IBM Business Analytics

IBM Software Solutions Group

IBM SPSS Predictive Analytics Overview

with IBM SPSS Modeler

Rebecca Young– Data Scientist, Advanced Analytics, IBM Canada
Soha Moosavi, Business Analyst, BCNet

April 26, 2017

Presenters

Rebecca Young
Data Scientist, IBM Canada

Soha Moosavi
Business Analyst, BCNET

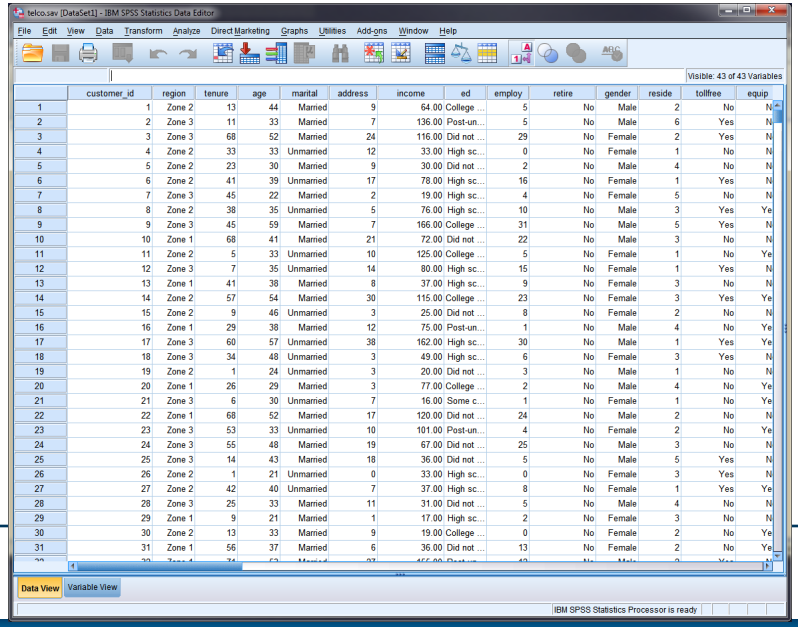
SPSS Products and Components – Statistics

SPSS Statistics

Statistics Desktop: rich desktop application for statistical analysis and hypothesis testing. Note that Statistics Desktop is self-sufficient; it does not require a connection to the Server. It is not strictly a "client" in the usual sense.

Statistics Server: server component used for statistical analysis and for executing in batch mode. Statistics Desktop can connect to this server. Statistics Server is also capable of running scheduled processes, invoked by SPSS Collaboration and Deployment Services or any other job scheduler.

IBM SPSS Statistics Desktop:



