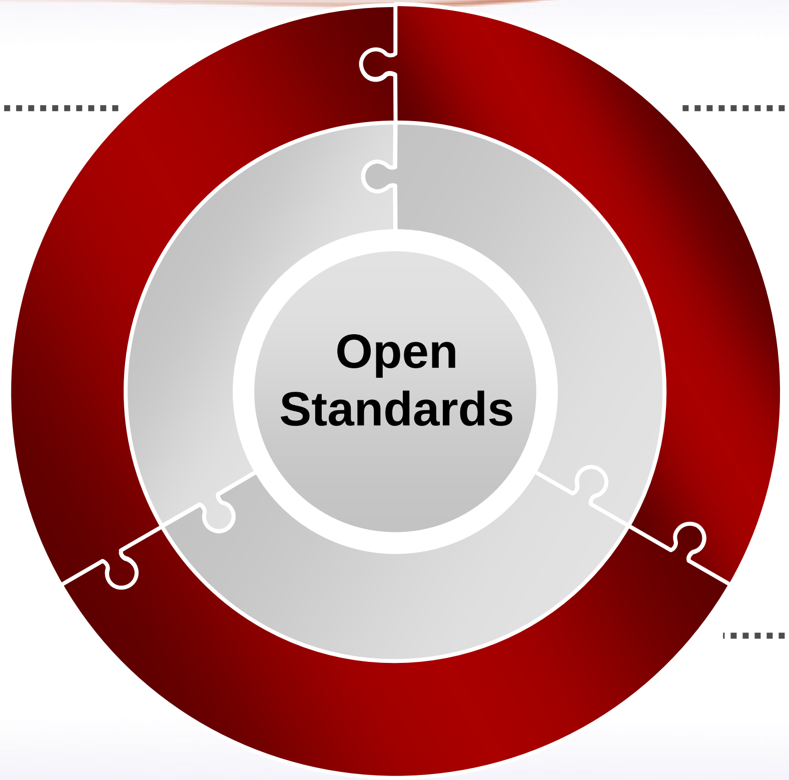
Easy Execution of Data Mining Models through PMML

Zementis, Inc.

UseR! 2009

Development, Deployment, and Execution

of Predictive Models



Development

R allows for reliable data manipulation and model building

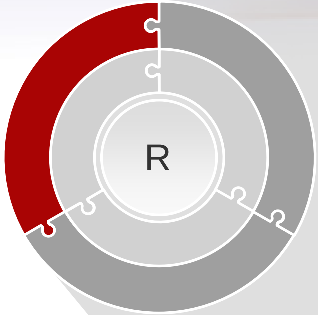
Deployment

PMML allows for easy expression and deployment of data transformations and data-mining models

Execution

Real-time execution of models via web-services calls

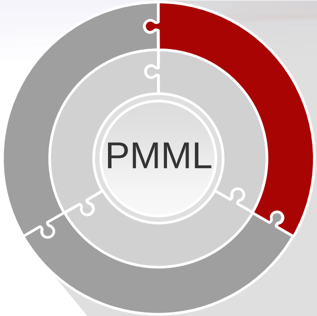
Model Development



The R Project

- R is an integrated suite of software facilities for data manipulation, calculation and graphical display.
- R provides a wide variety of statistical techniques and is highly extensible.
- R is similar to the S language and environment developed at Bell Labs.
- It is Open Source and a GNU project.
- R is available for free at http://www.r-project.org/

Model Deployment

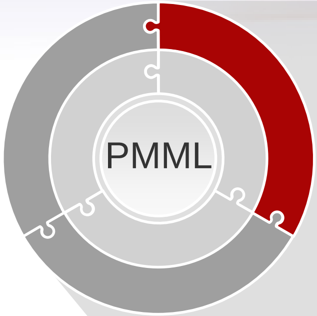


Predictive Model Markup Language (PMML)

- PMML is an XML-based language to
  - Define statistical and data mining models
  - Share models between compliant applications
- Standard for exchange of models to
  - Avoid proprietary issues and incompatibilities
  - Deploy models in operational infrastructure
- Clear separation of tasks
  - Model development vs. model deployment
  - Scientists focus on building the best model
  - Eliminates need for custom model deployment
  - Ensures scalability and reliability

