.

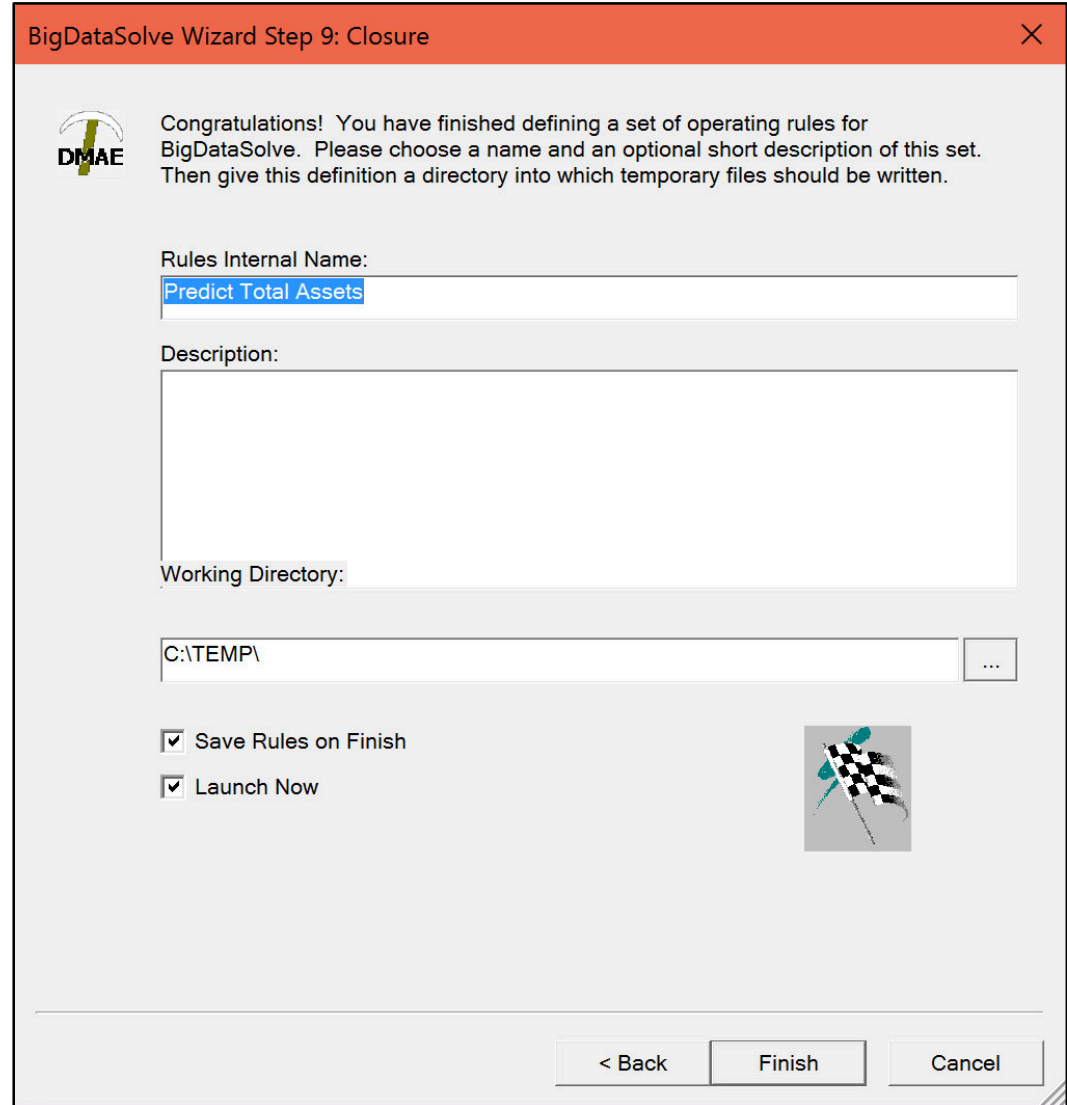


The screenshot shows a dialog box titled "BigDataSolve Wizard Step 8: Define Output". It features a red header bar with a close button (X) in the top right corner. On the left side, there is a logo for "DMAE" consisting of a stylized green and yellow graphic above the text "DMAE". The main text of the dialog reads: "Please indicate which files you would like BigDataSolve to save for further inspection. If you select to have output files saved, define the verbosity of their statistical reports." Below this text are three checked checkboxes: "Save syntax files", "Save output files", and "Save runtime statistics to a CSV file (comma separated values)". To the right of the "Save output files" checkbox is a text input field containing the text "verbose output". At the bottom of the dialog, there are three buttons: "< Back", "Next >", and "Cancel".

## Using the Rules Wizard: Step 9

### ■ Closure

- It is best to name and save your settings as a rules file in case you want to use the same or similar settings in the future.
- The 'Rules Internal Name' field is not a file name, so does not have standard file name restrictions; characters such as / \ , : ; and " are allowed.
- When the Finish button is selected, you will be asked if you want to save the Rules file. A Save As dialog box will appear, and you can specify a filename and directory.
- When a rules file is complete, select Execute from the Data Mining menu to run the analysis.



The screenshot shows the 'BigDataSolve Wizard Step 9: Closure' dialog box. It features a red title bar with a close button. On the left is the DMAE logo. The main text reads: 'Congratulations! You have finished defining a set of operating rules for BigDataSolve. Please choose a name and an optional short description of this set. Then give this definition a directory into which temporary files should be written.' Below this are three input fields: 'Rules Internal Name' (containing 'Predict Total Assets'), 'Description' (empty), and 'Working Directory' (containing 'C:\TEMP\'). There are two checked checkboxes: 'Save Rules on Finish' and 'Launch Now'. A checkered flag icon is positioned to the right of these checkboxes. At the bottom are three buttons: '< Back', 'Finish', and 'Cancel'.

---

## II. Data Mining



## Execute Data Mining

### ■ Execute

- A rules file must exist and be *open* to execute data mining.
- Make sure that SPSS is not open.
- From the Root Server, select Execute from the Data Mining menu.
- BigDataSolve™ servers will open the .sav data files that are linked to the open rules file, start the runs,, and then will build predictive models.
- If selected, IBM-SPSS will also be launched to perform the runs in parallel to ensure the results are mathematically identical. Select this in the preferences menu.
- (Note: BigDataSolve™'s native algorithms run 10-100 times faster than IBM-SPSS, but BigDataSolve™ is not as capable in handling off-nominal input data conditions such as matrices that are quite difficult to invert).
- The dialog at the right is the Server Information portion of the Runtime Information Center. All potential servers on the network are displayed, along with their current status. To see more information, select one of the options on the left side of the dialog.
- The analysis can be paused at any time by using the Suspend button. To resume analysis, select "Restart Unfinished Execution" from the Data Mining menu.

