

Name: Srinivasan Govindaraj

Title: Big Data Predictive Analytics



Please note the following

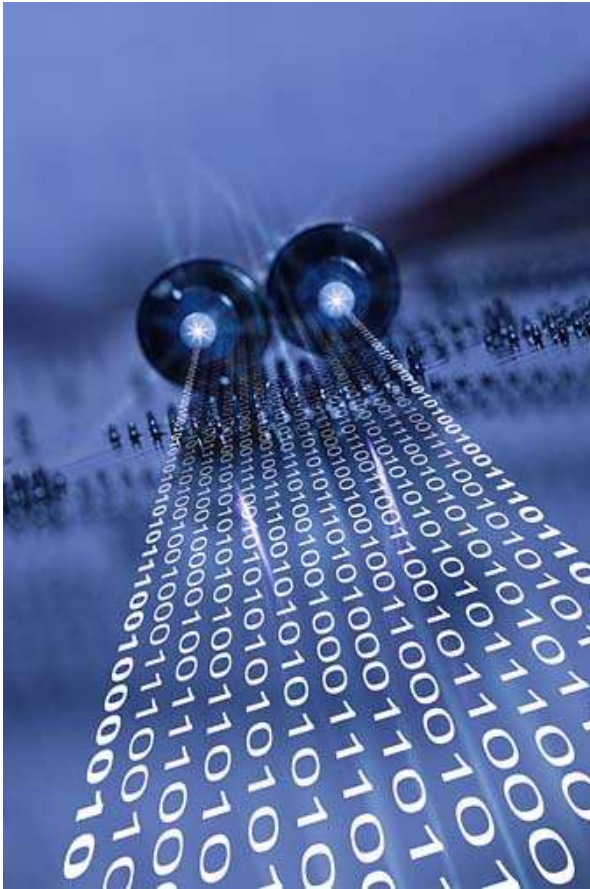
IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM's sole discretion.

Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision.

The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon many factors, including considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve results similar to those stated here.

What is Predictive Analytics?



The application of statistical analysis of historic data to predict future trends, patterns, and behaviors to improve the outcomes in automated and human processes

Predictive Algorithms

▪ **Classification**

- Uses a known outcome field (target) and its relationship to predictor inputs to build a model that can predict the target value in new data
- Linear Regression, Logistic Regression, C5.0, C&RT, QUEST, Neural Network, SVM

▪ **Association**

- Builds a model that shows the patterns of entities (events, purchases, attributes) in a data set
- Apriori, CARMA, Sequence

