TAKING YOUR APPLICATION DESIGN TO THE NEXT LEVEL WITH DATA MINING

Peter Myers

Mentor – SolidQ

Los Angeles SQL Server Professional Groups – 19 May, 2011

AGENDA

Data mining as a technology supports the discovery of patterns and statistics not easily visible in relational database queries, and in many cases these patterns can be used to deliver predictions. These patterns can be used to derive knowledge about data, and in turn this knowledge can be used to enhance application designs and the user experience.

In this presentation you will be introduced to Analysis Services 2005 and 2008 Data Mining and numerous demonstrations that show how to develop data mining models that can be embedded into your applications. Demonstrations will be based on Analysis Services 2008 Data Mining. This presentation is a must for any developer looking to embed "Artificial Intelligence" into their solution design to take their applications to the next level. It is designed to thrill you with potential, and excite you with the ease in which it can be accomplished. The demonstrations range from simple (involving no code!) to more sophisticated examples.

This presentation is targeted at developers with an interest in data mining, and equally for non-developers interested to understand and evaluate what data mining could achieve for them. There is no requirement to have any background or experience with data mining technologies. It will be advantageous, but not necessary, to attend the lunchtime presentation: Introduction to Analysis Services Data Mining.

PRESENTER INTRODUCTION

- Peter Myers
- Mentor, SolidQ
- BBus, MCITP (Dev, DBA, BI), MCT, MVP
- 14 years of experience designing and developing software solutions using Microsoft products, today specializing in Microsoft Business Intelligence
- Based in San Francisco
- pmyers@solidq.com

WHO WE ARE

- **Industry experts:**

  Growing, elite group of the world's best technical experts who, as reflected by the high concentration of Microsoft MVPs and RDs in our ranks, achieve excellence in their industry by maintaining the highest credentials

- **Published authors:**

  Technical reference books, Microsoft reference and training materials, industry white papers, technical magazine articles, and webcasts

- **Top technical speakers:**

  PASS Community Summit, Microsoft TechEd, The Microsoft BI Conference, SQL Server Connections, and countless user groups, international conferences and events

WHAT WE DO

Provide advanced, world-class expertise across the entire Microsoft relational data and development platforms and complimenting technologies

PRACTICE AREAS		SERVICES
Relational Database Management		Advanced, Public Training
Business Intelligence		Customized, Private Training
Development Methodologies		Solution Delivery & Tuning
SharePoint Collaboration		Enhanced, Mentoring Services