**Runtime Information Center**

Elapsed Time: 00:01:32  
 Estimated Time Remaining: 00:05:36  
 Runs Completed: 63 of 500  
 Servers Connected: 4  
 Progress:  **13%**

**Best Model**  
 Algorithm: Multiple Regression  
 Method: Forward Entry  
 Criteria: 0.16  
 Transformation: 0.50  
 OCCP: 92.53% Details...

**Server Information**  
 Communications  
 Performance  
 Best Run  
 Event Log

Computer Name	Runs Completed	Average Runtime	Status
MATTP	24	00:00:03	Sending
GRANT	21	00:00:03	Sending
JIM	10	00:00:06	Generating Model
SERVER2	8	00:00:09	Generating Model
AARON	0	00:00:00	Not connected

**Current Model**  
 Algorithm: MultReg  
 Method: Backward  
 Criteria: 0.08  
 Transformation: 0.90

**Best Model**  
 Algorithm: MultReg  
 Method: Forward  
 Criteria: 0.16  
 Transformation: 0.50  
 OCCP: 92.53%

Active Servers  
 Disconnected Servers  
 Never Connected Servers

Add Server View Network

Run #	OCCP	Lift	Algorithm	Method	Criteria	Transformation
42	92.53%	60.64%	Multiple Regression	Forward Entry	0.16	0.5
37	92.49%	59.15%	Multiple Regression	Forward Entry	0.03	0.8
34	91.82%	59.08%	Multiple Discriminant Analysis	Rao's V	0.13	0.5
52	90.82%	53.28%	Multiple Regression	Backward Elimination	0.07	0.6
5	89.61%	49.67%	Multiple Discriminant Analysis	Mahalanobis Distance	0.11	1.3
26	89.85%	49.57%	Multiple Regression	Stepwise Selection	0.08	1.3
10	88.89%	47.06%	Multiple Regression	Backward Elimination	0.11	1

Abort Suspend  All  Running 500 of 1783 possible runs 0 Update

## Execute Data Mining

### ■ Execute (cont'd)

- Displayed in the “Best Model” box at the top of the dialog are the specifications of the best model in the analysis thus far, across all servers.
- Select the Details button in the Best Model box for a prediction accuracy graph of each run. The graph shows the prediction levels in numerical order, not in the order in which they were performed.
- Use the “Add Server” button to add a server that is not on the network, or that BigDataSolve™ does not find automatically.
- **(BigDataSolve / DP™ Only):** To view other computers on the local area network executing the runs, use the “View Network” button.

The screenshot shows the **Runtime Information Center** dialog box. At the top, it displays progress information: Elapsed Time: 00:01:32, Estimated Time Remaining: 00:05:36, Runs Completed: 63 of 500, Servers Connected: 4, and a progress bar at 13%. The **Best Model** section shows: Algorithm: Multiple Regression, Method: Forward Entry, Criteria: 0.16, Transformation: 0.50, and OCCP: 92.53%. Below this is a table of server information:

Computer Name	Runs Completed	Average Runtime	Status
MATTP	24	00:00:03	Sending
GRANT	21	00:00:03	Sending
JIM	10	00:00:06	Generating Model
SERVER2	8	00:00:09	Generating Model
AARON	0	00:00:00	Not connected

Below the server table, there are sections for **Current Model** and **Best Model** with their respective specifications. The **Best Model** section also includes checkboxes for **Active Servers**, **Disconnected Servers**, and **Never Connected Servers**, along with **Add Server** and **View Network** buttons. At the bottom, there is a table of recent runs:

Run #	OCCP	Lift	Algorithm	Method	Criteria	Transformation
42	92.53%	60.64%	Multiple Regression	Forward Entry	0.16	0.5
37	92.49%	59.15%	Multiple Regression	Forward Entry	0.03	0.8

Overlaid on the bottom of the Runtime Information Center is the **Computers Located on the Network** dialog box. It shows a tree view of the network structure under "Entire Network" and "Computers Near Me". An arrow points from the text in the list to the "View Network" button in the Runtime Information Center. The "Computers Found (9)" list includes: AARON, DAVE, GRANT, JIM, KRISTI, LAPTOP3, LAPTOP4, MATT, and SERVER2.

## Execute Data Mining

### Execute, cont.

- There are several different ways to view information on a current BigDataSolve™ run.
- The Communications and Event log dialogs are shown on the right.
- (BigDataSolve / DP™ Only): On the communications dialog, potential servers are shown as transparent images until they are connected. When a server is active and connected to the root server, the image will be darkened and the line connecting the server to the root server will be blue. A ball in either direction indicates data being sent.
- The bottom dialog displays the Event Log screen. This shows the status of completed runs, and connection and analysis events on the servers and root server. The 'I' symbol on the left denotes informational stats, the 'E' symbol designates errors, and the '?' symbol signifies warnings. The time that each server connects or disconnects is also shown.
- Other screens with run information may be shown by selecting an option on the left side of the Runtime Information dialog.

**Runtime Information Center**

Elapsed Time: 00:00:18  
 Estimated Time Remaining: 00:00:14  
 Runs Completed: 2 of 30  
 Servers Connected: 3  
 Progress: 7%

Best Model:  
 Algorithm: Multiple Discriminant Analysis  
 Method: Smallest F-Ratio  
 Criteria: 0.03  
 Transformation: 1.40  
 OCCP: 88.10%

Server Information  
 Communications  
 Performance  
 Best Run  
 Event Log

Buttons: Abort, Suspend, All, Running 30 of 3113 possible runs, Update

**Runtime Information Center**

Elapsed Time: 00:01:01  
 Estimated Time Remaining: Done  
 Runs Completed: 30 of 30  
 Servers Connected: 0  
 Progress: 100%

Best Model:  
 Algorithm: Multiple Discriminant Analysis  
 Method: Smallest F-Ratio  
 Criteria: 0.12  
 Transformation: 0.60  
 OCCP: 91.82%

Time	Event Category	Server
9:16:54AM on Tue Oct 09, 2001	Run Completed	JON
9:16:55AM on Tue Oct 09, 2001	Run Started	JON
9:16:57AM on Tue Oct 09, 2001	Run Completed	JON
9:16:58AM on Tue Oct 09, 2001	Run Started	JON
9:16:58AM on Tue Oct 09, 2001	Run Completed	GRANT
9:16:58AM on Tue Oct 09, 2001	Run Started	JON
9:16:59AM on Tue Oct 09, 2001	Run Completed	GRANT
9:16:59AM on Tue Oct 09, 2001	Run Started	JON
9:17:00AM on Tue Oct 09, 2001	Run Completed	GRANT
9:17:01AM on Tue Oct 09, 2001	Run Started	GRANT
9:17:03AM on Tue Oct 09, 2001	Run Completed	GRANT
9:17:03AM on Tue Oct 09, 2001	MinimoSolv Execution Completed	

Buttons: Close

## Execute Data Mining

### ■ Execute, cont.

- The image to the right is the Performance tab from the Runtime Information Center. To change the performance metric being displayed, click on one of the tabs below the graph. The graphs are dynamically updated as the runs complete.
- The image below is the Performance tab from the Runtime Information Center. To change the performance metric being displayed, click on one of the tabs below the graph. The graphs are dynamically updated as the runs complete.