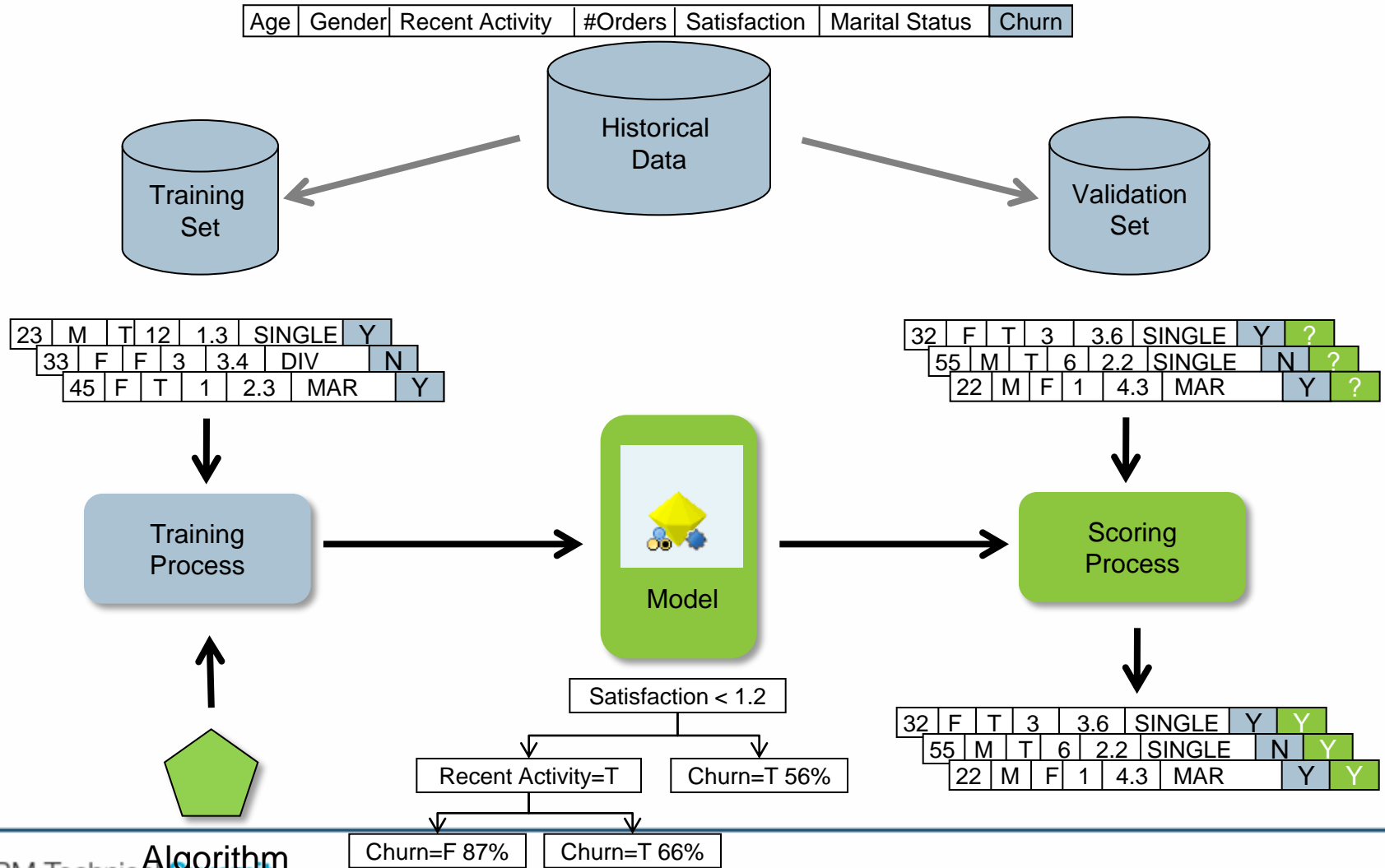
▪ **Segmentation**

- Divides the data set into clusters of records that are similar
- K-Means, Kohonen, TwoStep Cluster

