

# A Comparison of Leading Data Mining Tools

*John F. Elder IV & Dean W. Abbott*  
*Elder Research*

Fourth International Conference  
on Knowledge Discovery & Data Mining

Friday, August 28, 1998

New York, New York

KDD-98: A Comparison of Leading Data Mining Tools

Copyright © 1998,  
John F. Elder IV and Dean W. Abbott

All rights reserved

Manufactured in the United States of America

# Contacting Elder Research

<http://www.datamininglab.com>

Dr. John F. Elder IV  
1006 Wildmere Place  
Charlottesville, VA 22901

elder@datamininglab.com  
804-973-7673  
Fax: 804-995-0064

Dean W. Abbott  
3443 Villanova Avenue  
San Diego, CA 92122-2310

dean@datamininglab.com  
619-450-0313

# Tutorial Goals

- Compare and Summarize Data Mining Tools which:
  - Offer multiple modeling and classification algorithms
  - Support project stages surrounding model construction
  - Stand alone
  - Are general-purpose
  - Cost a lot
  - We could get our hands on
- Include some (focused) Desktop Tools

Other Reports: Two Crows, Aberdeen Group, Elder Research (forthcoming ), Data Mining Journal

# Topics

- Products covered
- Review of algorithms
- Comparative tables of properties
- Screen shots exemplifying qualities
- Summary of distinctives

## Caveats

- We don't know *every* tool well (and are sure to have missed some!)
  - Level of exposure noted for each tool
- Our background (biasing our perspective)
  - Very technical, “early adopters”
  - Emphasize solving real-world applications
  - More classification than estimation
- Field of tools is quite dynamic
  - New versions appear regularly

# Data Mining Products



# Tools Evaluated

Product	Company	URL	Version Tested	Our Experience
<i>Clementine</i>	Integral Solutions, Ltd.	<a href="http://www.isl.co.uk/clem.html">http://www.isl.co.uk/clem.html</a>	4	Moderate
<i>Darwin</i>	Thinking Machines, Corp.	<a href="http://www.think.com/html/products/products.htm">http://www.think.com/html/products/products.htm</a>	3.0.1	Moderate
<i>DataCruncher</i>	DataMind	<a href="http://www.datamindcorp.com">http://www.datamindcorp.com</a>	2.1.1	High
<i>Enterprise Miner</i>	SAS Institute	<a href="http://www.sas.com/software/components/miner.html">http://www.sas.com/software/components/miner.html</a>	Beta	Moderate
<i>GainSmarts</i>	Urban Science	<a href="http://www.urbanscience.com/main/gainpage.htm">http://www.urbanscience.com/main/gainpage.htm</a>	4.0.3	Low
<i>Intelligent Miner</i>	IBM	<a href="http://www.software.ibm.com/data/iminer/">http://www.software.ibm.com/data/iminer/</a>	2	Low
<i>MineSet</i>	Silicon Graphics, Inc.	<a href="http://www.sgi.com/Products/software/MineSet/">http://www.sgi.com/Products/software/MineSet/</a>	2.5	Low
<i>Model 1</i>	Group 1	<a href="http://www.unica-usa.com/model1.htm">http://www.unica-usa.com/model1.htm</a>	3.1	Moderate
<i>ModelQuest</i>	AbTech Corp.	<a href="http://www.abtech.com">http://www.abtech.com</a>	1	Moderate
<i>PRW</i>	Unica Technologies, Inc.	<a href="http://www.unica-usa.com/prodinfo.htm">http://www.unica-usa.com/prodinfo.htm</a>	2.5	High
<i>CART</i>	Salford Systems	<a href="http://www.salford-systems.com">http://www.salford-systems.com</a>	3.5	Moderate
<i>NeuroShell</i>	Ward Systems Group, Inc.	<a href="http://www.wardsystems.com/neuroshe.htm">http://www.wardsystems.com/neuroshe.htm</a>	3	Moderate
<i>OLPARS</i>	PAR Government Systems	<a href="mailto://olpars@partech.com">mailto://olpars@partech.com</a>	8.1	High
<i>Scenario</i>	Cognos	<a href="http://www.cognos.com/busintell/products/index.html">http://www.cognos.com/busintell/products/index.html</a>	2	Moderate
<i>See5</i>	RuleQuest Research	<a href="http://www.rulequest.com/see5-info.html">http://www.rulequest.com/see5-info.html</a>	1.07	Moderate
<i>S-Plus</i>	MathSoft	<a href="http://www.mathsoft.com/splus/">http://www.mathsoft.com/splus/</a>	4	High
<i>WizWhy</i>	WizSoft	<a href="http://www.wizsoft.com/why.html">http://www.wizsoft.com/why.html</a>	1.1	Moderate