**Runtime Information Center**

Elapsed Time: 00:21:52  
 Estimated Time Remaining: 00:22:57  
 Runs Completed: 61 of 750  
 Servers Connected: 1  
 Progress:  **8%**

**Best Model**  
 Algorithm: Multiple Regression  
 Method: Stepwise Selection  
 Criteria: 0.19  
 Transformation: 1.10  
 OCCP: 92.32% Details...

Server Information  
 Communications  
 Performance  
 Best Run  
 Event Log

**Lift**

OCCP Lift L1 + L2 A2 \* P2

Run #	OCCP	Lift	Algorithm	Method	Criteria	Transformation
15	87.88%	36.43%	Multiple Discriminant Analysis	Rao's V	0.05	0.5
9	91.19%	56.21%	Multiple Regression	Backward Elimination	0.07	0.5
11	89.47%	44.12%	Multiple Regression	Stepwise Selection	0.02	0.6
13	90.03%	45.78%	Multiple Discriminant Analysis	Mahalanobis Distance	0.02	0.6
14	89.38%	44.06%	Multiple Regression	Forward Entry	0.01	0.7
2	90.26%	52.82%	Multiple Discriminant Analysis	Mahalanobis Distance	0.1	0.8
10	88.77%	37.91%	Multiple Discriminant Analysis	Wilks' Lambda	0.15	0.8
17	88.76%	44.78%	Multiple Regression	Backward Elimination	0.02	0.9
6	89.65%	41.52%	Multiple Discriminant Analysis	Unexplained Variance	0.14	0.9

Abort Suspend  All  Running 750 of 1783 possible runs 750 Update

- The Overall Correct Classification Percentage (OCCP) tab shows the “lift” above chance alone. Chance is calculated by taking the sum of the squares of the number of cases in each group, divided by the square of the total number of cases.  $Lift = (100\% - OCCP) / (100\% - Chance)$
- In the bottom half run statistics are presented. These statistics can be sorted by any column by clicking on the column heading. Use the scroll bar to view all of the statistics. This list of statistics is also shown sorted by lift in the Execution Summary dialog when the runs have been completed.



## Execute Data Mining

### Execute, cont.

- The image on the right shows the Best Run dialog from the Runtime Information Center. This dialog displays detailed results for one run.
- The best run tends to change frequently during the beginning of the run, then less frequently as time passes. To change the criteria used to select the best run, use the mouse to click the arrow in the “Criteria for best run” box.
- Changing the criteria used to select the best run will also change the run that is displayed in the Best Model box at the top of the Runtime Information Center. This change will be reflected on all dialogs of the Runtime Information Center.

**Runtime Information Center**

Elapsed Time: 00:01:29  
 Estimated Time Remaining: 00:01:22  
 Runs Completed: 3 of 30  
 Servers Connected: 1  
 Progress:  **10%**

**Best Model**  
 Algorithm: Multiple Discriminant Analysis  
 Method: Smallest F-Ratio  
 Criteria: 0.10  
 Transformation: 0.80  
 OCCP: 96.94% Details...

Server Information  
 Communications  
 Performance  
 Best Run  
 Event Log

Criteria for best run: OCCP

**Run Settings**  
 Algorithm: Multiple Discriminant Analysis  
 Method: Smallest F-Ratio  
 Criteria: 0.10  
 Transformations: 0.80

**Performance Comparison**  
 OCCP: 96.94%  
 Lift: 74.90%  
 L1+L2: 0.00%  
 A2\*P2: 0.00%

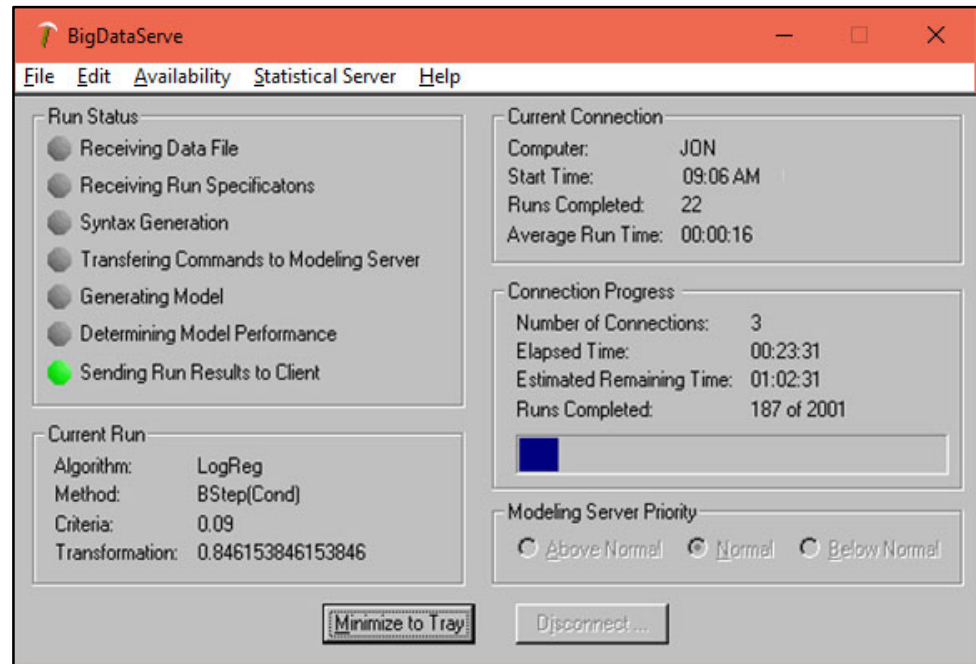
**Additional Run Info**  
 Server: KRISTI  
 Run Time: 00:00:04  
 Holdback size: 1505  
 Model Size: 495

Abort Suspend All Running 30 of 3113 possible runs 0 Update

## Execute Data Mining

### ■ Using the BigDataSolve™ Server

- Some terminology is helpful here.
  - The main application BigDataSolve™, is called the Root Server, because it is the root application that sends all the commands out and receives back results. The Root Server itself does not create models.
  - The BigDataSolve™ Server(s) are small, efficient applications (see the image to the right) on the network that “do the math”



- To generate models using the BigDataSolve™ Root Server, there must be at least one BigDataSolve™ server open.
- There are two versions of the BigDataSolve™ Root Server, BigDataSolve™ and BigDataSolve/DP™. BigDataSolve/DP™ allows only one server. **BigDataSolve/DP™ is the distributed computing superset of BigDataSolve™ – it allows up to 1,000 servers.**
- If the BigDataSolve™ server is open and connected through a network to the computer with the BigDataSolve™ Root Server, the Root Server will automatically locate the server.
- When the “Execute” command is run from the Data Mining Menu of BigDataSolve™, BigDataSolve™ will begin generating models with all available servers.