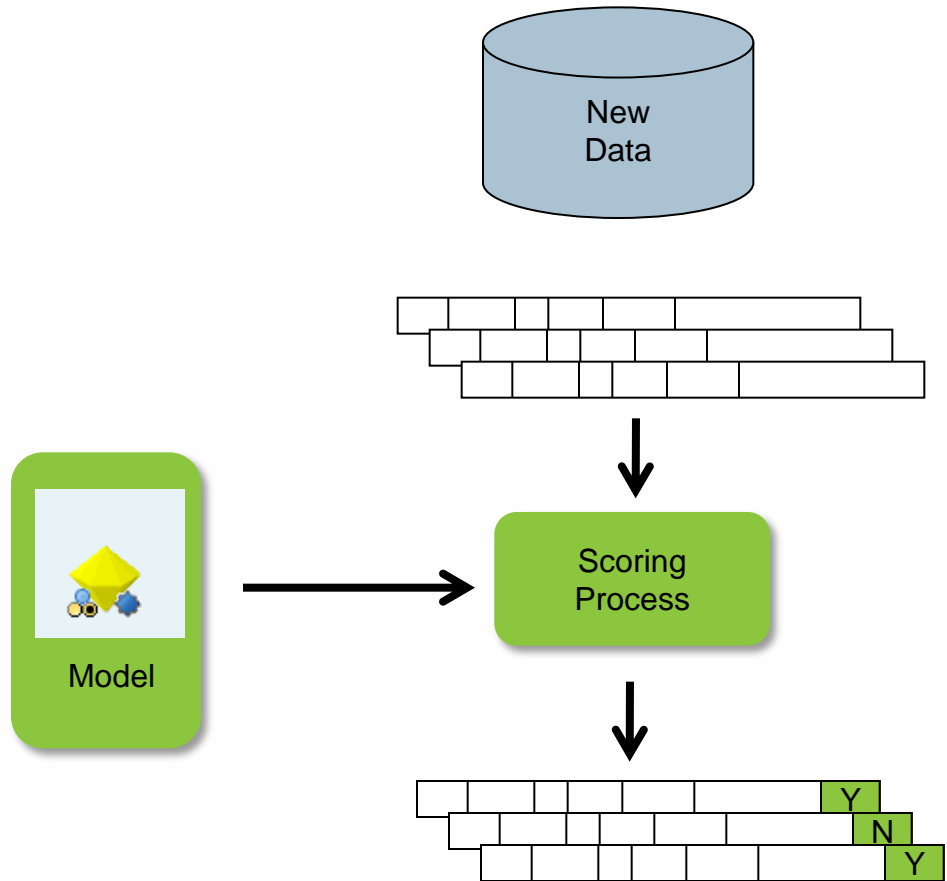
▪ **Forecasting**

- Produces future estimates for time based data
- ARIMA, Exponential Smoothing

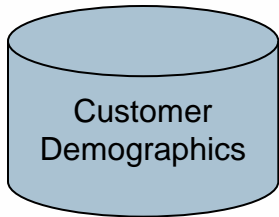
Model Building



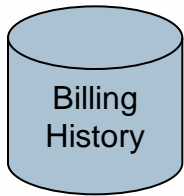
Predictive Scoring



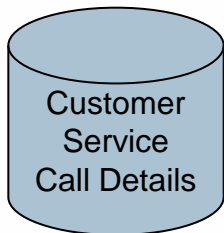
Data Exploration / Preparation



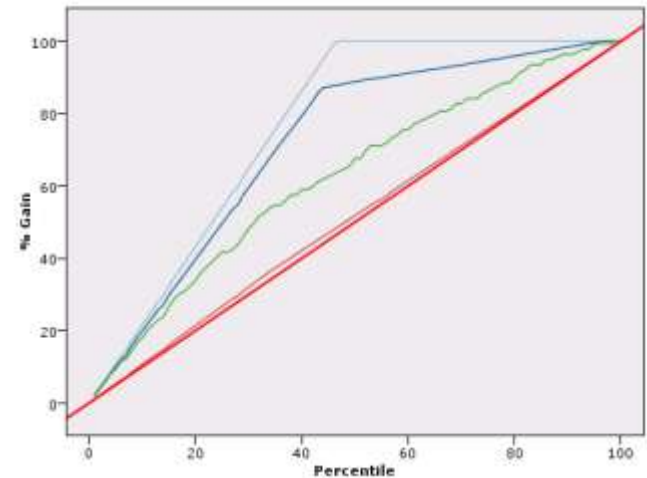
Age	Gender	Marital Status	Churn
-----	--------	----------------	-------



Age	Gender	#Late Pmt	#Orders	Marital Status	Churn
-----	--------	-----------	---------	----------------	-------

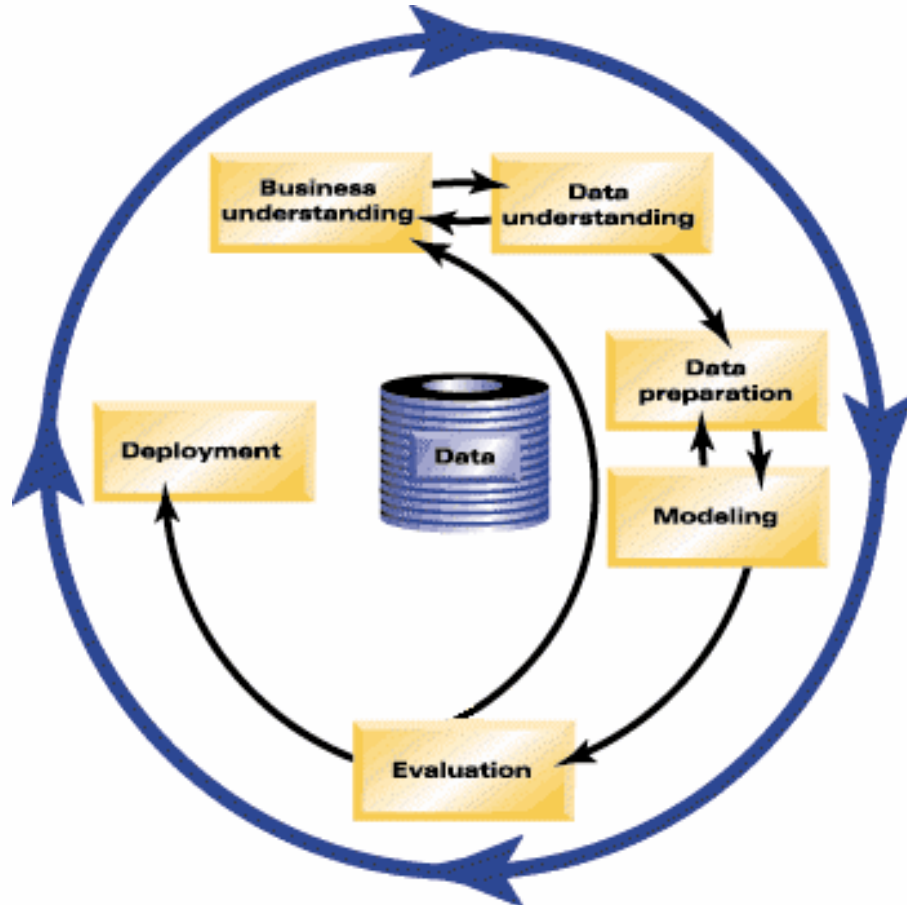


Age	Gender	#Late Pmt	#Orders	Marital Sts	Satisfaction	Churn
-----	--------	-----------	---------	-------------	--------------	-------