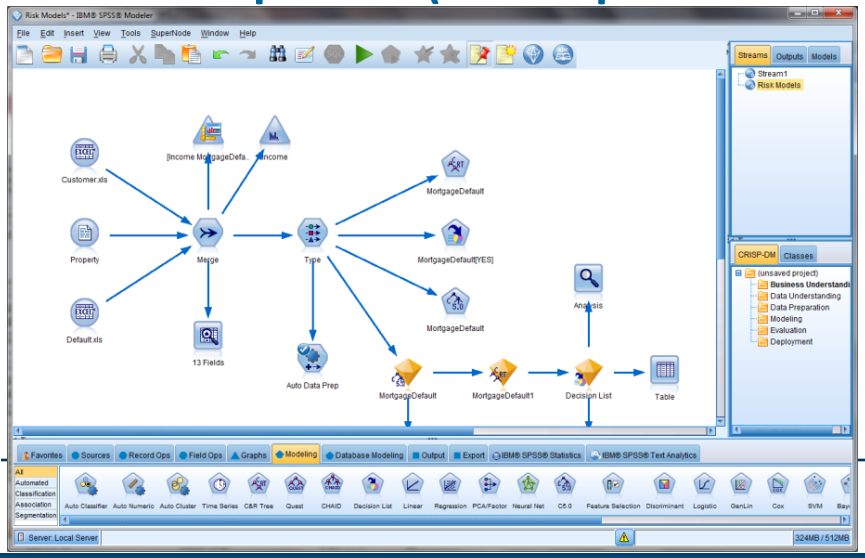
SPSS Products and Components – Modeler

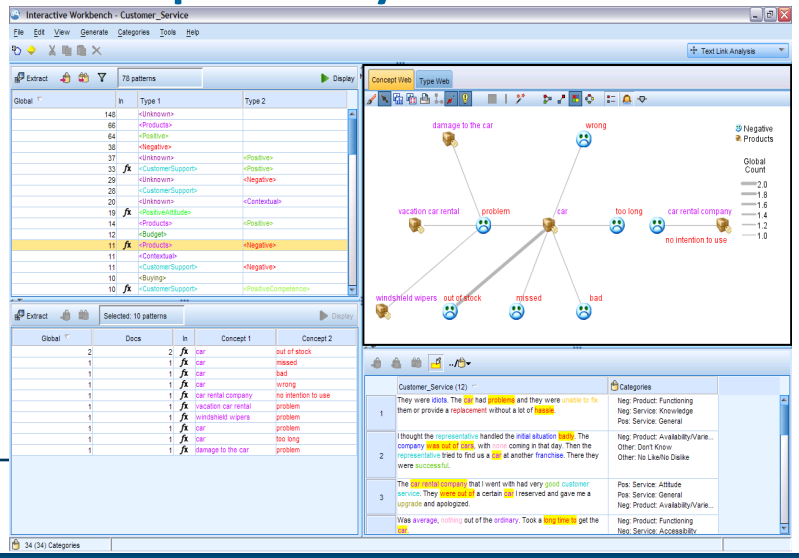
Modeler

Modeler Desktop: a rich desktop application for developing models. Note that Modeler Desktop is self-sufficient; it does not require a connection to the Server. It is not strictly a "client" in the usual sense.

Modeler Server: server component used for building models and for executing models in batch mode. Modeler Desktop can connect to this server. Modeler Server is also capable of performing batch scoring, Analytical Decision Management (together with C&DS), and other features.

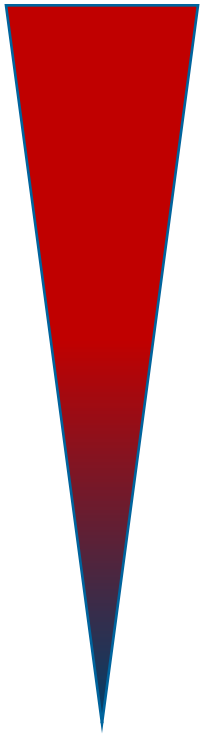
Each component (Desktop and Server) can run independently of the other.





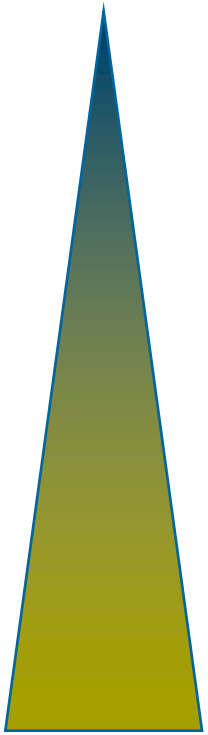
Statistical Analysis and Modeling Both Drive Predictive Analytics

Statistics Approach



- Statistical approach involves:
  - forming a theory about a possible relationship
  - converting the theory to a hypothesis
  - testing that hypothesis, using statistical methods
- It is a manual, user-driven, top-down approach to data analysis
- Used to ***test*** hypotheses

Modeling Approach

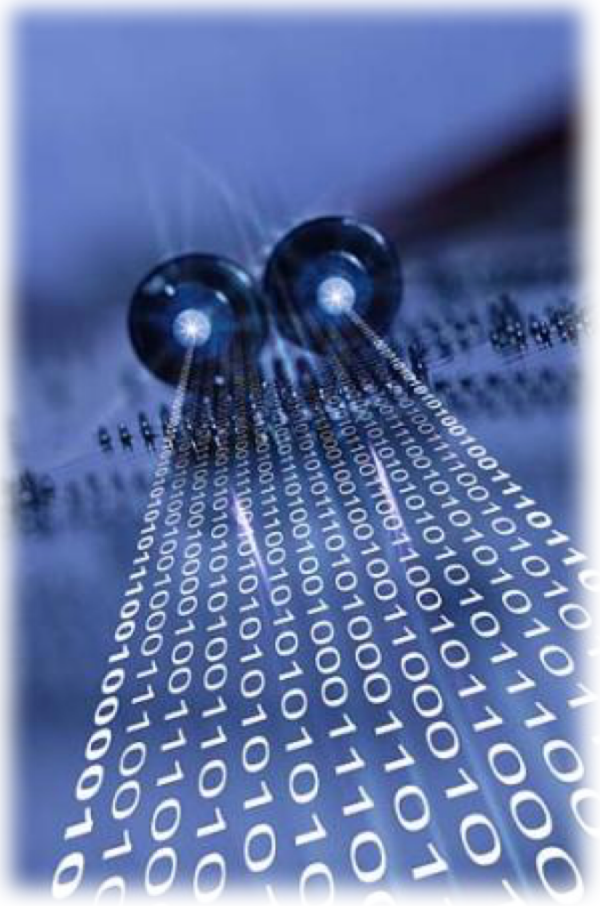


- Data mining involves:
  - the interrogation of the data
  - determined by the method and goal, rather than by the user
  - listening to the voice of the data
- It is a data-driven, self-organizing, bottom-up approach to data analysis
- Used to ***generate*** hypotheses

Note that Both Approaches Drive Predictive Analytics

Source: "Statistical Modeling: The Two Cultures," Leo Breiman, *Statistical Science*, 2001, Vol.16 (3), pp.199-231.