PMML

Industry Support

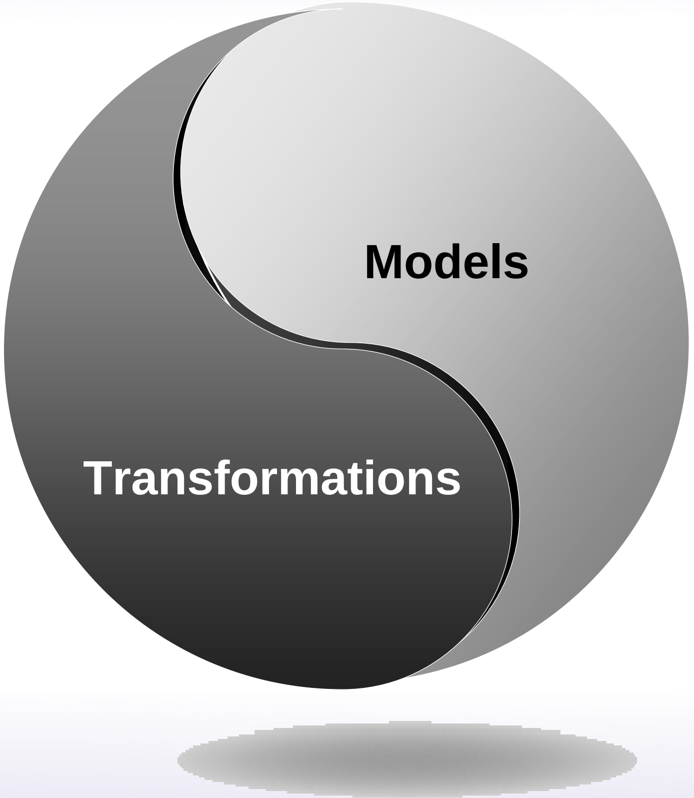


Matured and Supported by Industry

- Data Mining Group http://www.dmg.org
- Mature standard
  - Current version 4.0 (just released)
  - Active group and constant enhancements
- Vendor independent consortium
- Industry supporters
  - Major Players: IBM, Oracle, SAP, Microsoft
  - Analytics: SAS, SPSS, KXEN, Zementis
  - Business Intelligence: Microstrategy, Teradata
  - Open Source: R, KNIME