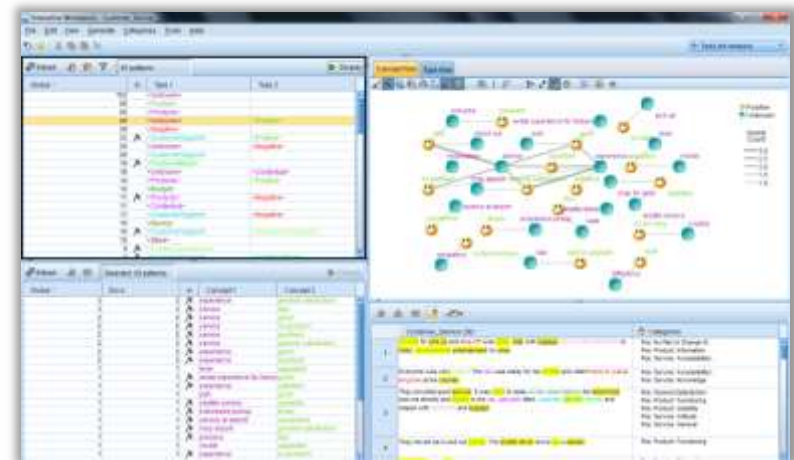
Data Mining Methodology – CRISP-DM

- Business understanding
- Data understanding
- Data preparation
- Modeling
- Evaluation
- Deployment



IBM SPSS Modeler

- High-performance data mining and text analytics workbench
- Easy-to-use, interactive interface without the need for programming
- Automated modeling and data preparation capabilities
- Access ALL data – structured and unstructured – from disparate sources
- Natural Language Processing (NLP) to extract concepts and sentiments in text
- Entity Analytics ensures the quality of the data and results in more accurate models
- Leverage existing investment in Cognos, Netezza, InfoSphere and System Z



Modeler User Interface

The screenshot displays the IBM SPSS Modeler user interface. At the top, the title bar reads "SampleStream - IBM® SPSS® Modeler". Below it is a menu bar with "File", "Edit", "Insert", "View", "Tools", "SuperNode", "Window", and "Help". A toolbar contains icons for file operations, editing, and analysis. The main workspace shows a workflow diagram with the following steps: "Merge" (receiving input from "IBM Cognos BI", "HistoricalData", and "Excel"), "Select", "Derive1", "Filter" (with "17 Fields" output), "Type" (with "Distribution" and "Web" outputs), "Partition" (with "Months with service" output), and "Model" (with "Auto Classifier" output). On the right, a "Streams" panel shows "SampleStream" and a "CRISP-DM" panel lists classes: "(unsaved project)", "Business Understanding", "Data Understanding", "Data Preparation", "Modeling", "Evaluation", and "Deployment". At the bottom, a "Favorites" bar includes "Sources", "Record Ops", "Field Ops", "Graphs", "Modeling", "Database Modeling", "Output", "Export", "IBM® SPSS® Statistics", and "IBM® SPSS® Text Analytics". A secondary toolbar at the very bottom lists tools: "Database", "Var. File", "Auto Data Prep", "Select", "Sample", "Aggregate", "Derive", "Type", "Filter", "Graphboard", "Auto Classifier", "Auto Numeric", "Auto Cluster", "Table", "Flat File", and "Database".

Modeler User Interface – Nodes Palette

The screenshot displays the IBM SPSS Modeler user interface. The main workspace shows a workflow diagram with the following nodes and connections:

- Inputs:** IBM Cognos BI, HistoricalData, and Excel.
- Merge:** Receives input from all three sources.
- Select:** Receives input from Merge.
- Derive1:** Receives input from Select.
- Filter:** Receives input from Derive1.
- Type:** Receives input from Filter and branches into three paths:
 - Distribution:** Receives input from Type.
 - Web:** Receives input from Type.
 - 17 Fields:** Receives input from Type.
- Partition:** Receives input from Type.
- Auto Classifier:** Receives input from Partition.
- Model:** Receives input from Auto Classifier.

The right-hand side of the interface features a **Streams** panel with a **SampleStream** entry and a **CRISP-DM** panel with a **Classes** list:

- Classes:**
 - (unsaved project)
 - Business Understanding
 - Data Understanding
 - Data Preparation
 - Modeling
 - Evaluation
 - Deployment

The bottom of the interface includes a **Favorites** bar and a **Nodes Palette** with the following categories and icons:

- Sources:** Database, Var. File
- Auto Data Prep:** Auto Data Prep
- Select:** Select
- Sample:** Sample
- Aggregate:** Aggregate
- Derive:** Derive
- Type:** Type
- Filter:** Filter
- Graphboard:** Graphboard
- Auto Classifier:** Auto Classifier
- Auto Numeric:** Auto Numeric
- Auto Cluster:** Auto Cluster
- Table:** Table
- Flat File:** Flat File
- Database:** Database

Modeler User Interface – Stream Canvas

The screenshot displays the IBM SPSS Modeler interface with a Stream Canvas workflow. The workflow starts with three data sources: IBM Cognos BI, HistoricalData (SQL), and Excel. These feed into a Merge node, followed by Select, Derive1, and Filter nodes. The Filter node branches into Distribution, Web, and 17 Fields. The 17 Fields node feeds into a Type node, which then feeds into a Partition node. The Partition node feeds into an Auto Classifier node, which finally feeds into a Model node. The right-hand side of the interface shows a Streams panel with 'SampleStream' and a CRISP-DM Classes panel with a project structure including Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. The bottom of the interface features a toolbar with various tool icons and a status bar with tabs for Favorites, Sources, Record Ops, Field Ops, Graphs, Modeling, Database Modeling, Output, Export, IBM SPSS Statistics, and IBM SPSS Text Analytics.