## Execute Data Mining

### ■ Scheduling the BigDataSolve™ Server

- Availability of the server can be customized by selecting 'Advanced' from the Availability menu on the server. In the top portion of the dialog, a specific time can be specified to allow a connection. To set up a schedule of availability, use the Custom option, and the bottom section of the dialog will become available.
- *Either* the times to allow a connection or the times to block a connection are shown. Select the Allow Connection or Don't Allow Connection options to change which options are viewed. Select one or more days and a range of times, then select the Add button to add an additional set of criteria for the server availability.

When to Allow Connection

Always  
 Never  
 At 10:26 AM  
 In 0 hours 0 minutes  
 Custom

Custom Connection Schedule

Day	Start Time	Stop Time
+ Weekdays	5:50 PM	5:50 AM NextDay
+ Weekends	All Day	All Day

Sunday  Monday  Tuesday  Wednesday  Thursday  Friday  Saturday  
 Weekdays  Weekends

Between 5:50 PM and 5:50 AM  Next Day  Allow Connect  
 All Day  Don't Allow Connection

## Execute Data Mining

### ■ Execution Summary

- When a run has been completed, the BigDataSolve™ Execution Summary will be displayed.
- This information is sorted in descending order of Lift, which puts the best runs at the top.
- The Run # column shows the order that the runs were completed. The SPS and SPO files are saved for each run and named by the run number, so if you want to rerun a specific syntax file, or inspect the output later, the run number will be necessary.
- OCCP stands for Overall Correct Classification Percentage: the percentage of the holdback sample that was correctly classified.
- Lift is the correct classification percentage **above** chance alone. This is determined by comparing the OCCP to the percentage achieved by chance alone.
- Transformation refers to the power to which the predictors were raised. The range and steps for transformation can be set in the Define Parameters step of the Rules wizard, by using the Advanced button.
- The specifications of each run are listed, so that algorithms that performed well can be explored further for the current data set.

BigDataSolve Execution Summary

Run #	Server	OCCP	Lift	Algorithm	Method	Criteria	Transformation
0	KRISTI	93.81	46.7	Multiple Discriminant Analysis	Rao's V	0.03	1.32
30	JON	93.81	46.7	Multiple Discriminant Analysis	Rao's V	0.03	1.34
27	KRISTI	93.75	46.67	Multiple Discriminant Analysis	Mahalanobis Distance	0.09	0.98
8	KRISTI	93.65	46.61	Multiple Discriminant Analysis	Mahalanobis Distance	0.04	1.36
41	JON	93.65	46.61	Multiple Discriminant Analysis	Mahalanobis Distance	0.09	0.66
24	KRISTI	93.64	46.6	Multiple Discriminant Analysis	Rao's V	0.08	1.3
65	JON	93.64	46.6	Multiple Discriminant Analysis	Rao's V	0.08	1.4
33	KRISTI	93.63	46.6	Multiple Discriminant Analysis	Wilks' Lambda	0.11	0.62
73	JON	93.63	46.6	Multiple Discriminant Analysis	Wilks' Lambda	0.12	0.72
36	KRISTI	93.61	46.58	Multiple Discriminant Analysis	Wilks' Lambda	0.03	0.74
20	KRISTI	93.59	46.57	Multiple Discriminant Analysis	Smallest F-Ratio	0.14	1.4
59	JON	93.59	46.57	Logistic Regression	Backward Stepwise (Conditional Statistic)	0.11	1.3
1	KRISTI	93.59	46.57	Multiple Discriminant Analysis	Wilks' Lambda	0.18	0.82
31	JON	93.59	46.57	Multiple Discriminant Analysis	Mahalanobis Distance	0.05	0.52
2	KRISTI	93.53	46.54	Multiple Discriminant Analysis	Mahalanobis Distance	0.11	1.32
32	JON	93.53	46.54	Multiple Discriminant Analysis	Mahalanobis Distance	0.12	1.26
42	KRISTI	93.5	46.52	Multiple Discriminant Analysis	Smallest F-Ratio	0.17	1.1
77	JON	93.5	46.52	Logistic Regression	Backward Stepwise (Conditional Statistic)	0.05	0.5
17	KRISTI	93.49	46.52	Logistic Regression	Forward Stepwise (Conditional Statistic)	0.1	0.6
55	JON	93.49	46.52	Logistic Regression	Backward Stepwise (Likelihood Ratio)	0.04	0.8
5	KRISTI	93.47	46.51	Multiple Discriminant Analysis	Smallest F-Ratio	0.15	0.8
37	JON	93.47	46.51	Multiple Discriminant Analysis	Smallest F-Ratio	0.17	1.22
6	KRISTI	93.46	46.5	Multiple Discriminant Analysis	Smallest F-Ratio	0.14	0.96