What is predictive analytics?



Predictive Analytics helps connect **data** to **effective action** by drawing reliable conclusions about current conditions and future events

Gareth Herschel, Research Director, Gartner Group

Enabling businesses to use **predictive models** to exploit patterns found in historical data to **identify** potential **risks** and **opportunities** before they occur.

Areas for Predictive Analytics

Operations

Warranty Claims
Product quality assurance
Predictive maintenance to predict equipment failure

Human Resources

Who are our best employees?
How do we keep our best employees from leaving?
Which prospects to recruit?

Risk Management

Network Intrusion

Education

Which students are likely to drop out and why?
Who are the best students and how to recruit more of them?
Which alumni are likely to make larger donations?

Public Safety

Crime analysis
Predictive policing

Finance

Money laundering
Tax audits & collections

Marketing

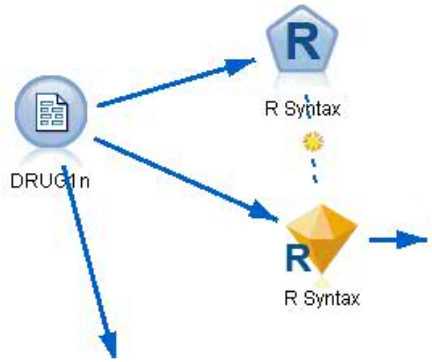
Who are our best customers?
Can we get more like that?
What/why do they buy?
Why do they leave?

.... and many more

Modeling Techniques in IBM SPSS Modeler

Technique	Usage	Algorithms
Classification (or prediction)	• Used to predict group membership (e.g., will this employee leave?) or a number (e.g., how many widgets will I sell?)	• Auto Classifiers, Decision Trees, Logistic, SVM, Time Series, etc.
Segmentation	• Used to classify data points into groups that are internally homogenous and externally heterogeneous. • Identify cases that are unusual	• Auto Clustering, K-means, etc. • Anomaly detection
Association	• Used to find events that occur together or in a sequence (e.g., market basket)	• APRIORI, Carma, Sequence

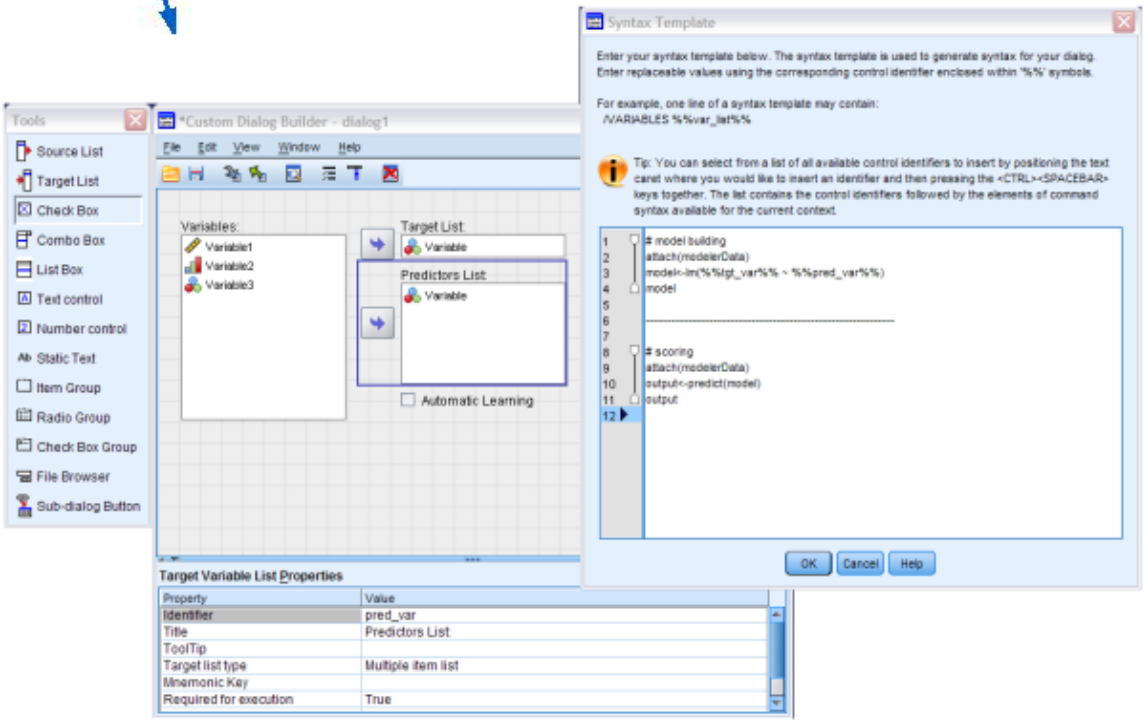
Extend Capabilities through Open Source (R, Python)