SampleStream - IBM® SPSS® Modeler

File Edit Insert View Tools SuperNode Window Help

Streams Outputs Models

SampleStream

CRISP-DM Classes

- (unsaved project)
- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

Database Var. File Auto Data Prep Select Sample Aggregate Derive Type Filter Graphboard Auto Classifier Auto Numeric Auto Cluster Table Flat File Database

Modeler User Interface – Managers Pane & Project Pane

The screenshot displays the IBM SPSS Modeler user interface. The main workspace shows a workflow diagram with the following nodes: IBM Cognos BI, HistoricalData, Excel, Merge, Select, Derive1, Filter, Type, Distribution, Web, 17 Fields, Months with service, Auto Classifier, Partition, and Model. The right-hand pane is divided into two sections: 'Streams' (containing 'SampleStream') and 'CRISP-DM Classes' (listing Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment). The bottom toolbar includes various tool icons such as Database, Var. File, Auto Data Prep, Select, Sample, Aggregate, Derive, Type, Filter, Graphboard, Auto Classifier, Auto Numeric, Auto Cluster, Table, Flat File, and Database.

Big Data Analytics

- **What capabilities are important for organization's dealing with big data?**
 - Better outcomes
 - Utilizing existing IT investment
 - High performance and Scalable

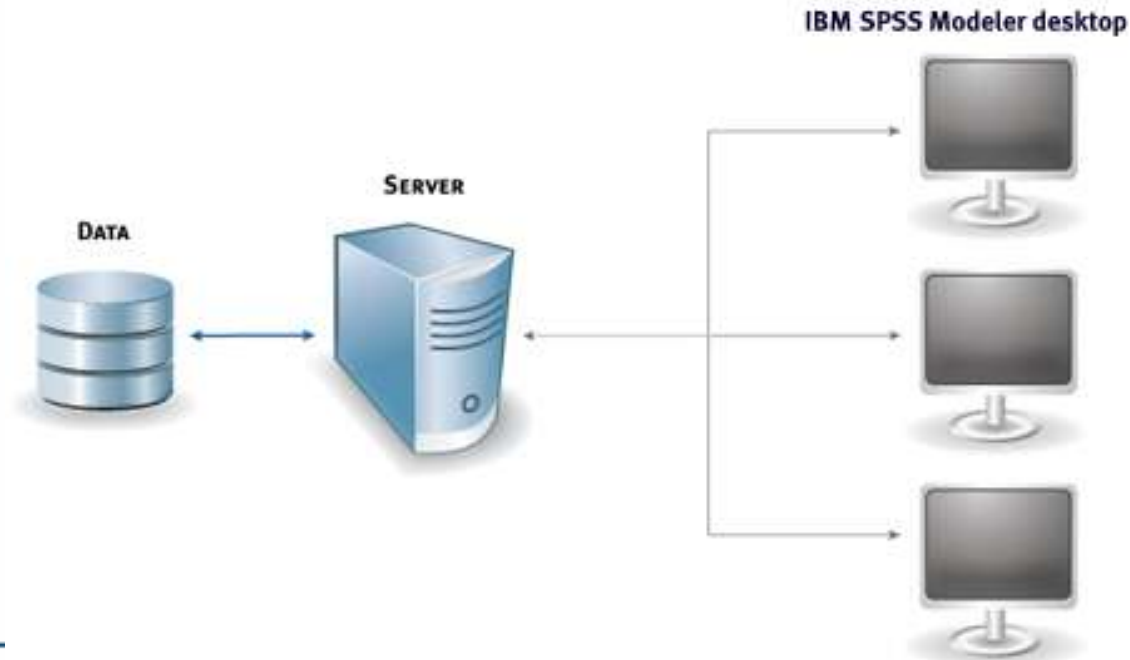
- **Where would you use it?**
 - Performance on large volumes of data
 - Deployment and execution of analytics within an organization
 - Benefit of in-database mining
 - Support for database vendor algorithms via helper applications
 - Scoring adapters

Big Data Analytics

- Volume
 - Move Analytics to the Data
 - Hadoop & MPP Databases
- Velocity
 - Deploy Analytics Inside of Transaction Systems
 - InfoSphere Streams – WebSphere – DB2 z/OS
- Variety
 - Text Analytics – extract concepts and sentiment
 - Video & Image – identify subject – facial recognition

In-Database Support with SPSS Modeler Server

- All the features of IBM SPSS Modeler
 - Large volumes of data
 - High performance
 - Administration and security options
-
- **In-Database via...**
 - SQL pushback
 - In Database Algorithms
 - Scoring Adapters
 - SQL scoring



Helper Applications in SPSS Modeler

- Modeler Server supports integration with data mining and modeling tools that are available from database vendors, including
 - IBM Netezza
 - IBM DB2 InfoSphere Warehouse
 - Oracle Data Miner
 - Microsoft Analysis Services