Done

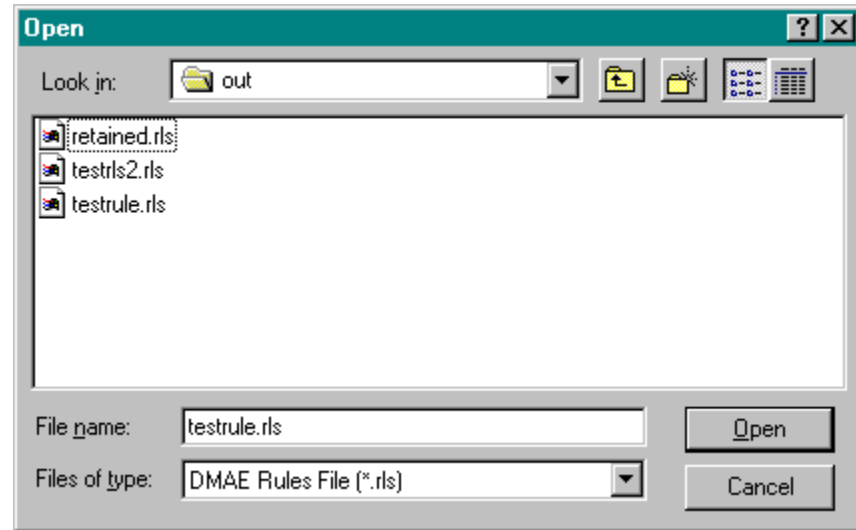
---

### III. Appendix - Reference Manual

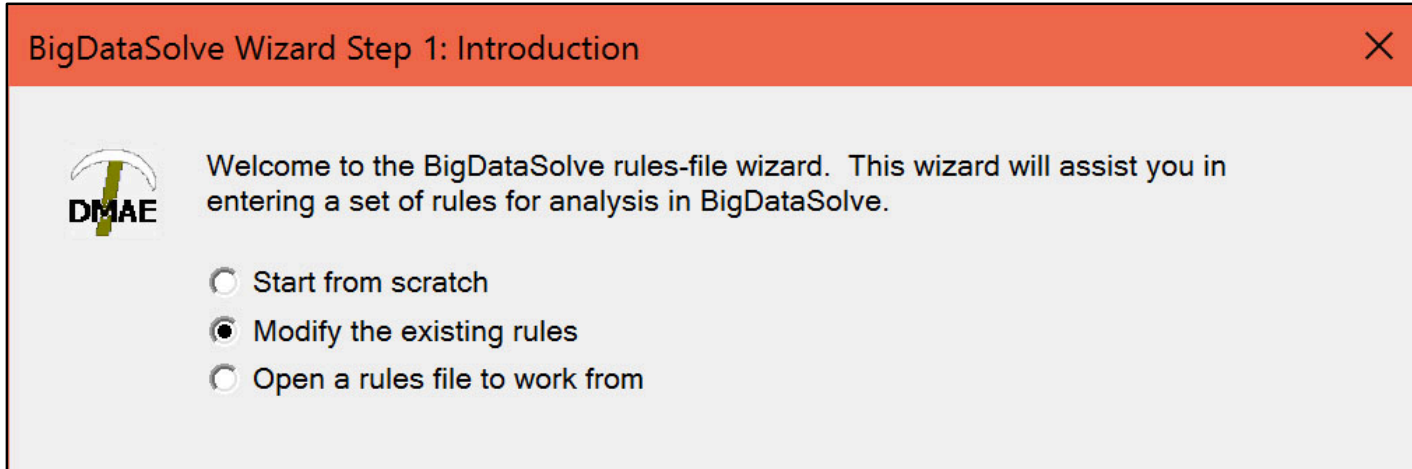


## 1. File: Open Rules File

- A rules file stores all of the files, options, and variables that have been selected while using BigDataSolve™. A rules file must be opened or created to begin the analysis.
- To Create a rules file, select New Rules from the File menu before you start the analysis.
- To use an existing rules file, select Open Rules from the File menu.
- This dialog box will be opened that allows you to browse through all directories. When the desired file has been selected, click the Open button.
- If the Cancel button is selected, BigDataSolve™ will exit this dialog box without opening a rules file.



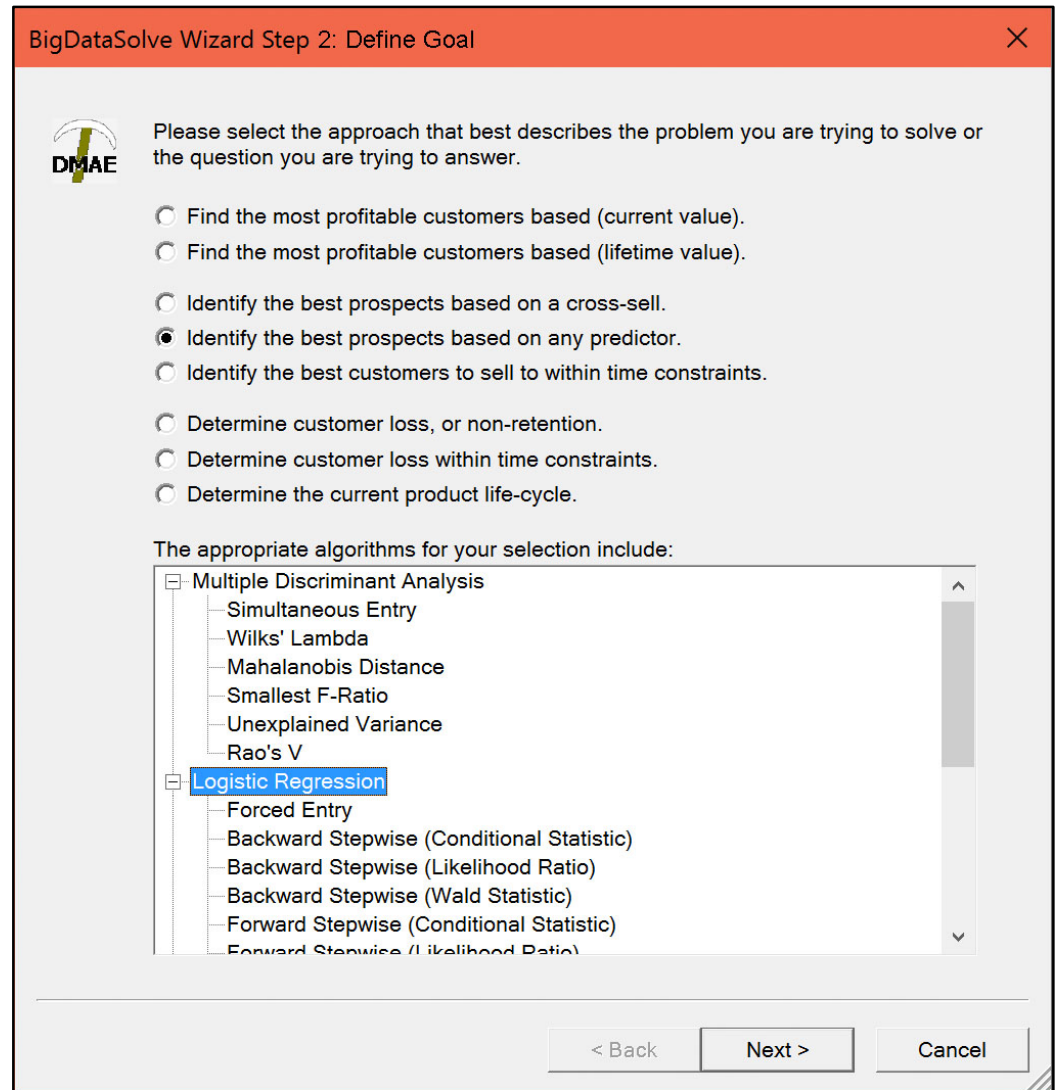
## 2. Rules Wizard Step 1: Introduction



- The rules file is the easiest and most efficient way to step through all of the customizable options for the data mining.
- To begin the rules wizard, either select the pencil icon from the toolbar, or “Start Rules Wizard” from the Rules submenu of the Data Mining menu. The rules wizard will present options on each screen that must be completed to execute the data mining analysis. Make your selection, then click on the Next button.
- At any point beyond the second step, you can also go back to the previous rules wizard pages to modify your selections by using the Back button.
- If the Cancel button is selected, BigDataSolve™ will exit the wizard and cancel any changes made.