# Categories for Comparisons

- Platforms Supported
- Algorithms Included
  - Decision Trees
  - Neural Networks
  - Other
- Data Input and Model Output Options
- Usability Ratings
- Visualization Capabilities
- Modeling Automation Methods

KDD-98: A Comparison of Leading Data Mining Tools

Platforms	PC Standalone (95/NT)	Unix Standalone	Unix Server / PC Client	NT Server / PC Client	Database Connectivity
<i>Clementine</i>	√	√+			√
<i>Darwin</i>			√		√
<i>DataCruncher</i>	√		√		√
<i>Enterprise Miner</i>	√		√+	√	√
<i>GainSmarts</i>	√	√			√
<i>Intelligent Miner</i>			√		√
<i>MineSet</i>		√			√
<i>Model 1</i>	√			√	√
<i>ModelQuest</i>	√	√			√
<i>PRW</i>	√				√
<i>CART</i>	√	√+			
<i>Scenario</i>	√				√
<i>NeuroShell</i>	√				
<i>OLPARS</i>	√	√			
<i>See5</i>	√	√+			
<i>S-Plus</i>	√				√-
<i>WizWhy</i>	√				

Key	
blank	no capability
√-	some capability
√	good capability
√+	excellent capability

# Tool Groupings

## Desktop

- PC (standalone)
- Flat Files
- One or Two Algorithms
- Data Fits into RAM

## High End

- Multiple Platforms, Client-Server
- Flat Files or Direct Database Access
- Multiple Algorithm Types
- Large Databases

# End User Perspectives

## Business

- Intuitive Interface
  - Clear steps in data mining process
  - Non-technical terminology
  - Familiar environment
- Descriptive Reporting
  - Domain terminology
  - Graphical representations