SPSS Modeler and Netezza

- Modeler supports integration with IBM Netezza, providing the ability to run data mining algorithms to be directly in the IBM Netezza environment from the Modeler user interface.
- The following algorithms from Netezza Analytics are supported within Modeler

- Bayes Net
- Decision Trees
- Divisive Clustering
- Generalized Linear
- K-Means
- KNN
- Linear Regression
- Naive Bayes
- PCA
- Regression Tree
- Time Series



Netezza Bayes Net



Netezza Generalized Linear



Netezza Linear Regression



Netezza Decision Tree



Netezza K-Means



Netezza Naive Bayes



Netezza Regression Tree



Netezza Divisive Clustering



Netezza KNN

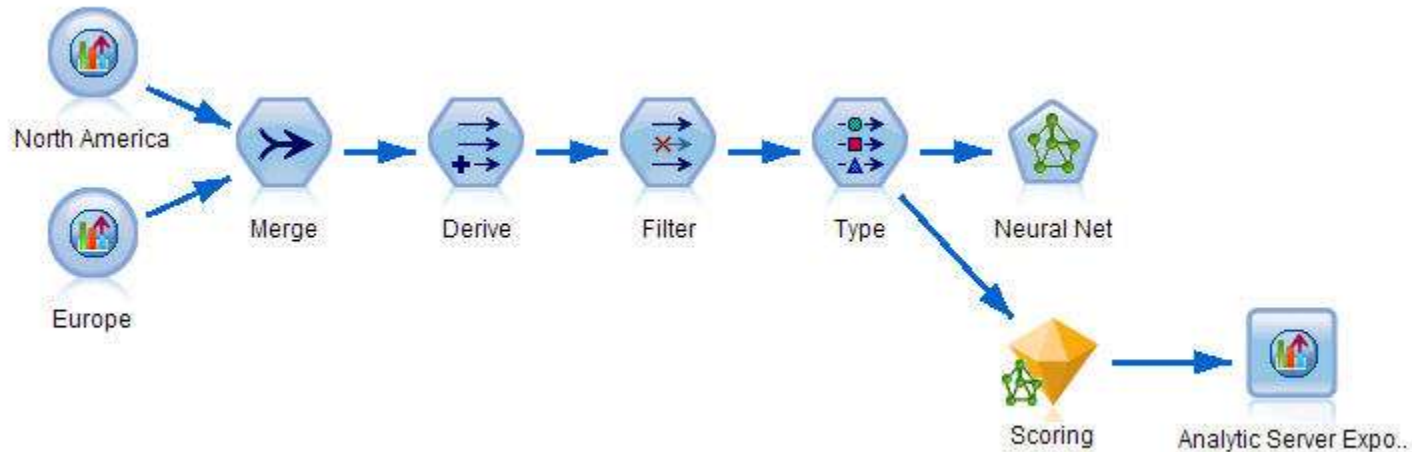


Netezza PCA



Netezza Time Series

Predictive Analytics for Hadoop-based Distributed Systems



- **IBM SPSS Modeler enabled for big data analytics**

- Multiple data access paths: file, database, HCatalog
- Distributed processing of transformations, select model building & most model scoring
- No Map/Reduce coding required
- Federation with traditional RDBMS
- R function support for distributed systems