### 3. Rules Wizard Step 2: Define Goal

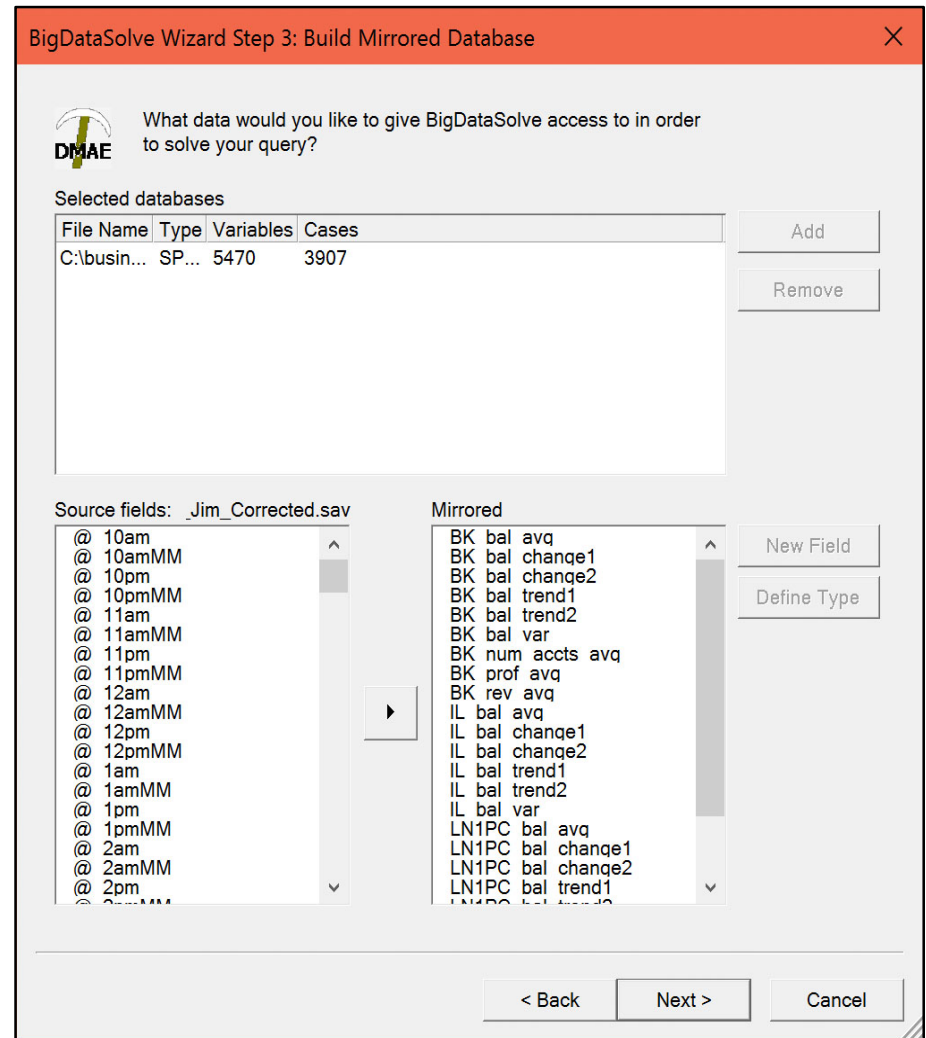
- This dialog allows you to select the desired business problem, and displays the appropriate algorithms for that problem.
- Select the desired business problem. Algorithms will be refreshed to show only those that would be used to determine solutions.
- When you are satisfied with your selection, click on the Next button.
- If the Cancel button is selected, BigDataSolve™ will exit this dialog and cancel any changes made.





## 4. Rules Wizard Step 3: Build Mirrored Database

- In this step, the database to be used for the analysis is selected and imported.
- To open a database, select the Add button. An Open File dialog will appear. Select a file and click on the Open button, and it will be imported into BigDataSolve™.
- If a database name was already selected for the current rules file, but has the wrong path or file name, use the Remove button to delete the name and reselect it.
- To select the variables to be used from a database, highlight the database name on the wizard screen. The available variables will appear in the Source Fields box. Highlight any number of fields from the Source Fields box, then click on the arrow to move the fields to the Mirrored fields box. The mirrored fields are those that will be included in the data mining execution.
- To select fields from a different database, highlight the database name, and select the fields to mirror.
- Select the Next box to move to the next screen.



## 5. Rules Wizard Step 4: Select Variables

- Once the variables from all of the databases have been selected, the fields are divided into predictor and predicted fields.
- Select the variables in the Mirrored fields box, then use the arrows to move the fields to either the “Fields to predict” or the “Fields to be used in prediction” boxes.
- There should only be one field to predict, and at least two predictor variables.
- When the desired fields have been selected, click on the Next button.

BigDataSolve Wizard Step 4: Select Variables

Which data fields would you like BigDataSolve to schedule for prediction, and which fields should be used to predict those variables?

Mirrored

Field to Predict  
TotalAssetDollars

Fields to be used in prediction

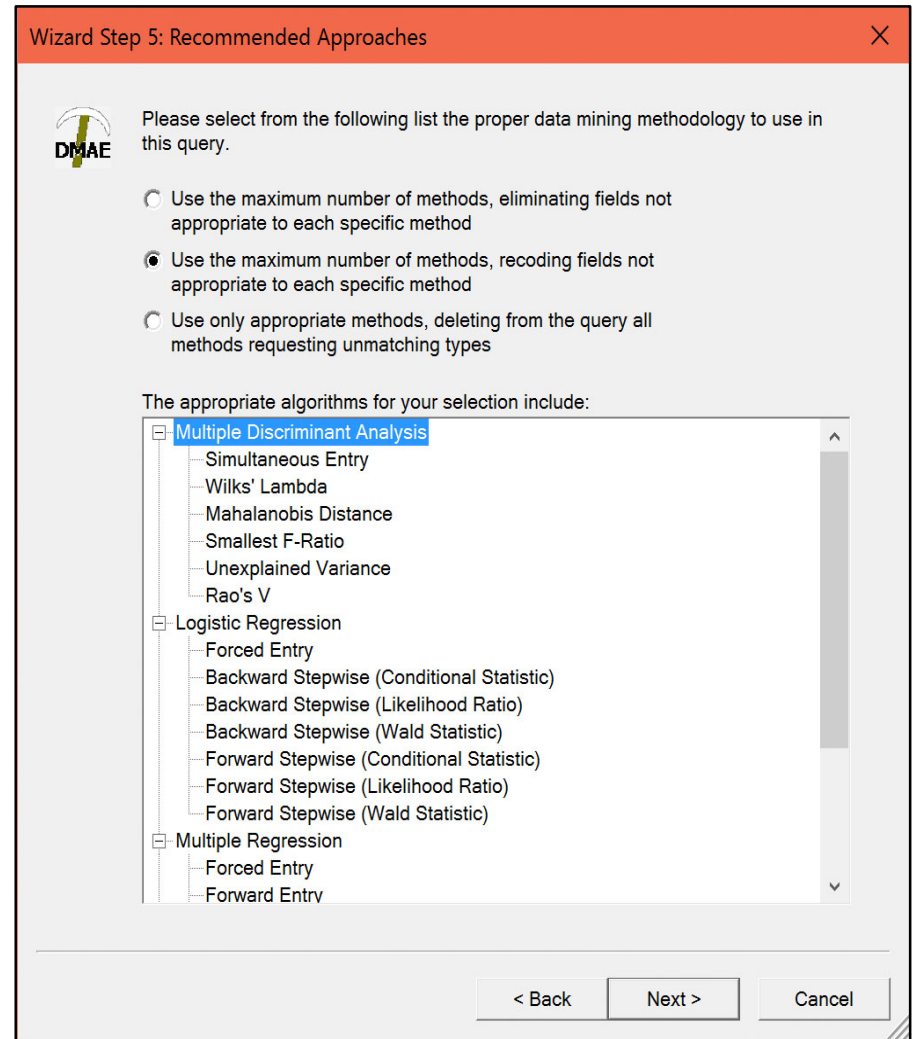
- BK bal avq
- BK bal change1
- BK bal change2
- BK bal trend1
- BK bal trend2
- BK bal var
- BK num accts avq
- BK prof avq
- BK rev avq
- IL bal avq
- IL bal change1
- IL bal change2
- IL bal trend1
- IL bal trend2
- IL bal var
- LN1PC bal avq
- LN1PC bal change1
- LN1PC bal change2
- LN1PC bal trend1
- LN1PC bal trend2
- LN1PC bal var
- LN1PC num accts avq