R Integration

R Build/Score, Process and Output node support

Scale R execution by leveraging database vendor provided R engines



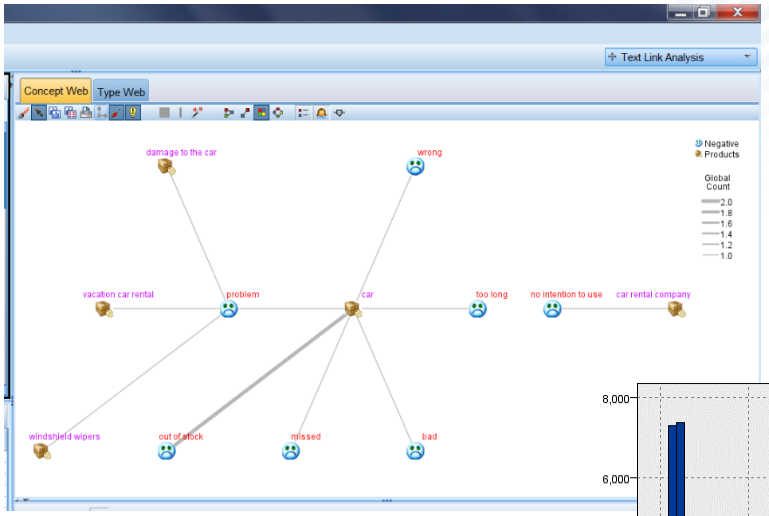
Custom Dialog Builder for R or Python

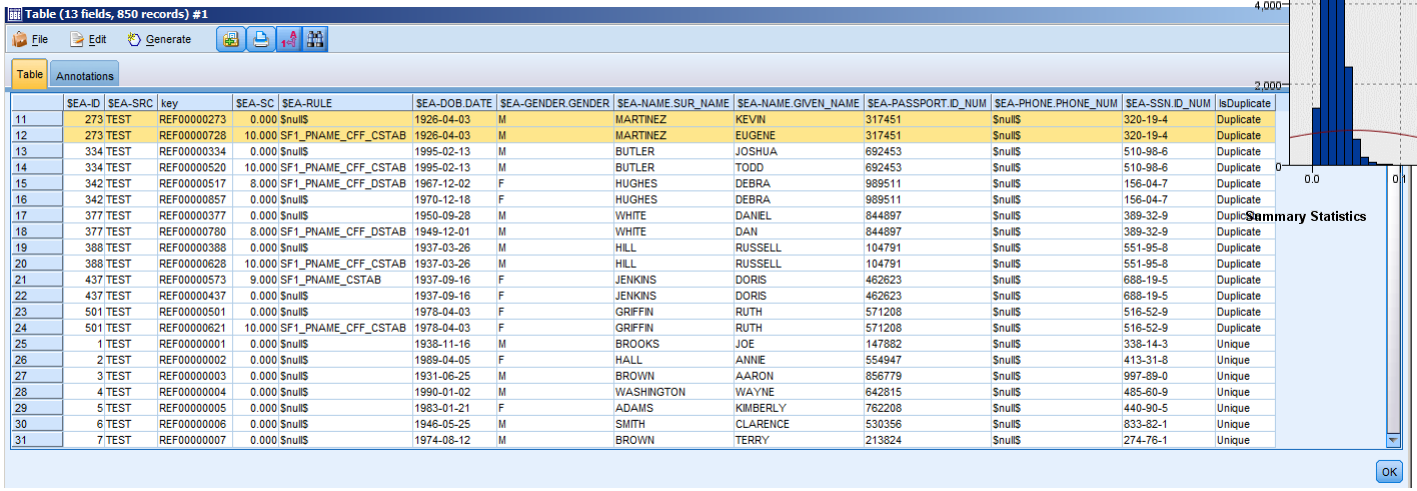
Provides the ability to create new Modeler Algorithm nodes and dialogs that run R processes