For more information visit www.solidq.com

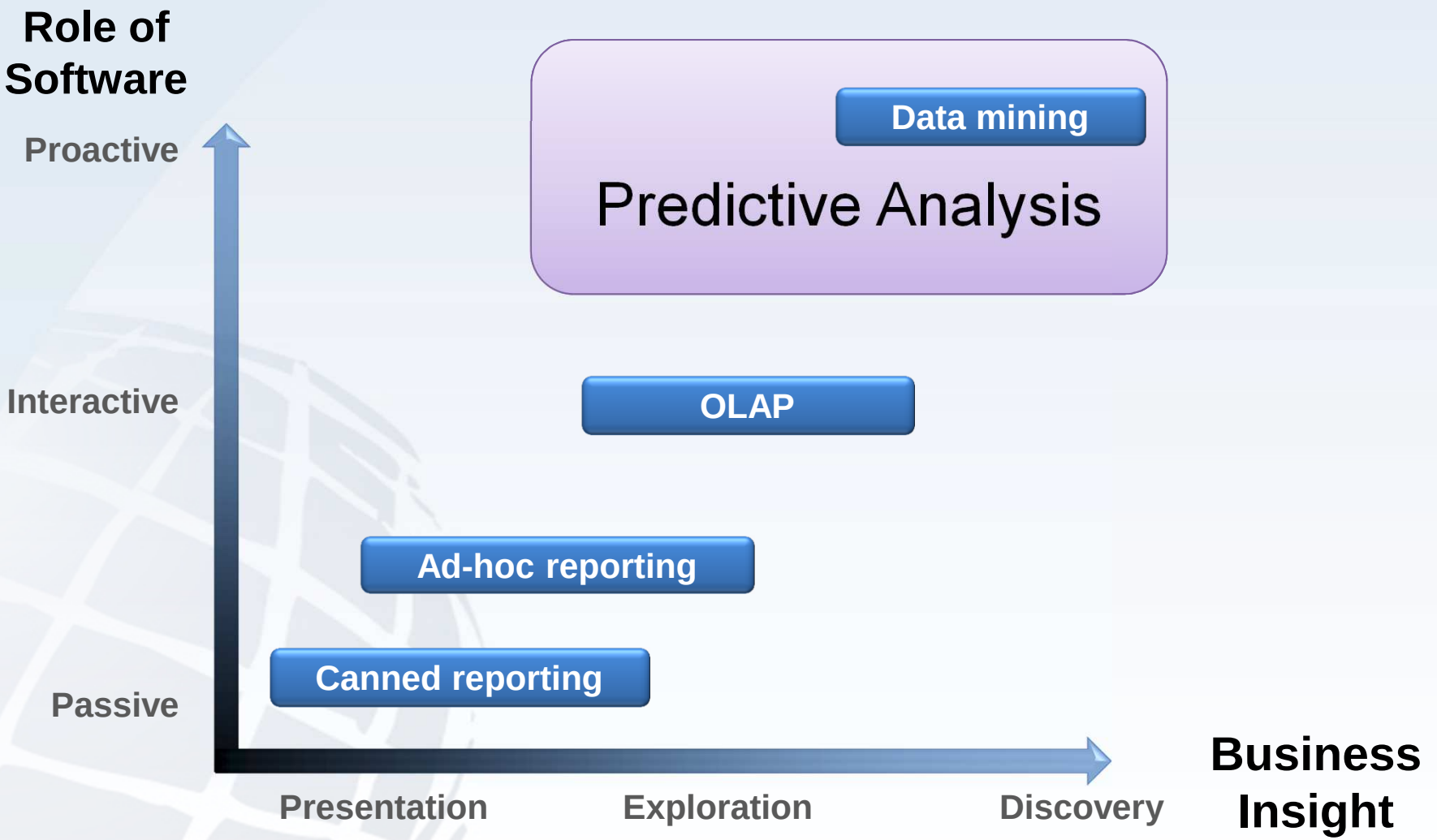
AGENDA

- Introducing Data Mining
- Describing the Data Mining Process
- SQL Server 2008 Data Mining
- Data Preparation
- Data Mining Visualization
- Data Mining Programmability
- Demonstrations