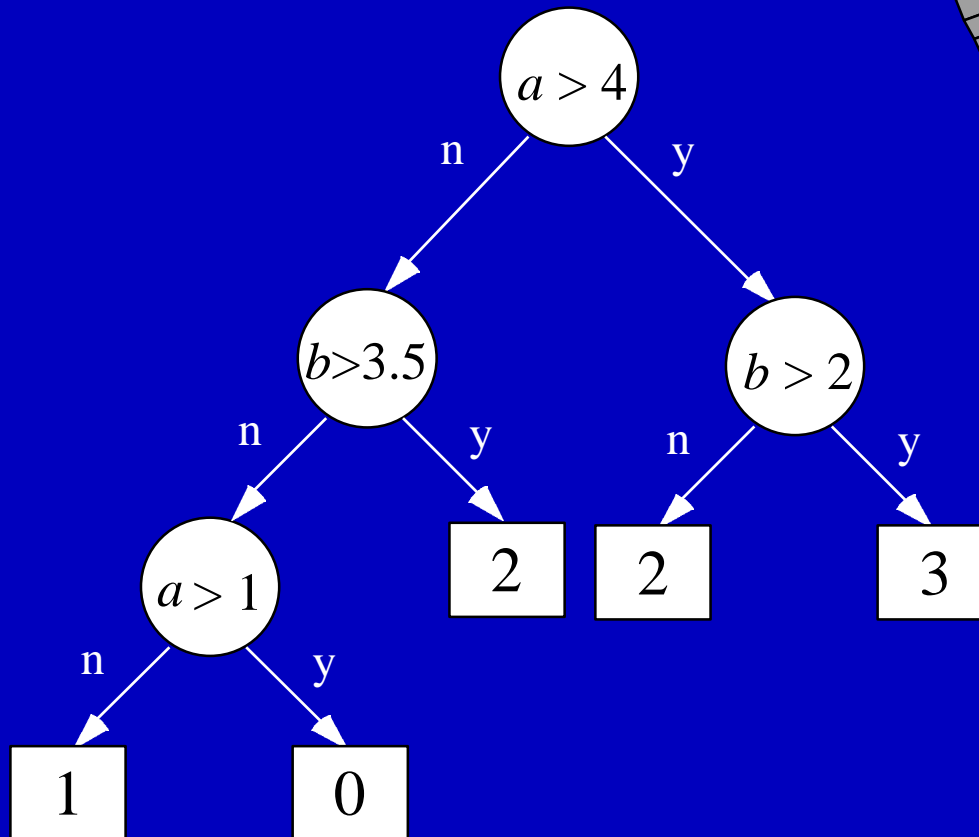
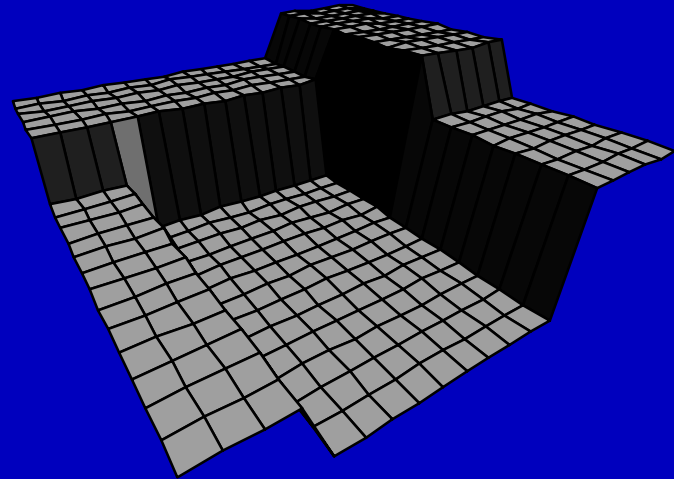
## Technical

- Algorithm Options
  - Knobs to enhance model performance
- Model Automation
  - Simplify model design cycle
  - Documentation of steps used in generating models (repeatability)