PMML

Bringing data and Models Together

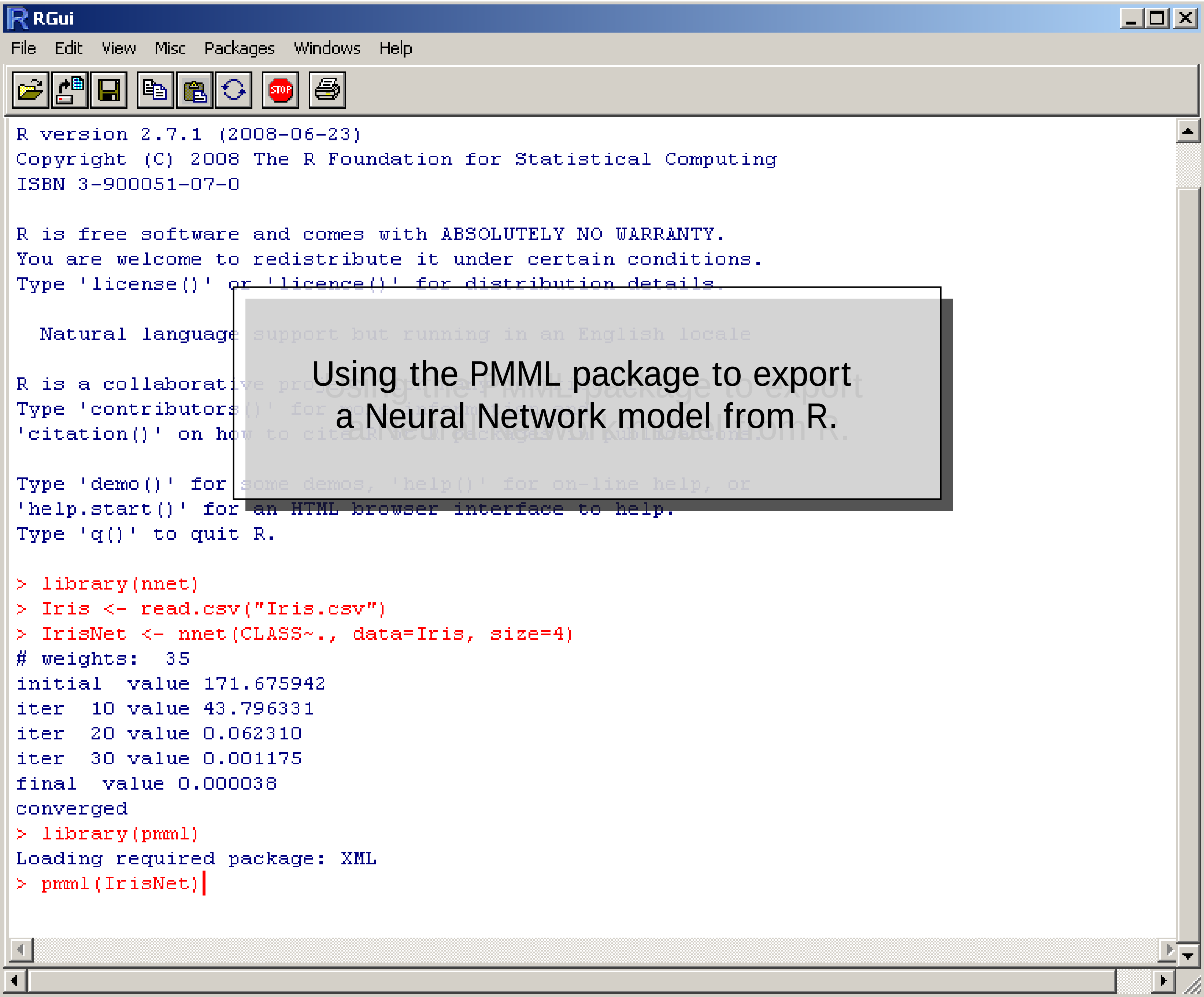


Data Transformations and Data-Mining Models come together in PMML.