Continuous Dependent Variable  
Classification is correct if  
5 or 10 %  
of actual value, whichever is greater

< Back Next > Cancel

## 6. Rules Wizard Step 5: Recommended Approaches

- There may be some data fields that are not compatible with an algorithm that will be used in the analysis. This step allows the user to choose how BigDataSolve™ will handle the incompatible fields. There are three options:
  - The first option leaves the incompatible data field out of the particular method for which it is inappropriate.
  - The second option attempts to use all of the selected data fields. Data that is not compatible with a specific method will be recoded (i.e. a continuous variable will be transformed using exponents to be used as a discrete variable).
  - The third option eliminates the analysis method, rather than the data field, if there is incompatible data.
- Select the option by clicking on it, then select the Next button.



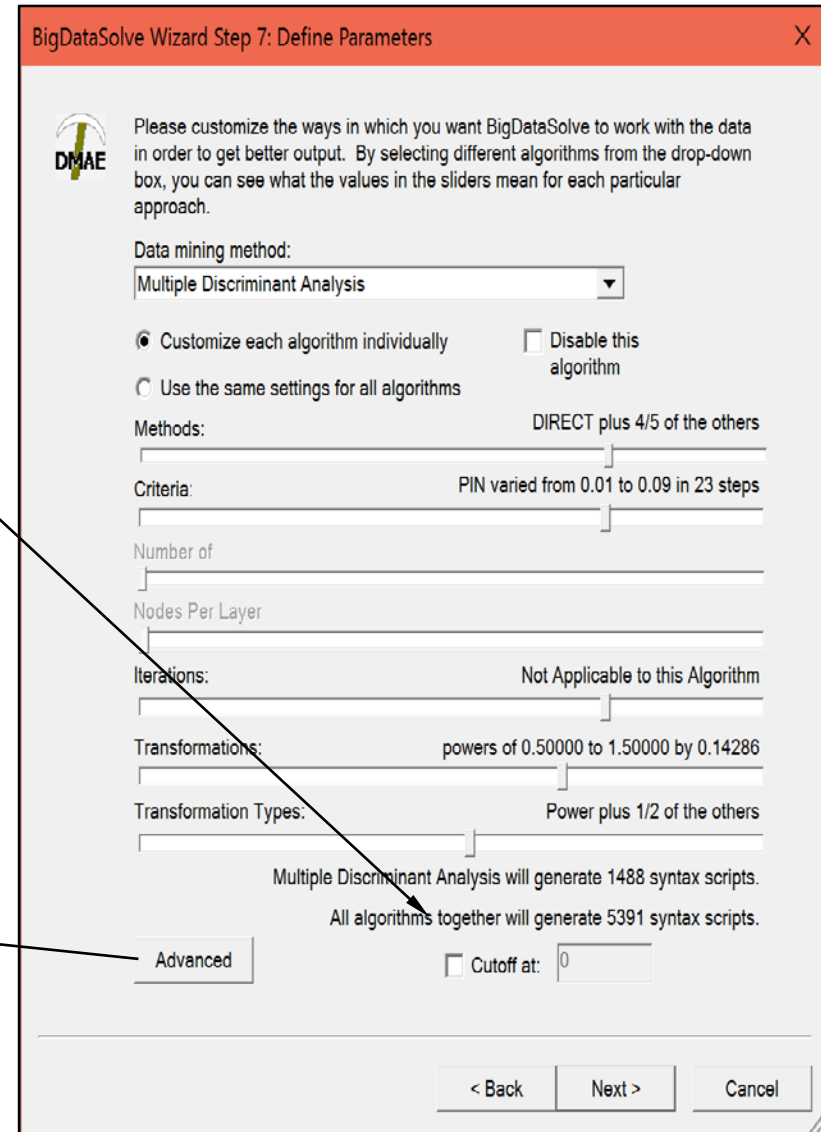
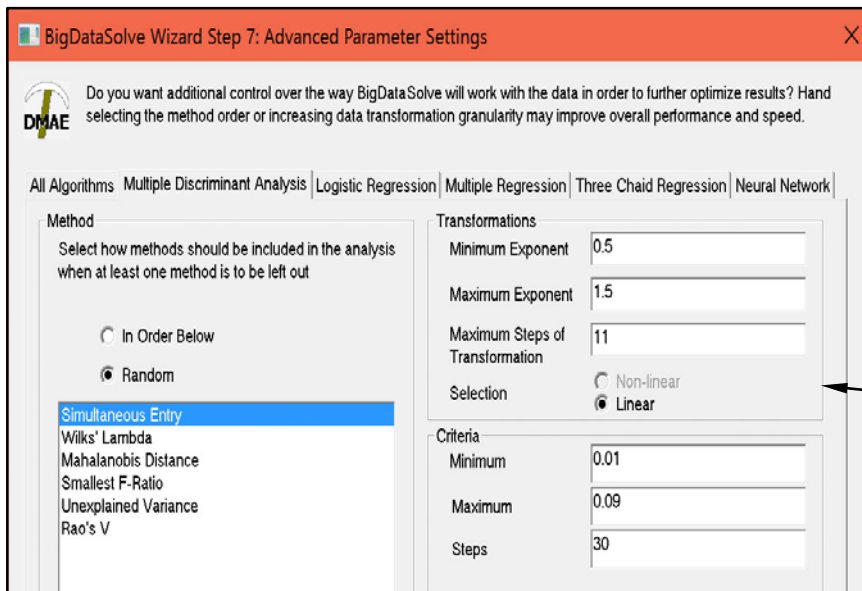
## 7. Rules Wizard Step 6: Select Cases for Inclusion

- The data file can be filtered based on segment variables. Multiple segments can be selected so that for example, only people with a loan account who have been with the bank for more than 10 years are included in the run.
- Missing data is data that is either system missing or marked as missing in SPSS. Select one of the options to change how missing data is handled by BigDataSolve™.
- Missing data can be handled differently for discrete or continuous variables.
- If segments have been selected, the Working File Size reflects the number of cases included in the selected segments. Use the slider to change the selected number of cases to use from the working data file and for the holdback sample.