KDD-98: A Comparison of Leading Data Mining Tools

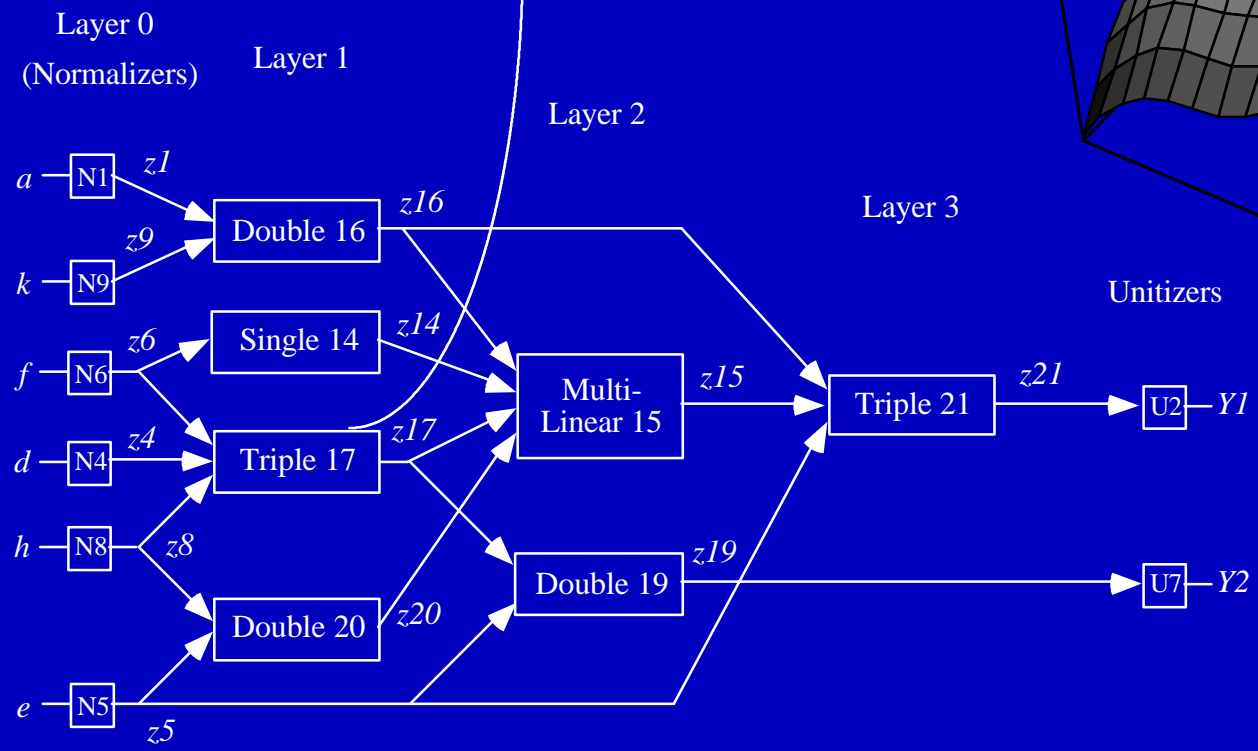
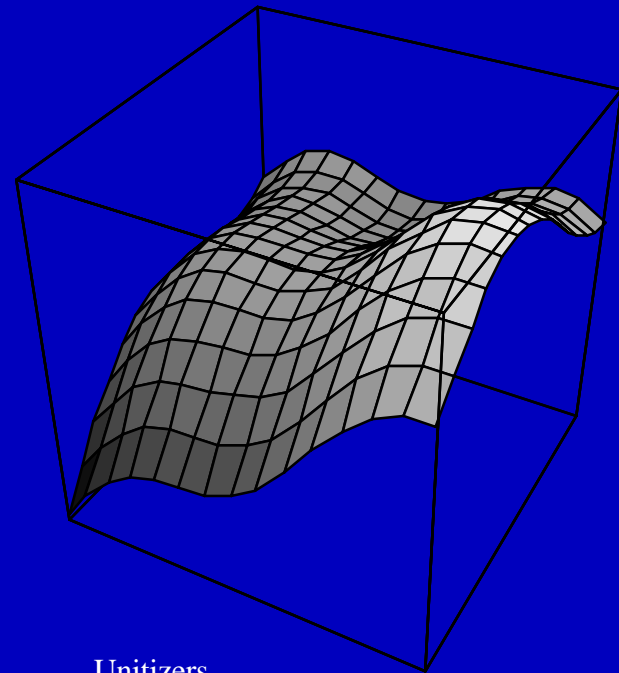
<b>Data Input &amp; Model Output</b>	<b>Automatic Header</b>	<b>Save Data Format</b>	<b>ODBC</b>	<b>Native Database Drivers</b>	<b>Summary Reports</b>	<b>Output Source Code</b>
<i>Clementine</i>	√		√			√
<i>Darwin</i>		√	√			√
<i>DataCruncher</i>	√	√	√	√	√	
<i>Enterprise Miner</i>	√-		√	√	√-	√
<i>GainSmarts</i>	√	√		√	√	√
<i>Intelligent Miner</i>				√-		√
<i>MineSet</i>		√		√		
<i>Model 1</i>	√	√	√		√	√
<i>ModelQuest</i>	√			√	√	√
<i>PRW</i>	√	√	√			√
<i>CART</i>	√					
<i>Scenario</i>	√				√	
<i>NeuroShell</i>	√					
<i>OLPARS</i>		√				
<i>See5</i>	√-					
<i>S-Plus</i>	√		√		√	√
<i>WizWhy</i>	√				√	