Makes R usable for non-programmers

Uncovering Patterns in Unstructured Data

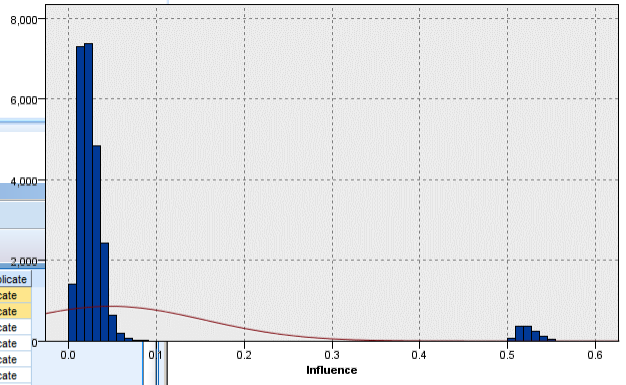
- **Text Analytics**
  - Natural Language
  - Sentiment Analysis
- **Social Network Analysis**
  - Uncover relationships
  - Find leaders and followers





	$EA-ID	$EA-SRC	key	$EA-SC	$EA-RULE	$EA-DOB.DATE	$EA-GENDER.GENDER	$EA-NAME.SUR_NAME	$EA-NAME.GIVEN_NAME	$EA-PASSPORT.ID_NUM	$EA-PHONE.PHONE_NUM	$EA-SSN.ID_NUM	IsDuplicate
11	273	TEST	REF00000273	0.000	$null$	1926-04-03	M	MARTINEZ	KEVIN	317451	$null$	320-19-4	Duplicate
12	273	TEST	REF00000728	10.000	SF1_PNAME_CFF_CSTAB	1926-04-03	M	MARTINEZ	EUGENE	317451	$null$	320-19-4	Duplicate
13	334	TEST	REF00000334	0.000	$null$	1995-02-13	M	BUTLER	JOSHUA	692453	$null$	510-98-6	Duplicate
14	334	TEST	REF00000520	10.000	SF1_PNAME_CFF_CSTAB	1995-02-13	M	BUTLER	TODD	692453	$null$	510-98-6	Duplicate
15	342	TEST	REF00000517	8.000	SF1_PNAME_CFF_DSTAB	1967-12-02	F	HUGHES	DEBRA	989511	$null$	156-04-7	Duplicate
16	342	TEST	REF00000857	0.000	$null$	1970-12-18	F	HUGHES	DEBRA	989511	$null$	156-04-7	Duplicate
17	377	TEST	REF00000377	0.000	$null$	1950-09-28	M	WHITE	DANIEL	844897	$null$	389-32-9	Duplicate
18	377	TEST	REF00000780	8.000	SF1_PNAME_CFF_DSTAB	1949-12-01	M	WHITE	DAN	844897	$null$	389-32-9	Duplicate
19	388	TEST	REF00000388	0.000	$null$	1937-03-26	M	HILL	RUSSELL	104791	$null$	551-95-8	Duplicate
20	388	TEST	REF00000628	10.000	SF1_PNAME_CFF_CSTAB	1937-03-26	M	HILL	RUSSELL	104791	$null$	551-95-8	Duplicate
21	437	TEST	REF00000573	9.000	SF1_PNAME_CSTAB	1937-09-16	F	JENKINS	DORIS	462623	$null$	688-19-5	Duplicate
22	437	TEST	REF00000437	0.000	$null$	1937-09-16	F	JENKINS	DORIS	462623	$null$	688-19-5	Duplicate
23	501	TEST	REF00000501	0.000	$null$	1978-04-03	F	GRIFFIN	RUTH	571208	$null$	516-52-9	Duplicate
24	501	TEST	REF00000621	10.000	SF1_PNAME_CFF_CSTAB	1978-04-03	F	GRIFFIN	RUTH	571208	$null$	516-52-9	Duplicate
25	1	TEST	REF00000001	0.000	$null$	1938-11-16	M	BROOKS	JOE	147882	$null$	338-14-3	Unique
26	2	TEST	REF00000002	0.000	$null$	1989-04-05	F	HALL	ANNIE	554947	$null$	413-31-8	Unique
27	3	TEST	REF00000003	0.000	$null$	1931-06-25	M	BROWN	AARON	856779	$null$	997-89-0	Unique
28	4	TEST	REF00000004	0.000	$null$	1990-01-02	M	WASHINGTON	WAYNE	642815	$null$	485-60-9	Unique
29	5	TEST	REF00000005	0.000	$null$	1983-01-21	F	ADAMS	KIMBERLY	762208	$null$	440-90-5	Unique
30	6	TEST	REF00000006	0.000	$null$	1946-05-25	M	SMITH	CLARENCE	530356	$null$	833-82-1	Unique
31	7	TEST	REF00000007	0.000	$null$	1974-08-12	M	BROWN	TERRY	213824	$null$	274-76-1	Unique

