

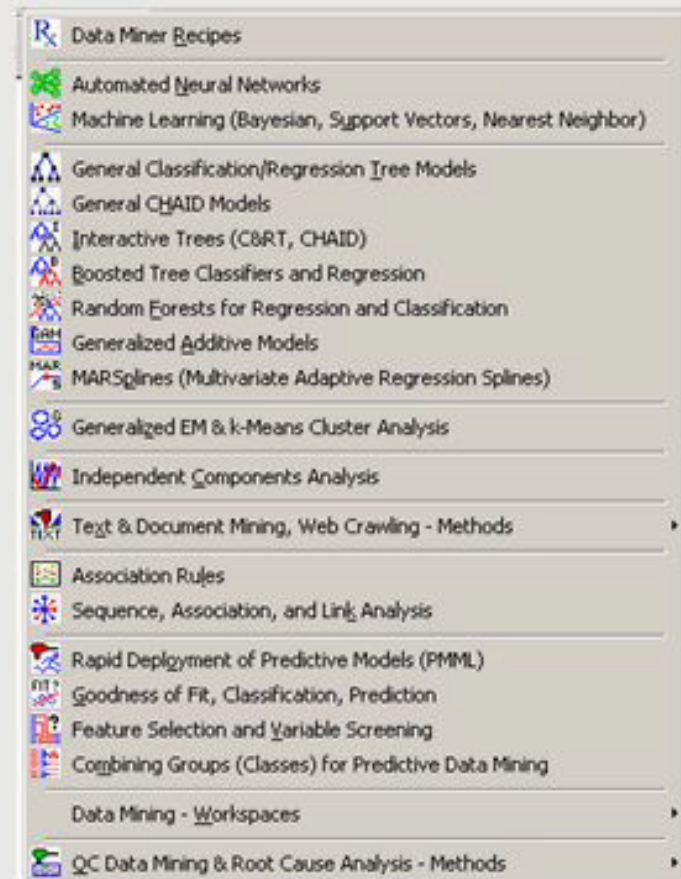
Introduction to Data Mining




















Knowledge Discovery vs. Statistical Analysis

- **Statistical Analysis**
 - Focuses on “hypothesis testing” and “parameter estimation”
 - Fits “parsimonious statistical models” with the goal to “explain” complex relationships with fewer parameters
 - Examples: Regression, nonparametric statistics, factor analysis, traditional quality control
- **Data Mining**
 - Focuses on knowledge discovery, detection of patterns, clusters, and so on; we only have data and no (or few) expectations and hypotheses
 - Fits simple models (such as regression) or complex models (such as neural nets) to enable valid prediction
 - Examples: Neural nets, stochastic gradient boosting of tree classifiers, random forests, support vector machines