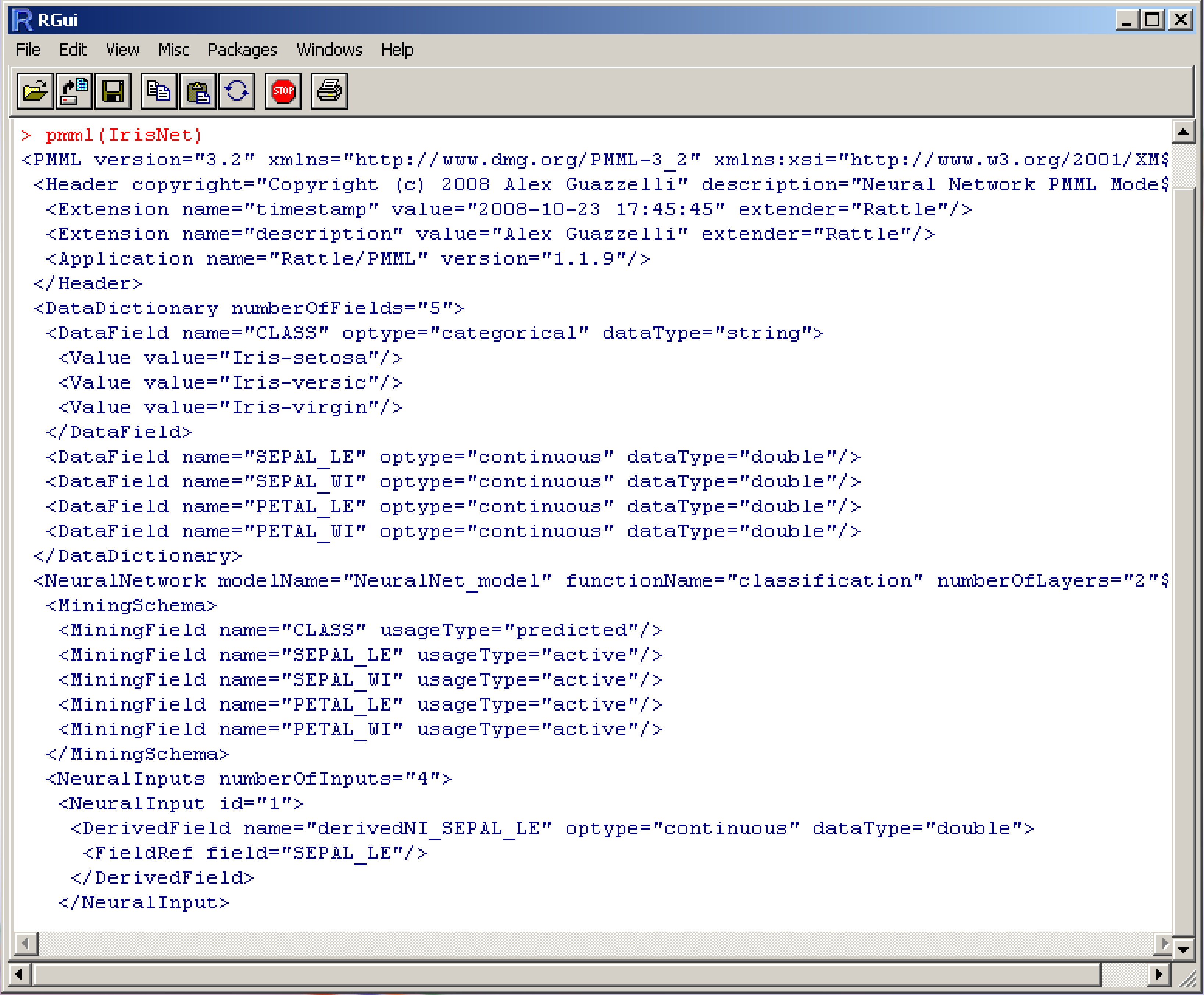
Predictive Modeling Markup Language

- A **Data Dictionary** defines all the raw data fields (including missing value strategy and outlier treatment).
- Several **Data Transformations** strategies allow for intelligent extraction of feature detectors from raw data ("data massaging").
- A comprehensive list of **Data-Mining Models** offers power and flexibility.
- Post-processing of results allow for tailored decisions

Using the PMML package to export a Neural Network model from R.

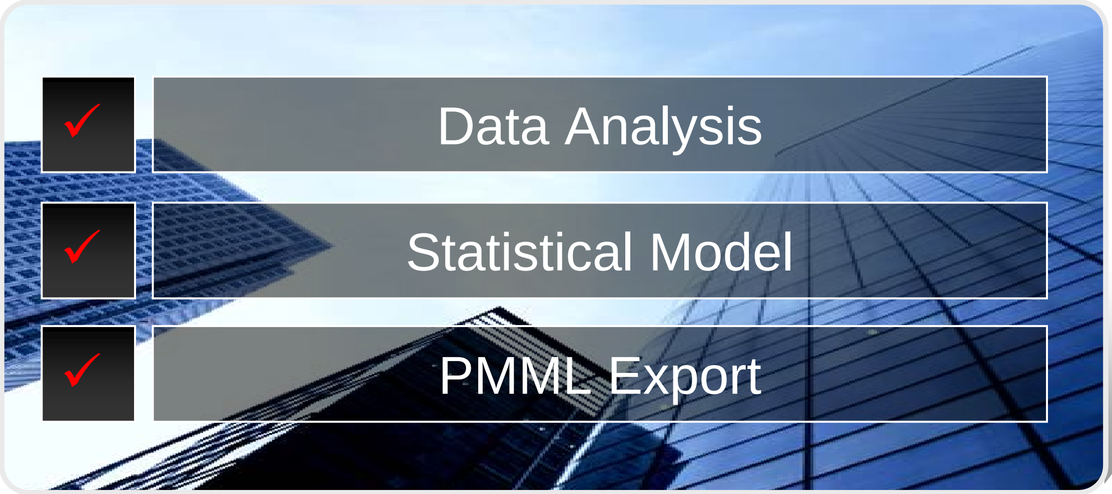


```
R version 2.7.1 (2008-06-23)
Copyright (C) 2008 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> library(nnet)
> Iris <- read.csv("Iris.csv")
> IrisNet <- nnet(CLASS~., data=Iris, size=4)
# weights:  35
initial  value 171.675942
iter  10 value 43.796331
iter  20 value 0.062310
iter  30 value 0.001175
final  value 0.000038
converged
> library(pmml)
Loading required package: XML
> pmml(IrisNet)
```



```
> pmml(IrisNet)
<PMML version="3.2" xmlns="http://www.dmg.org/PMML-3_2" xmlns:xsi="http://www.w3.org/2001/XM$
 <Header copyright="Copyright (c) 2008 Alex Guazzelli" description="Neural Network PMML Mode$
  <Extension name="timestamp" value="2008-10-23 17:45:45" extender="Rattle"/>