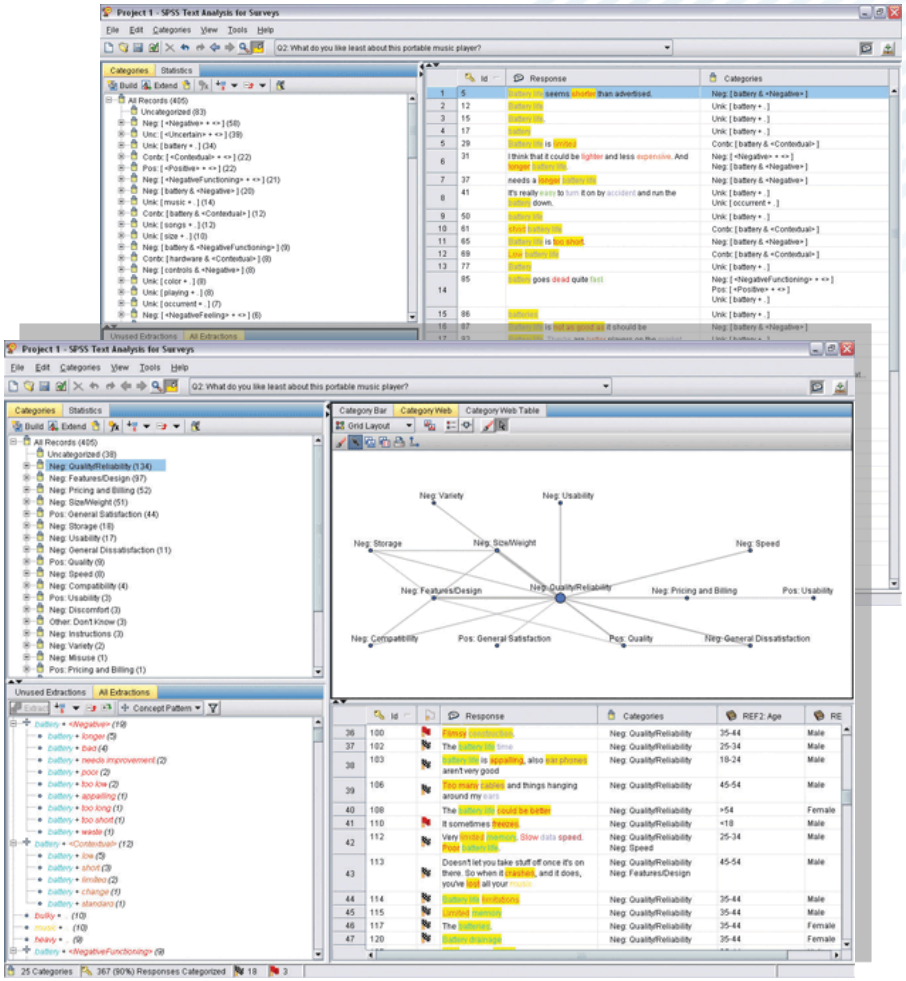
OK



IBM Software Solutions Group

Text Analytics

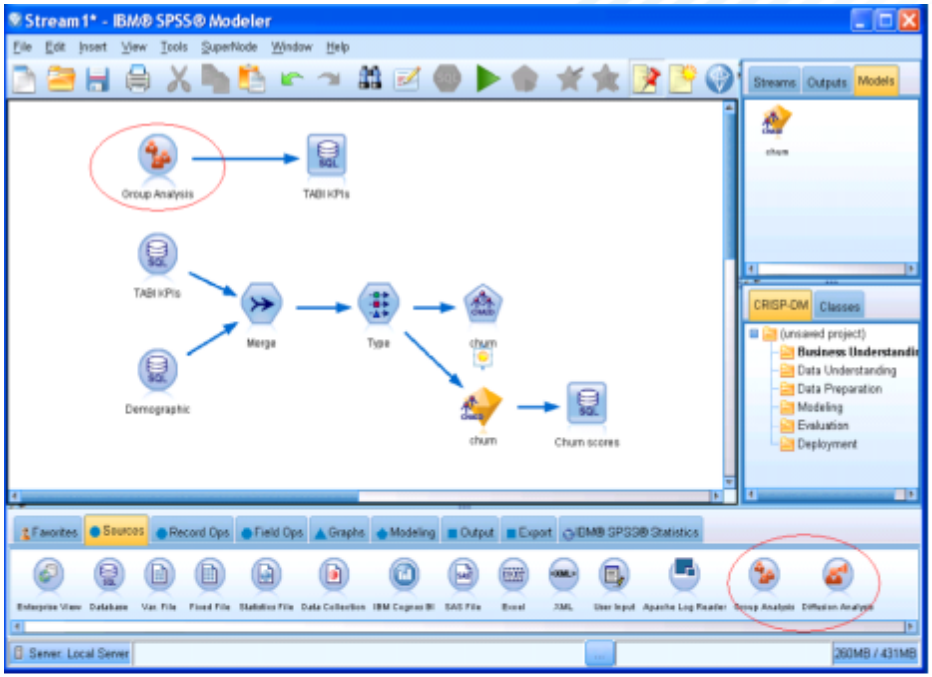
- Uses natural language processing heuristic rules and statistical techniques to reveal conceptual meaning in text
- Extracts concepts from text and categorizes them
- Makes unstructured qualitative data more quantifiable, enabling the discovery of key insights from sources such as survey responses, documents, emails, call center notes, web pages, blogs, forums and more