STATISTICA Data Miner

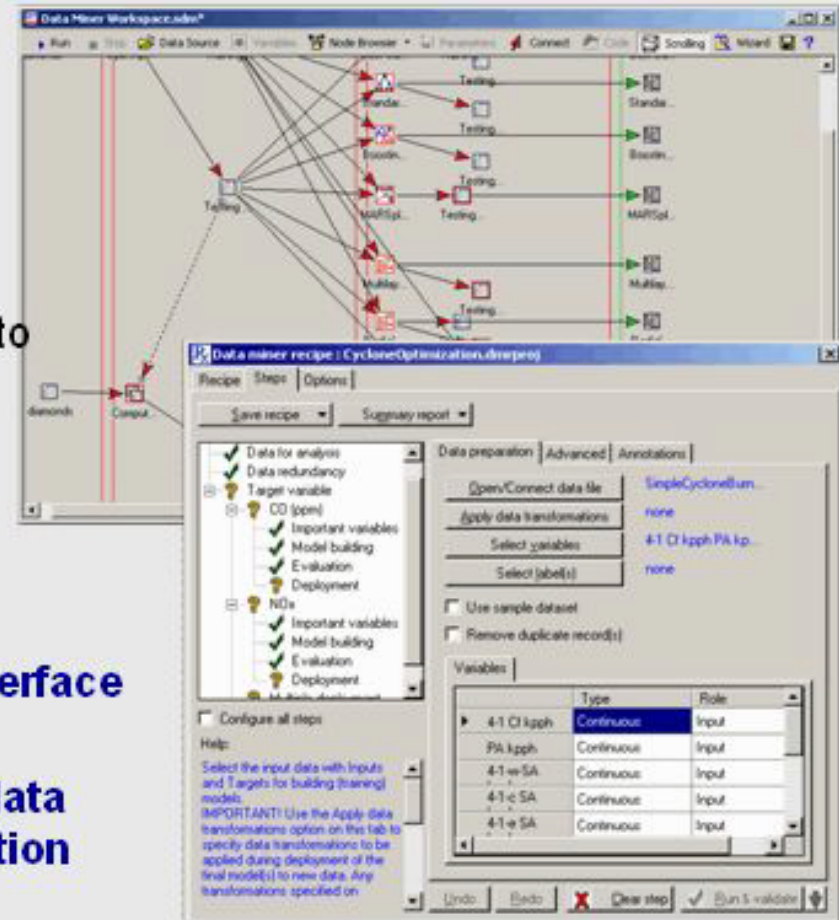
- **Most comprehensive collection of data mining algorithms in a single package**
 - Automatic Feature Selection
 - Automated Neural Network Problem Solver
 - Classification & Regression Trees
 - CHAID Trees (recursive partitioning)
 - Stochastic Gradient Boosted Trees
 - Voting of Trees (Forests)
 - MAR (Regression) Splines
 - Support Vector Machines
 - k-Nearest Neighbors
 - Naive Bayes Classifiers
 - Generalized Additive Models
 - k-Means and EM Clustering
 - Association/Sequence Rules
 - (continuously updated)