  <Extension name="description" value="Alex Guazzelli" extender="Rattle"/>
  <Application name="Rattle/PMML" version="1.1.9"/>
 </Header>
 <DataDictionary numberOfFields="5">
  <DataField name="CLASS" optype="categorical" dataType="string">
   <Value value="Iris-setosa"/>
   <Value value="Iris-versic"/>
   <Value value="Iris-virgin"/>
  </DataField>
  <DataField name="SEPAL_LE" optype="continuous" dataType="double"/>
  <DataField name="SEPAL_WI" optype="continuous" dataType="double"/>
  <DataField name="PETAL_LE" optype="continuous" dataType="double"/>
  <DataField name="PETAL_WI" optype="continuous" dataType="double"/>
 </DataDictionary>
 <NeuralNetwork modelName="NeuralNet_model" functionName="classification" numberOfLayers="2"$
  <MiningSchema>
   <MiningField name="CLASS" usageType="predicted"/>
   <MiningField name="SEPAL_LE" usageType="active"/>
   <MiningField name="SEPAL_WI" usageType="active"/>
   <MiningField name="PETAL_LE" usageType="active"/>
   <MiningField name="PETAL_WI" usageType="active"/>
  </MiningSchema>
  <NeuralInputs numberOfInputs="4">
   <NeuralInput id="1">
    <DerivedField name="derivedNI_SEPAL_LE" optype="continuous" dataType="double">
     <FieldRef field="SEPAL_LE"/>
    </DerivedField>
   </NeuralInput>
```