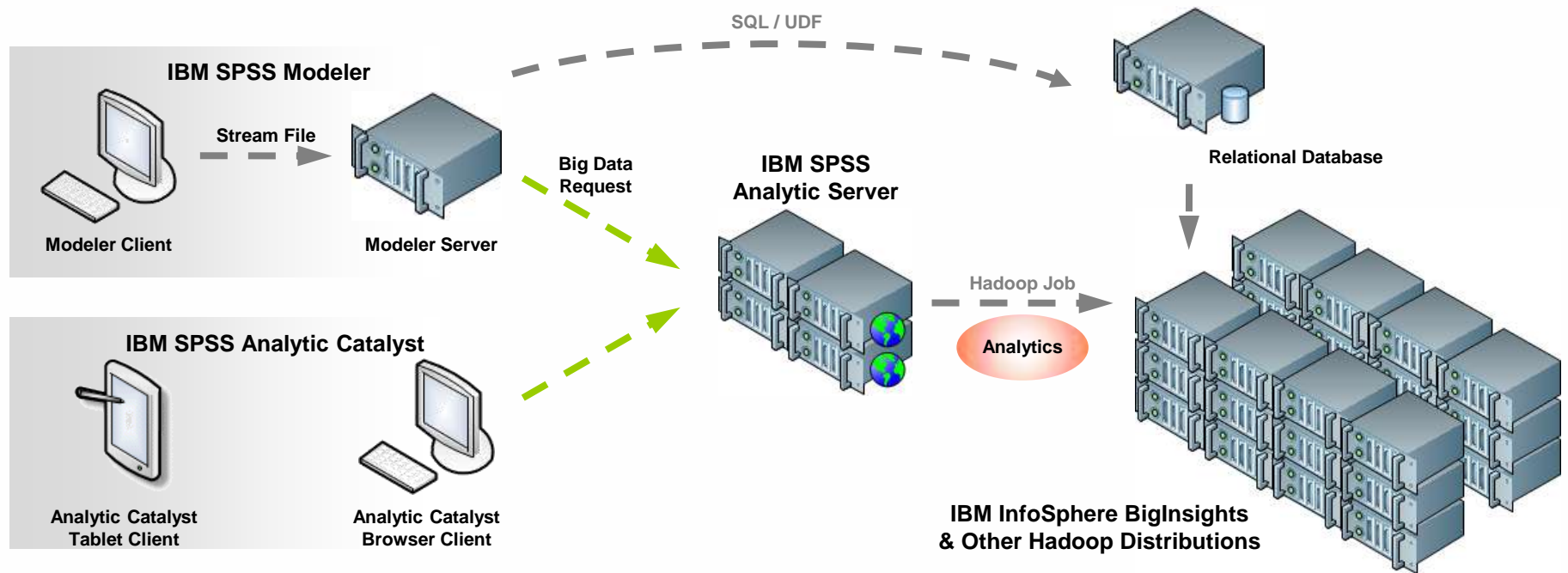
- **Text Analytics for big data**

- Scoring of text analytics models against distributed data sets

- **Integrated & open architecture**

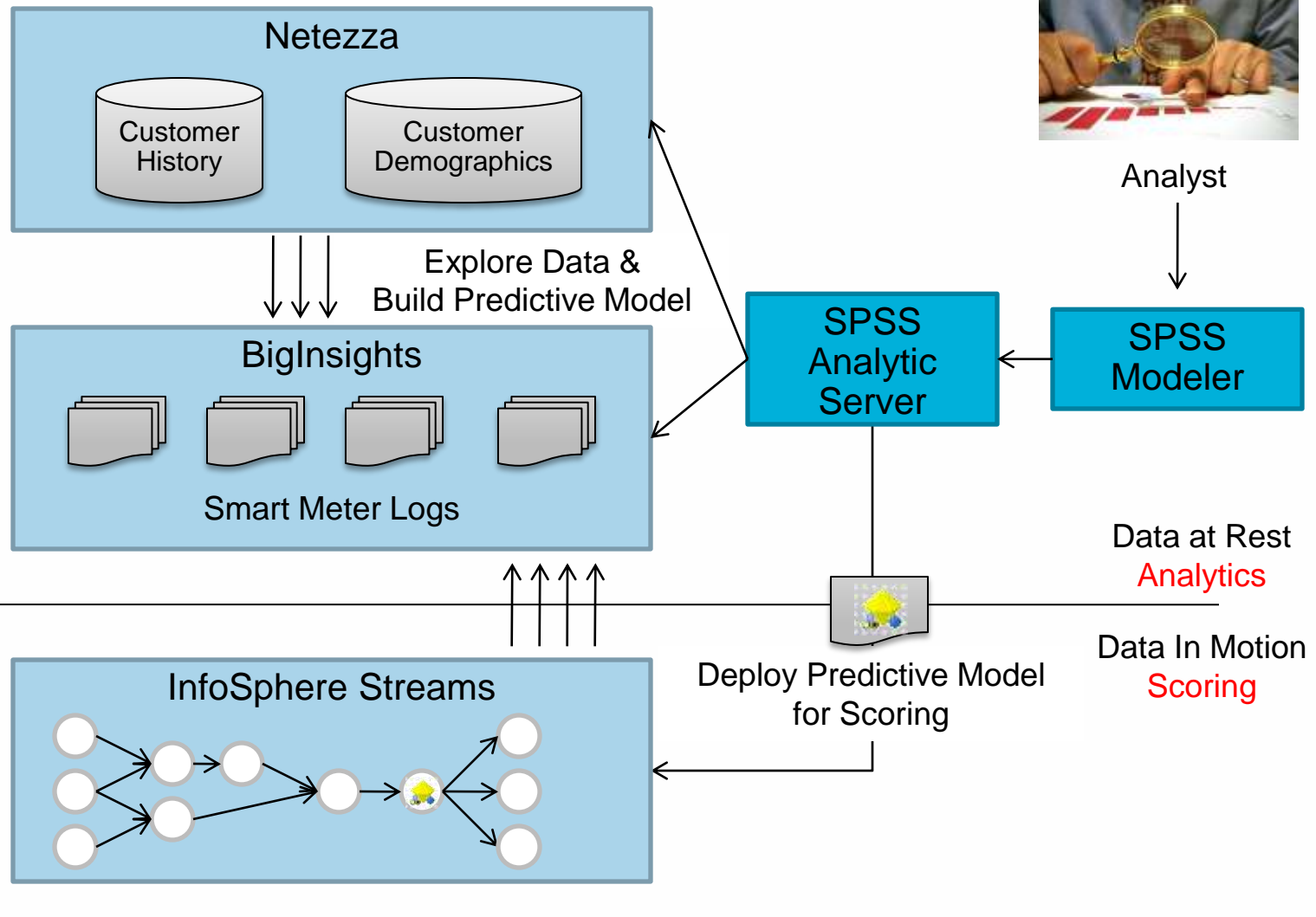
- Asset lifecycle management via Collaboration and Deployment Services
- Support for BigInsights, Cloudera, Hortonworks & Apache Hadoop distributions