# Decision Trees



# Polynomial Networks

$$Z_{17} = 3.1 + 0.4a - .15b^2 + 0.9bc - 0.62abc + 0.5c^3$$

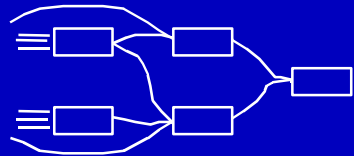
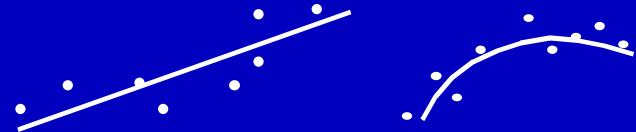


# “Consensus” Models

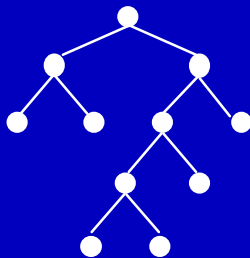
## Parametrically Summarize Data Points

orders, terms

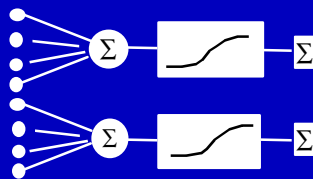
Regression



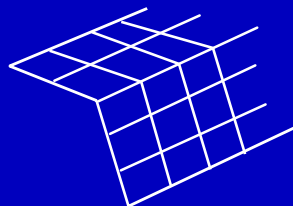
Polynomial Networks  
(e.g. GMDH, ASPN)



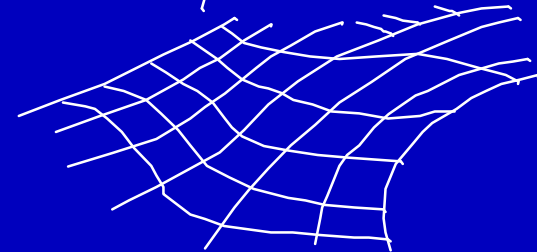
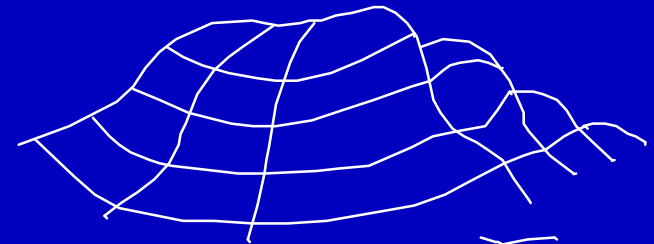
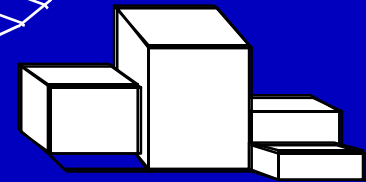
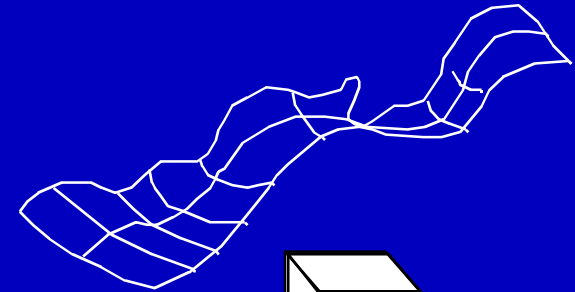
Decision Trees  
(e.g., CART, CHAID, C5)



Logistic or Sigmoidal  
Networks (ANNs)