Got Models…

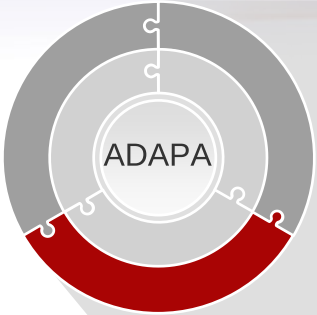


- ✓ Data Analysis
- ✓ Statistical Model
- ✓ PMML Export

What Now?

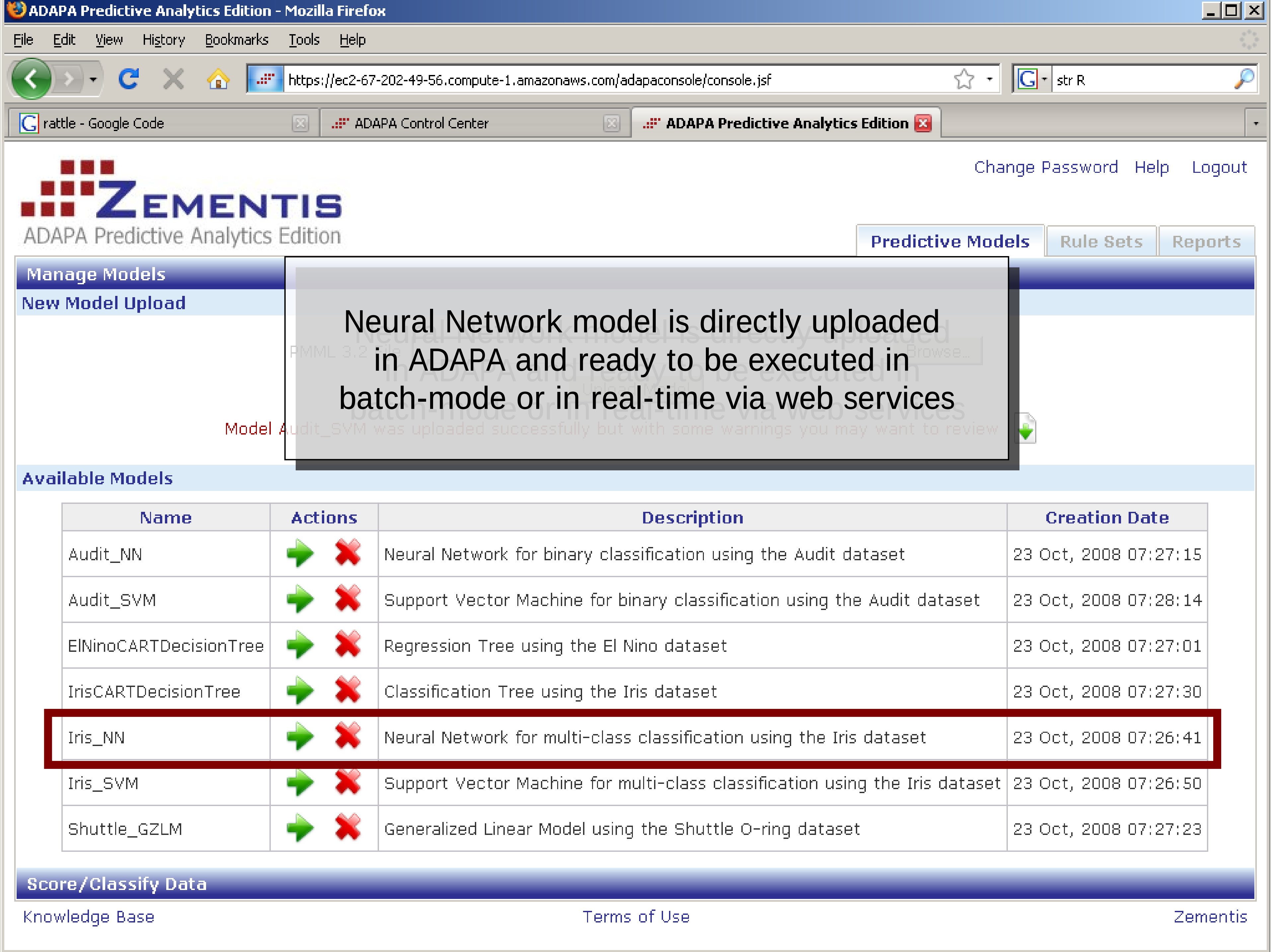
Model Execution

The ADAPA Example



Predictive Analytics Scoring Engine

- Data transformations and model execution in real-time (via web-services calls) or batch-mode.
- Environment to manage and deploy many predictive models.
- Framework for SOA-based IT integration
  - Completely standards based and easily integrated with any existing infrastructure.
- Not a model building environment.
- Available as a Service in the Amazon Cloud (EC2).


ADAPA Predictive Analytics Edition - Mozilla Firefox
File
Edit
View
History
Bookmarks
Tools
Help
https://ec2-67-202-49-56.compute-1.amazonaws.com/adapaconsole/console.jsf
str R
rattle - Google Code
ADAPA Control Center
ADAPA Predictive Analytics Edition
ZEMENTIS
ADAPA Predictive Analytics Edition
Change Password
Help
Logout
Predictive Models
Rule Sets
Reports
PMML 3.2 file
Browse...
Upload

Neural Network model is directly uploaded in ADAPA and ready to be executed in batch-mode or in real-time via web services

Manage Models

New Model Upload

Model Audit_SVM was uploaded successfully but with some warnings you may want to review

Available Models

Name	Actions	Description	Creation Date
Audit_NN	➡ ✖	Neural Network for binary classification using the Audit dataset	23 Oct, 2008 07:27:15
Audit_SVM	➡ ✖	Support Vector Machine for binary classification using the Audit dataset	23 Oct, 2008 07:28:14
ElNinoCARTDecisionTree	➡ ✖	Regression Tree using the El Nino dataset	23 Oct, 2008 07:27:01
IrisCARTDecisionTree	➡ ✖	Classification Tree using the Iris dataset	23 Oct, 2008 07:27:30
Iris_NN	➡ ✖	Neural Network for multi-class classification using the Iris dataset	23 Oct, 2008 07:26:41
Iris_SVM	➡ ✖	Support Vector Machine for multi-class classification using the Iris dataset	23 Oct, 2008 07:26:50
Shuttle_GZLM	➡ ✖	Generalized Linear Model using the Shuttle O-ring dataset	23 Oct, 2008 07:27:23

Score/Classify Data

Thank You!

E-mail: *info@zementis.com*

U.S.A

6125 Cornerstone Court East
Suite 250
San Diego, CA, 92121

Tel: +1 619 330-0780
Fax: +1 858 535-0227

Asia

19/F., Unit A
Ho Lee Commercial Building
38-44 D'Aguilar Street
Central, Hong Kong (S.A.R.)

Tel: +852 2868-0878
Fax: +852 2845-6027