Big Data Architecture



- Modeler Server utilizes Analytic Server for Big Data
- Analysts define analysis in a familiar & accessible workbench to conduct analysis, modeling & scoring over high volumes of varied data
 - Federation of heterogeneous data sources to use legacy & external data in model building & scoring
 - Transformations, sampling & write-back of output to big data systems

Big Data Analytics Architecture



IBM SPSS Modeler Script

- Macro language for building / editing / running a Modeler Stream
- Execution Options
 - Command Line
 - Interactively from IBM SPSS Modeler Interface
 - Scheduled from IBM SPSS Collaboration & Deployment Services
 - Web Service invocation from IBM SPSS Collaboration & Deployment Services

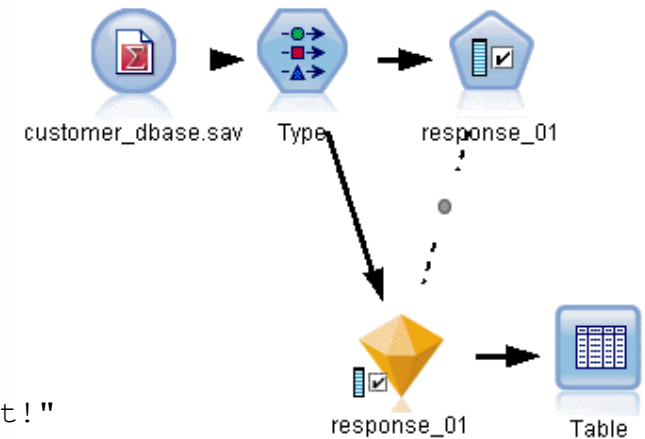
Script Example

```
create stream 'featureselection'  
create statisticsimportnode  
position :statisticsimportnode at 50 50  
set :statisticsimportnode.full_filename = "$CLEO_DEMOS/customer_dbase.sav"
```

```
create typenode  
position :typenode at 150 50  
set :typenode.direction.'response_01' = Target  
connect :statisticsimportnode to :typenode
```

```
create featureselectionnode  
position :featureselectionnode at 250 50  
set :featureselectionnode.screen_missing_values=true  
set :featureselectionnode.max_missing_values=80  
set :featureselectionnode.criteria = Likelihood  
set :featureselectionnode.important_label = "Check Me Out!"  
set :featureselectionnode.selection_mode = TopN  
set :featureselectionnode.top_n = 15  
connect :typenode to :featureselectionnode  
execute :featureselectionnode
```

```
create tablenode  
position :tablenode at 250 250  
connect response_01:applyfeatureselectionnode to :tablenode  
execute :tablenode
```



Questions

Acknowledgements and disclaimers

Availability: References in this presentation to IBM products, programs, or services do not imply that they will be available in all countries in which IBM operates.

The workshops, sessions and materials have been prepared by IBM or the session speakers and reflect their own views. They are provided for informational purposes only, and are neither intended to, nor shall have the effect of being, legal or other guidance or advice to any participant. While efforts were made to verify the completeness and accuracy of the information contained in this presentation, it is provided AS-IS without warranty of any kind, express or implied. IBM shall not be responsible for any damages arising out of the use of, or otherwise related to, this presentation or any other materials. Nothing contained in this presentation is intended to, nor shall have the effect of, creating any warranties or representations from IBM or its suppliers or licensors, or altering the terms and conditions of the applicable license agreement governing the use of IBM software.

All customer examples described are presented as illustrations of how those customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics may vary by customer. Nothing contained in these materials is intended to, nor shall have the effect of, stating or implying that any activities undertaken by you will result in any specific sales, revenue growth or other results.

© **Copyright IBM Corporation 2013. All rights reserved.**

– **U.S. Government Users Restricted Rights - Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.**

IBM, the IBM logo, ibm.com, Rational, the Rational logo, Telelogic, the Telelogic logo, Green Hat, the Green Hat logo, and other IBM products and services are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. If these and other IBM trademarked terms are marked on their first occurrence in this information with a trademark symbol (® or ™), these symbols indicate U.S. registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at “Copyright and trademark information” at www.ibm.com/legal/copytrade.shtml

If you have mentioned trademarks that are not from IBM, please update and add the following lines:

[Insert any special third-party trademark names/attributions here]

Other company, product, or service names may be trademarks or service marks of others.

Thank You