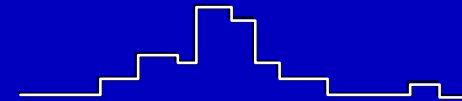
Hinging Hyperplanes,  
MARS



## “Consensus” Models (continued)

orientation, bin width

Histogram



function 

Radial Basis Function



 family, order

Wavelets



# “Contributory” Models

retain data points; each potentially affects estimate at new point

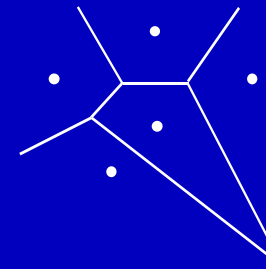
shape, spread

Kernels



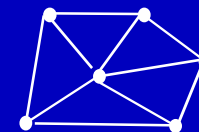
k, distance metric

k-Nearest Neighbor



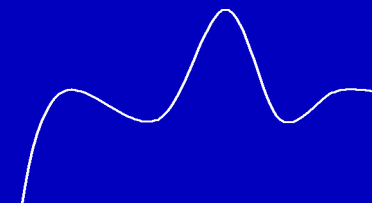
Goal, iterations

Delaunay Planes



Spread, index

Projection Pursuit Regression



# Properties of Algorithms

Algorithm	Accurate	Scalable	Interpret-able	Useable	Robust	Versatile	Fast	Hot
Classical (LR, LDA)	—	👍	👍—	👍	—	—	👍	👎
Neural Networks	👍	👎	👎	👎	—	👎	👎👎	👍
Visualization	👍	👎👎	👍	👍	👍👍	👎	👎👎👎	👍—
Decision Trees	👎	👍	👍	👍—	👍	👍	👍—	👍—
Polynomial Networks	👍	—	👎	👍—	—👎	—	—👎	—
K-Nearest Neighbors	👎	👎👎	👍—	—	—👎	👎	👍	👎
Kernels	👍	👎👎	👎	—👎	👎	👎	👍	👎

## Key

👍 good

— neutral

👎 bad

KDD-98: A Comparison of Leading Data Mining Tools

Algorithms	Decision Trees	Linear/Statistical	Multi-layer Perceptrons	Nearest Neighbor	Radial Basis Functions	Bayes	Rule Induction	Polynomial Networks	Generalized Linear Models	Time Series	Sequential Discovery	K Means	Association Rules	Kohonen
<i>Clementine</i>	√	√	√				√					√	√	√
<i>Darwin</i>	√		√	√										
<i>Datamind</i>							√							
<i>Enterprise Miner</i>	√	√	√		√				√	√		√	√	
<i>GainSmarts</i>	√	√+												
<i>Intelligent Miner</i>	√	√-	√		√-					√	√	√+	√	
<i>MineSet</i>	√					√						√	√	
<i>Model 1</i>	√+	√	√											
<i>ModelQuest</i>	√	√		√				√		√-				
<i>PRW</i>		√+	√	√	√	√						√		
<i>CART</i>	√													
<i>Cognos</i>	√													
<i>NeuroShell</i>			√+		√					√-				
<i>OLPARS</i>		√	√	√	√	√						√		√
<i>See5</i>	√						√							
<i>SPlus</i>	√	√+							√	√		√		
<i>WizWhy</i>							√							

Multi-Layer Perceptrons	Learning Rate	Learning Rate Decay	Momentum	Multiple Activation Functions	Multiple Stop Criteria	Cross-Validation	Normalize Inputs	Advanced Learning Alg.	Other Cost functions	Automatic Model Selection	Network Visual	Parameter Summary
<i>Clementine</i>	✓	✓	✓							✓		
<i>Darwin</i>	✓			✓		✓		✓	✓			✓
<i>Enterprise Miner</i>	✓	✓	✓	✓	✓		✓	✓	✓		✓	✓
<i>Intelligent Miner</i>	✓									✓		
<i>Model 1</i>	✓									✓		
<i>PRW</i>	✓		✓	✓			✓			✓		✓
<i>NeuroShell</i>	✓	✓	✓	✓	✓							
<i>OLPARS</i>	✓		✓	✓	✓		✓				✓	✓