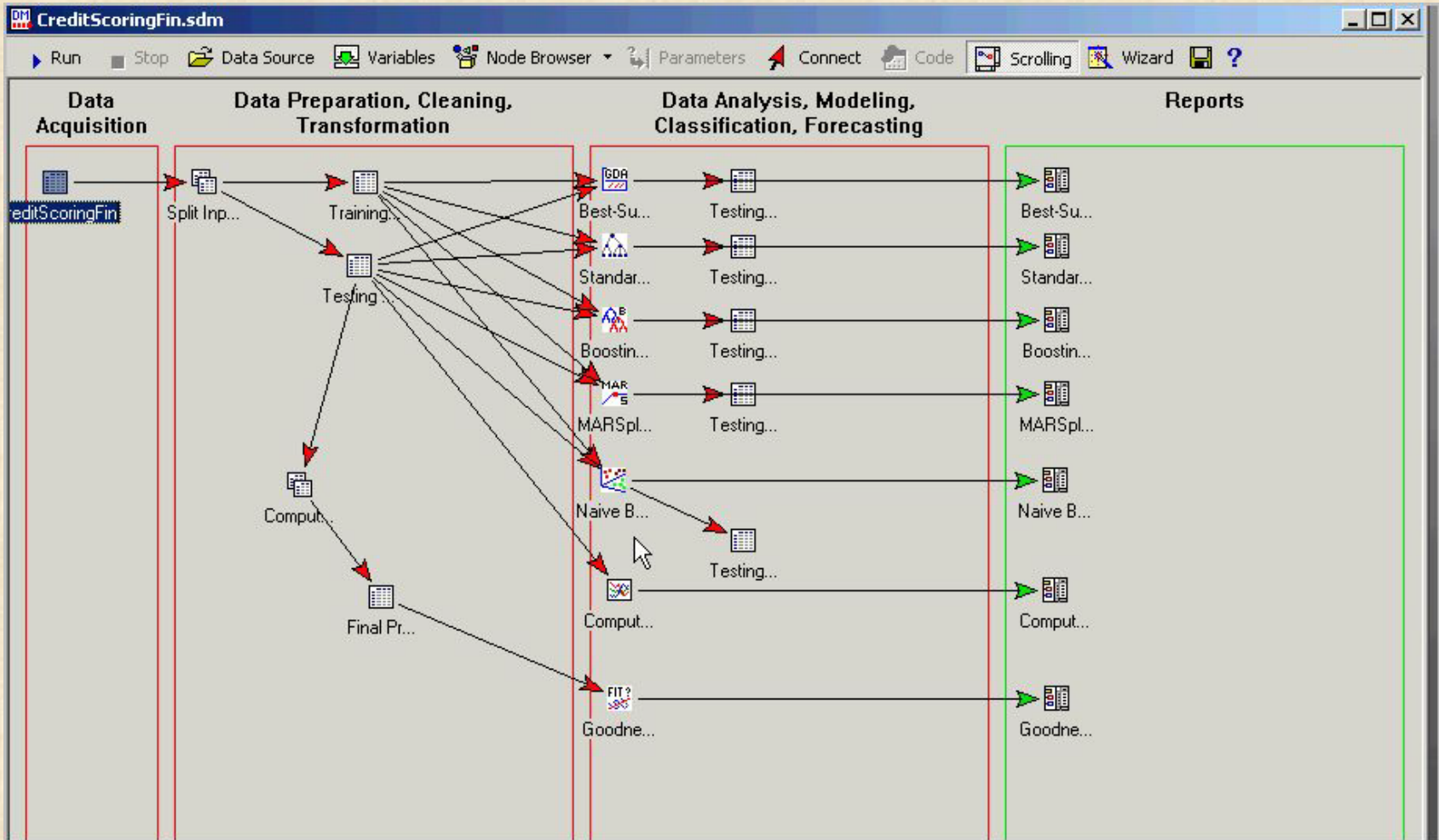


	Data Miner Recipes
	Automated Neural Networks
	Machine Learning (Bayesian, Support Vectors, Nearest Neighbor)
	General Classification/Regression Tree Models
	General CHAID Models
	Interactive Trees (C&RT, CHAID)
	Boosted Tree Classifiers and Regression
	Random Forests for Regression and Classification
	Generalized Additive Models
	MARSplines (Multivariate Adaptive Regression Splines)
	Generalized EM & k-Means Cluster Analysis
	Independent Components Analysis
	Text & Document Mining, Web Crawling - Methods
	Association Rules
	Sequence, Association, and Link Analysis
	Rapid Deployment of Predictive Models (PMML)
	Goodness of Fit, Classification, Prediction
	Feature Selection and Variable Screening
	Combining Groups (Classes) for Predictive Data Mining
	Data Mining - Workspaces
	QC Data Mining & Root Cause Analysis - Methods

Data Miner Workspaces, Recipes

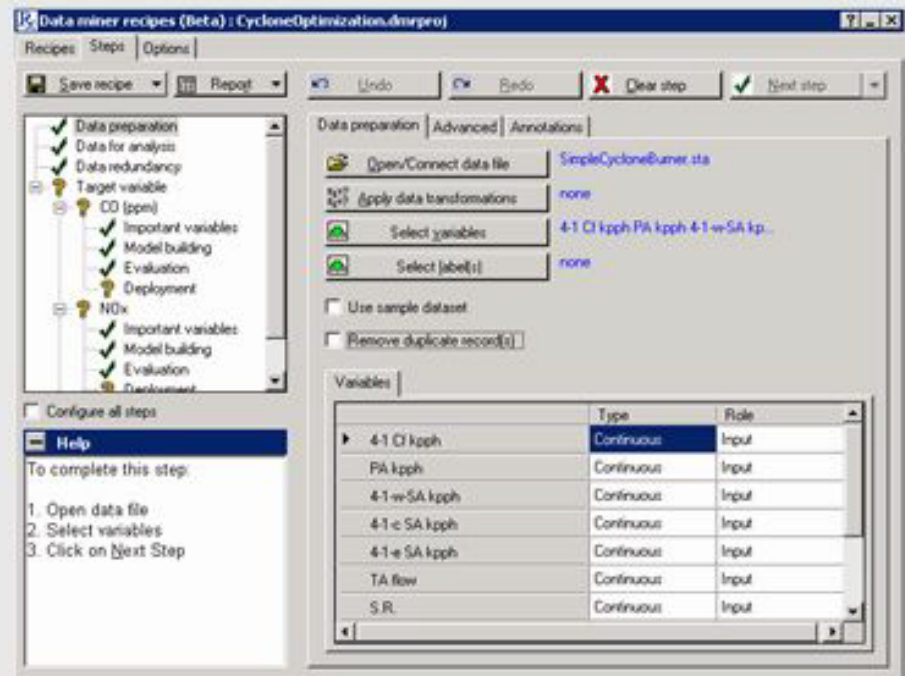
- Efficient UI solutions for effective data mining
- Data Miner Workspaces
 - Custom data mining workflows or templates (best practices)
 - Use *all* STATISTICA functionality to build work flows that go beyond traditional data mining (e.g., perform Predictive QC-Mining,...)
- Run on desktop, offload to server
- Data Miner Recipes
 - **Revolutionary, efficient user interface for both novices and experts**
 - **Single-click data mining, from data definition through model validation and deployment/ scoring**