The screenshot shows the 'BigDataSolve Wizard Step 6: Select Cases for Inclusion' dialog box. It features a title bar with a close button (X) and a logo on the left. The main text asks: 'You can further clarify the model BigDataSolve will use by selecting specific respondents by either region or segment?'. There are two radio button options: 'Select respondents by region' (checked) with a 'Map...' button, and 'Select respondents by market segment' (unchecked) with a 'Segments...' button. Below this is a section for 'Missing Data Replacement Methodology' with two sub-sections: 'Discrete Data' and 'Continuous Data'. Each sub-section has three radio button options: 'Eliminate Cases' (selected), 'Replace With Median (For Two Group Variables Only)', and 'Replace Using Multiple Discriminant Analysis' (for discrete) or 'Replace with Mean', 'Replace Using Multiple Regression', and 'Replace Using Maximum Likelihood Estimation' (for continuous). At the bottom, there is a 'Sample Size (3672 Valid Cases)' section with two rows: 'Working File Size' showing a value of 3672 (100%) and a slider; and 'Holdback Sample (75% recommended):' showing a value of 2754 (75.00%) and a slider. At the very bottom are three buttons: '< Back', 'Next >', and 'Cancel'.

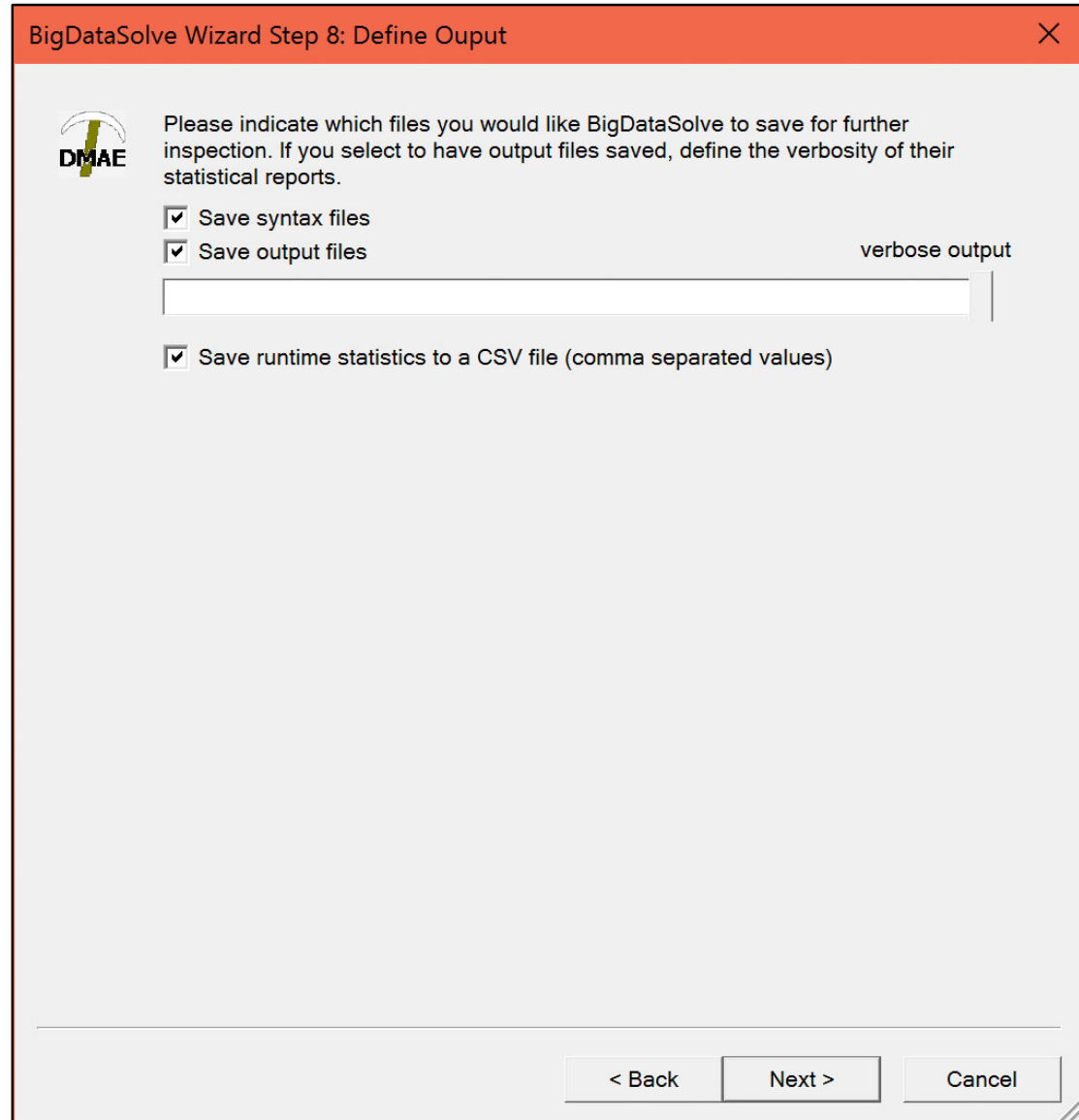
## 8. Rules Wizard Step 7: Define Parameters

- All of the methods and criteria used for the analysis can be customized using this dialog. Select one of the Data Mining Methods to customize the options for that method. Use the advanced button at the bottom to set specific exponent parameters or change the order of the methods to be used within each algorithm.
- The number of model runs generated given the current settings is displayed at the bottom of the dialog. A random selection of these possible scripts can be selected to be run by checking the “Cutoff at” box and entering a number of scripts to be run.



## 9. Rules Wizard Step 8: Define Output


- Syntax and Output files that are saved are saved in the working directory specified in Step 9 of the wizard.
- A folder is created in the working directory for each of the selected methods, and the syntax and output files are saved in the subdirectories.
- The runtime statistics are saved in a CSV file in the working directory.



## 10. Rules Wizard Step 9: Closure

- The rules internal name is not a file name, but a reference for the rules file that is used within BigDataSolve™ and used for a run title. The rules internal name and description are optional.
- The working directory specifies where all of the output from the runs will be saved.

BigDataSolve Wizard Step 9: Closure


 Congratulations! You have finished defining a set of operating rules for BigDataSolve. Please choose a name and an optional short description of this set. Then give this definition a directory into which temporary files should be written.

Rules Internal Name:

Description:

Working Directory:  
 ...

Save Rules on Finish  
 Launch Now



< Back   Finish   Cancel