Decision Trees	"CART"	C5 or C4.5	CHAID	Other	Priors	Classification Costs	Missing Data	Pruning Severity	Visual Trees
<i>Clementine</i>		✓				✓	✓	✓	✓-
<i>Darwin</i>	✓				✓	✓	✓		
<i>Enterprise Miner</i>	✓	✓-	✓		✓+	✓	✓	✓	✓
<i>GainSmarts</i>	✓		✓	✓			✓		✓
<i>Intelligent Miner</i>				✓			✓		✓
<i>MineSet</i>	✓		✓			✓	✓	✓	✓
<i>Model 1</i>	✓		✓				✓-		
<i>ModelQuest</i>		✓-					✓	✓	
<i>CART</i>	✓+				✓	✓	✓		✓
<i>Scenario</i>				✓			✓		
<i>S-Plus</i>	✓						✓	✓	✓
<i>See5</i>		✓+				✓	✓	✓	

KDD-98: A Comparison of Leading Data Mining Tools

Regression / Stats	Linear	Logistic	Complexity Penalty	Cross-Validation	Input Selection	Factor Analysis
<i>Clementine</i>	√					
<i>Enterprise Miner</i>	√+	√+	√	√	√	√
<i>GainSmarts</i>	√+	√+	√			
<i>Intelligent Miner</i>	√-				√	√
<i>MineSet</i>	√					
<i>Model 1</i>	√	√		√	√+	
<i>ModelQuest Enterprise</i>	√	√	√	√	√	
<i>PRW</i>	√	√		√	√+	
<i>S-Plus</i>	√+	√+	√	√	√	√
<i>Scenario</i>						√

KDD-98: A Comparison of Leading Data Mining Tools

Usability	Data Loading and Manipulation	Model Building	Model Understanding	Technical Support	Overall
<i>Clementine</i>	√+	√+	√+	√+	√+
<i>Darwin</i>	√	√	√+	√	√
<i>DataCruncher</i>	√+	√+	√	√	√
<i>Enterprise Miner</i>	√	√	√	√	√
<i>GainSmarts</i>	√+	√	√	√	√
<i>Intelligent Miner</i>	√	√	√	√	√
<i>MineSet</i>	√	√+	√+	√	√+
<i>Model 1</i>	√+	√+	√+	√+	√+
<i>ModelQuest Enterprise</i>	√	√+	√+	√+	√+
<i>PRW</i>	√+	√+	√+	√+	√+
<i>CART</i>	√-	√	√	√	√
<i>Scenario</i>	√	√+	√+	√	√+
<i>NeuroShell</i>	√	√	√	√	√
<i>OLPARS</i>	√-	√	√	√	√
<i>See5</i>	√	√	√	√	√
<i>S-Plus</i>	√	√	√+	√	√
<i>WizWhy</i>	√	√	√+	√	√

KDD-98: A Comparison of Leading Data Mining Tools

Visualization	Histograms	Pie Charts	Scatter/ Line Plots	Rotating Scatter	Conditional Plots	Classification Decision Regions	Correlation Plots
<i>Clementine</i>	√		√		√	√-	√
<i>Darwin</i>	√-	√-	√-				
<i>DataCruncher</i>	√	√	√		√		
<i>Enterprise Miner</i>	√	√	√	√-	√		√
<i>GainSmarts</i>	√-		√-				
<i>Intelligent Miner</i>	√	√	√		√		
<i>MineSet</i>	√	√	√	√	√		
<i>Model 1</i>	√	√	√	√			
<i>ModelQuest Enterprise</i>	√		√				
<i>PRW</i>	√	√	√	√			
<i>CART</i>							
<i>Scenario</i>							√
<i>NeuroShell</i>			√				
<i>OLPARS</i>	√	√	√	√-	√	√	
<i>See5</i>	√						
<i>S-Plus</i>	√	√	√		√		√
<i>WizWhy</i>							

KDD-98: A Comparison of Leading Data Mining Tools

<b>Automation</b>	<b>Method of Automation</b>	<b>Free Text Annotation of Steps</b>
<i>Clementine</i>	Visual Programming, Programming Language	√
<i>Darwin</i>	Programming Language	√
<i>DataCruncher</i>	(Task manager)	
<i>Enterprise Miner</i>	Visual Programming, Programming Language	√
<i>GainSmarts</i>	Macro Language, Wizards	√-
<i>Intelligent Miner</i>	(Wizards)	
<i>MineSet</i>	Data History, Log	
<i>Model 1</i>	Model Wizard	
<i>ModelQuest</i>	Batch Agenda	
<i>PRW</i>	Experiment Manager; Macros	√
<i>CART</i>	Built-in Basic Scripting	
<i>Scenario</i>		
<i>NeuroShell</i>		
<i>OLPARS</i>		
<i>See5</i>		
<i>S-Plus</i>	Scripting (S); C/C++	
<i>WizWhy</i>		