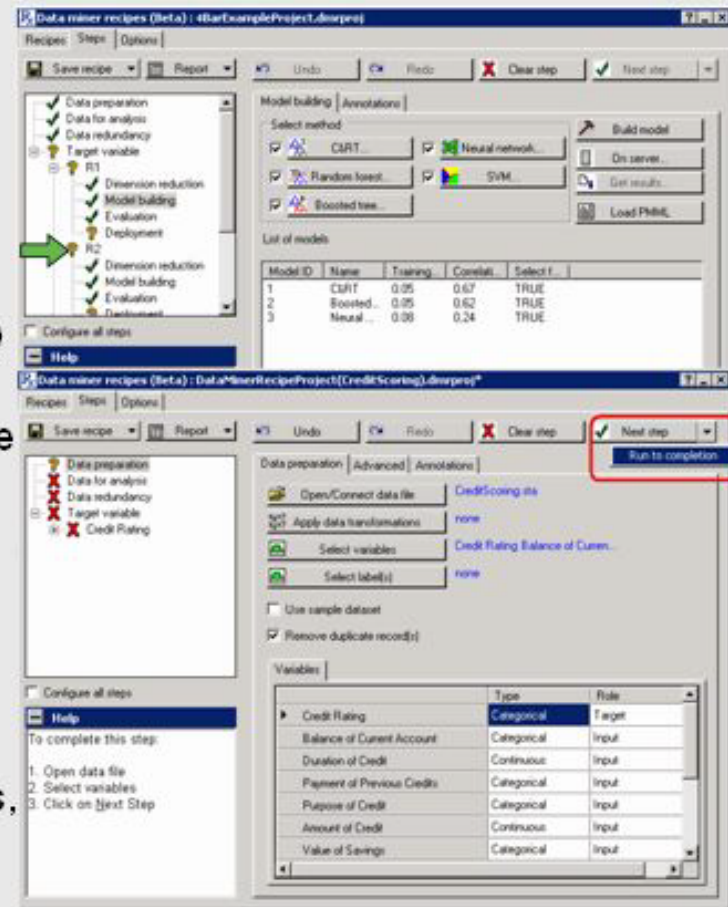
Data Miner Recipe *STATISTICA DMR*

- Goals:
 - Make a Very Simplified UI
 - Make data mining quick and easy
 - Make data mining simple and fast
 - One-click data mining
 - Integration with *STATISTICA* platform
 - Easy scoring of new data
- Fool-proof until better fools are invented



DMR: Make Data Mining Quick and Easy

- The "recipe" metaphor: There are a limited number of "recipes"
 - Predictive data mining for a continuous response (regression)
 - Predictive data mining for a categorical response (classification)
 - Predictive data mining for a rare (continuous or categorical) response
- All of these typical workflows can be accomplished using *STATISTICA Data Miner Recipes*
 - And often with a single click!
- Learning how to apply advanced data mining methods to solve real-world problems is now reduced to a few hours, at most



DMR Example: Predicting Credit Risk

- Example:
 - 100,000 records and 19 variables
- Select variables for predictive modeling
- Click run-to-completion
- An excellent prediction model will appear in under 5 minutes
- Models can be deployed to *STATISTICA Enterprise* for automatic scoring for new data