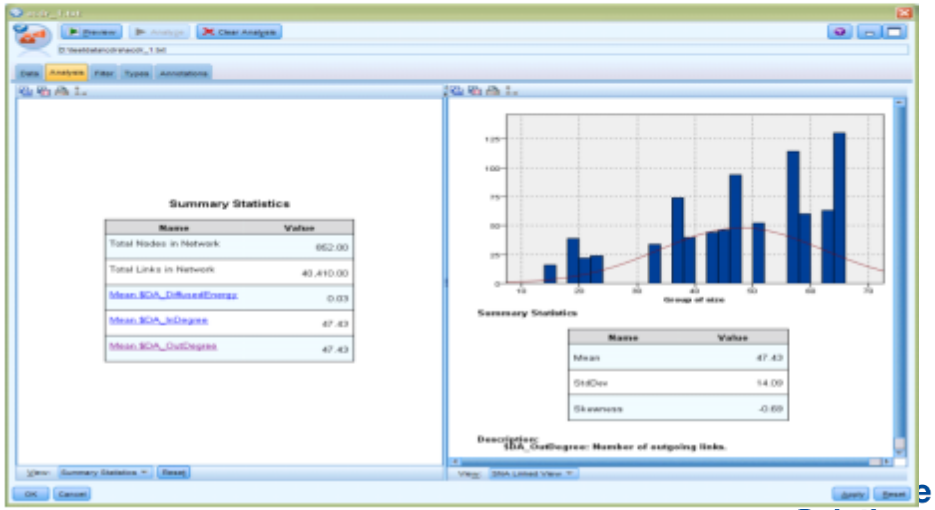


Brings repeatability to ongoing decision making

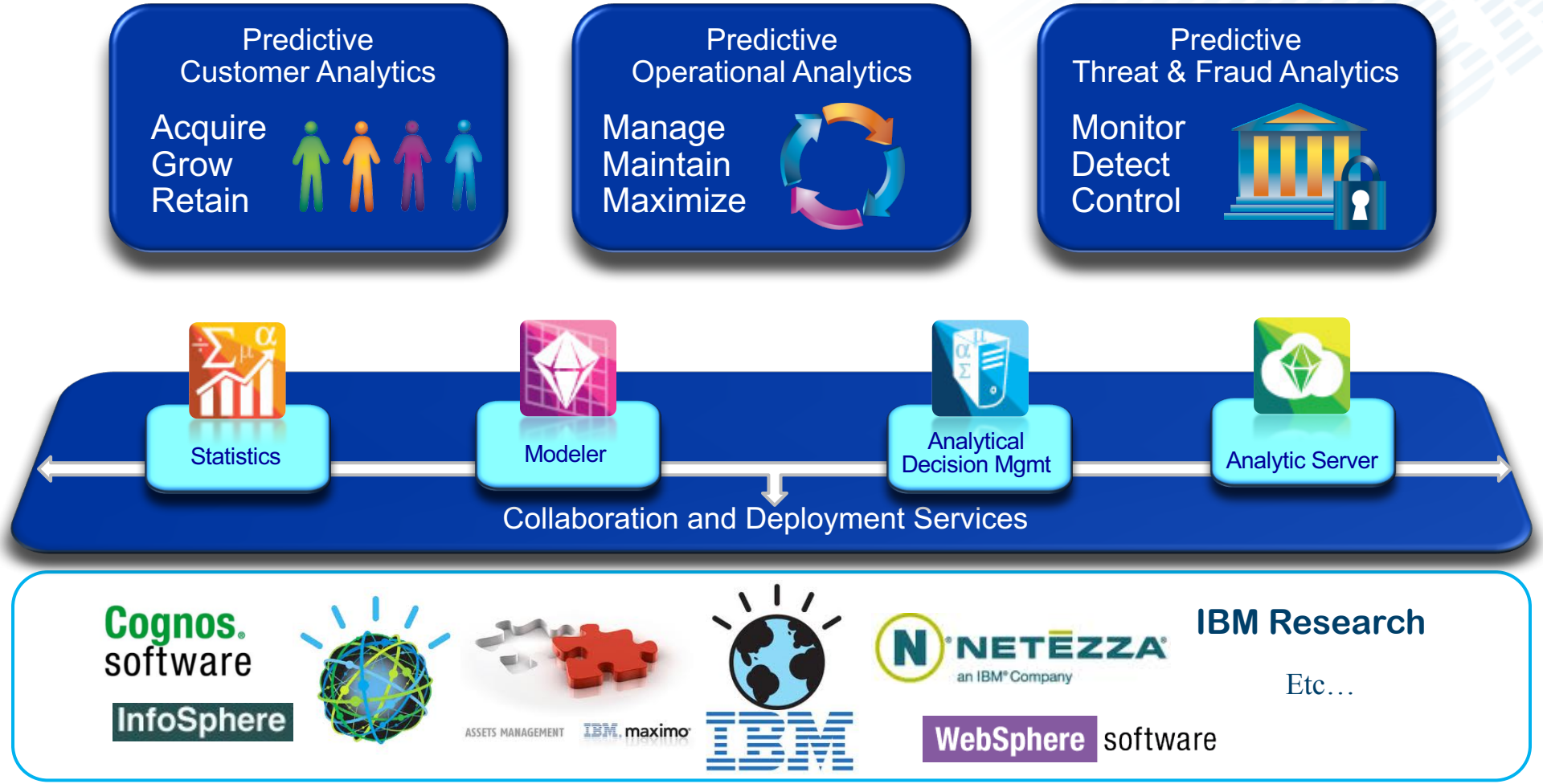
Social Network Analysis

- Processes CDR (Call Data Record) data companies to produce social analysis
- Focuses around identifying groups, leaders and probabilities that others will churn based on influence
- Enhances existing churn predictions of Modeler
- Expressed as two new nodes in the Sources Palette
  - Group Analysis – what are the groups in my data and who are the leaders
  - Diffusion Analysis – uses existing churn information to determine who else that churner is likely to influence to leave



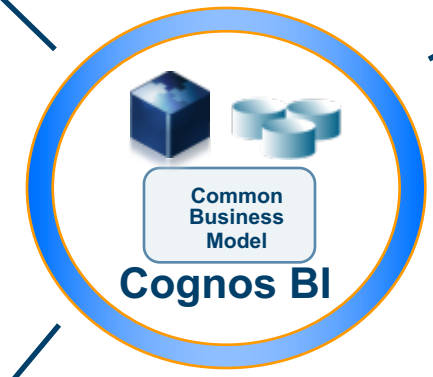


IBM SPSS Modeler Within the Business Analytics Ecosystem

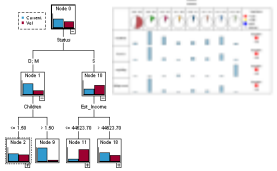


Integration with IBM Cognos

1) Leveraging BI, identify problem or situation needing attention



2) SPSS predictive analytics feed results back into the BI layer



3) Results widely distributed via BI for consumption by business Users

Overview takeaways

Easy to use, visual interface

- Short timeframe to be productive with actionable results
- Does not require knowledge of programming language
- No proprietary data formats
- Open architecture

Business results focused

- Leverages the investments already made in technology
- Cost effective solution that delivers powerful results across organization
- Full range of algorithms for your business problems
- Big Data enabled (Hadoop, SQL Pushback)

End-to-end solution

- Data preparation through real time interactions
- Use structured, unstructured and semi-structured data
- Integrated portfolio for business analytics
- Scales from a single desktop to an enterprise deployments

SPSS Agreement with BCNET

- Members pay the same fee as before, for three years, without the 6% annual increase
- After year three members can exit the service if they wish
- After three years, members only pay the maintenance fee component and a small BCNET administration fee for their seats (less than 25% of current fees)
- How you use the licenses will not change
- Any renewals or new SPSS purchases will be completed through BCNET