## A Recent Breakthrough: Bundling

- 1) Construct varied models, and
- 2) Combine their estimates

Generate component models by varying:

- Case Weights
- Data Values
- Guiding Parameters
- Variable Subsets

Combine estimates using:

- Estimator Weights
- Voting
- Advisor Perceptrons
- Partitions of Design Space

## Example Bundling Techniques

- *Bayes*: sum estimates of possible models, weighted by priors
- *GMDH* (Ivakhenko 68) -- multiple layers of quadratic polynomials, using two inputs each, fit by LR
- *Stacking* (Wolpert 92) -- train a 2nd-level (LR) model using leave-1-out estimates of 1st-level (neural net) models
- *Bagging* (Breiman 96) (*bootstrap aggregating*) -- bootstrap data (to build trees mostly); take majority vote or average
- *Bumping* (Tibshirani 97) -- bootstrap, select single best
- *Boosting* (Freund & Shapire 96) -- weight error cases by  $\beta\tau = (1-e(t))/e(t)$ , iteratively re-model; weight model  $t$  by  $\ln(\beta\tau)$
- *Crumpling* (Anderson & Elder 98) -- average cross-validations
- *Born-Again* (Breiman 98) -- invent new  $X$  data...

KDD-98: A Comparison of Leading Data Mining Tools

Distinctives	Strengths	Weaknesses
<i>Clementine</i>	visual interface; algorithm breadth	scalability
<i>Darwin</i>	efficient client-server; intuitive interface options	no unsupervised; limited visualization
<i>DataCruncher</i>	ease of use	single algorithm
<i>Enterprise Miner</i>	depth of algorithms; visual interface	harder to use; new product issues
<i>GainSmarts</i>	data transformations, built on SAS; algorithm option depth	no unsupervised; limited visualization
<i>Intelligent Miner</i>	algorithm breadth; graphical tree/cluster output	few algorithm options; no automation
<i>MineSet</i>	data visualization	few algorithms; no model export
<i>Model 1</i>	ease of use; automated model discovery	really a vertical tool
<i>ModelQuest</i>	breadth of algorithms	some non-intuitive interface options
<i>PRW</i>	extensive algorithms; automated model selection	limited visualization
<i>CART</i>	depth of tree options	difficult file I/O; limited visualization
<i>Scenario</i>	ease of use	narrow analysis path
<i>NeuroShell</i>	multiple neural network architectures	unorthodox interface; only neural networks
<i>OLPARS</i>	multiple statistical algorithms; class-based visualization	dated interface; difficult file I/O
<i>See5</i>	depth of tree options	limited visualization; few data options
<i>S-Plus</i>	depth of algorithms; visualization; programable/extendable	limited inductive methods; steep learning curve
<i>WizWhy</i>	ease of use; ease of model understanding	limited visualization

## Closing Observations

- Data Mining Tools Can:
  - Enhance inference process
  - Speed up design cycle
- Data Mining Tools Can Not:
  - Substitute for statistical and domain expertise
- Users are advised to:
  - Get training on tools
  - Be alert for product upgrades

## Forthcoming Report

- Report provides detailed comparison of high-end data mining tools, including capabilities, ease of use, and practical tips.
- Available for \$695 from Elder Research (<http://www.datamininglab.com>), Q4 1998.
- Purchasers receive brief free consulting session to explore report findings in more detail, if desired.

Note: The analyses and reviews were performed completely independently, and were made possible by the cooperation of the vendors, for which Elder Research is very grateful. The companies, however, provided no financial support, and had no influence on its editorial content.