INTRODUCING DATA MINING

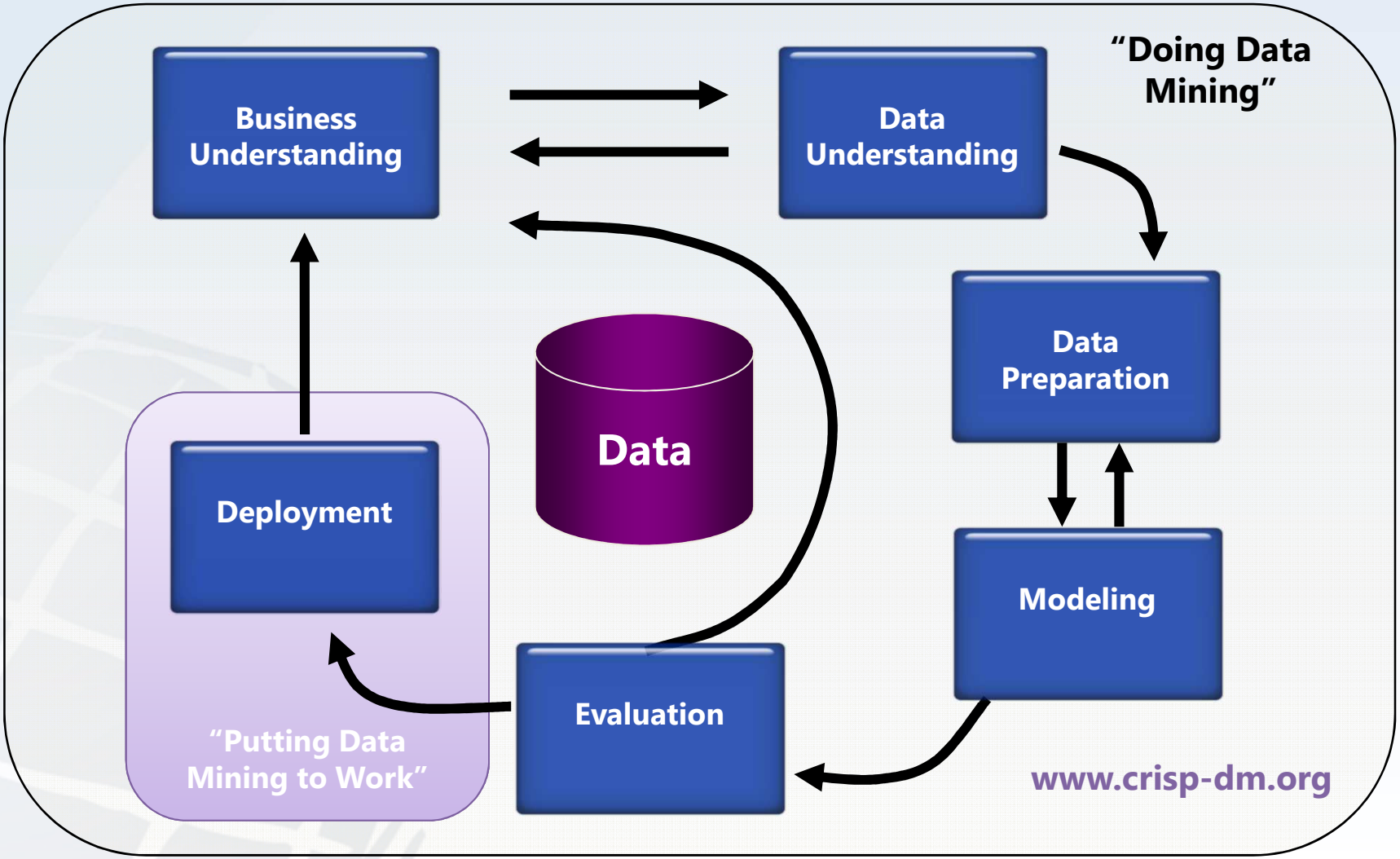
- Addresses the problem:
  "Too much data and not enough information"
- Enables data exploration, pattern discovery, and pattern prediction—which lead to knowledge discovery
- Forms a key part of a BI solution

DATA MINING ENABLES PREDICTIVE ANALYSIS



BUSINESS SCENARIOS

DESCRIBING THE DATA MINING PROCESS



DATA PREPARATION

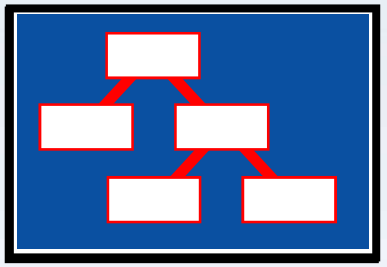
- Often significant amounts of effort are required to prepare data for mining:
  - Transforming for cleaning and reformatting
  - Isolating and flagging abnormal data
  - Appropriately substituting missing values
  - Discretizing continuous values into ranges
  - Normalizing values between 0 and 1
- Of course, having the required data to begin with is important:
  - When designing systems, give consideration to attributes that may be required as inputs for classification
    - For example, demographic data: Age, Gender, Region, etc

MODELING

Design time

Process time

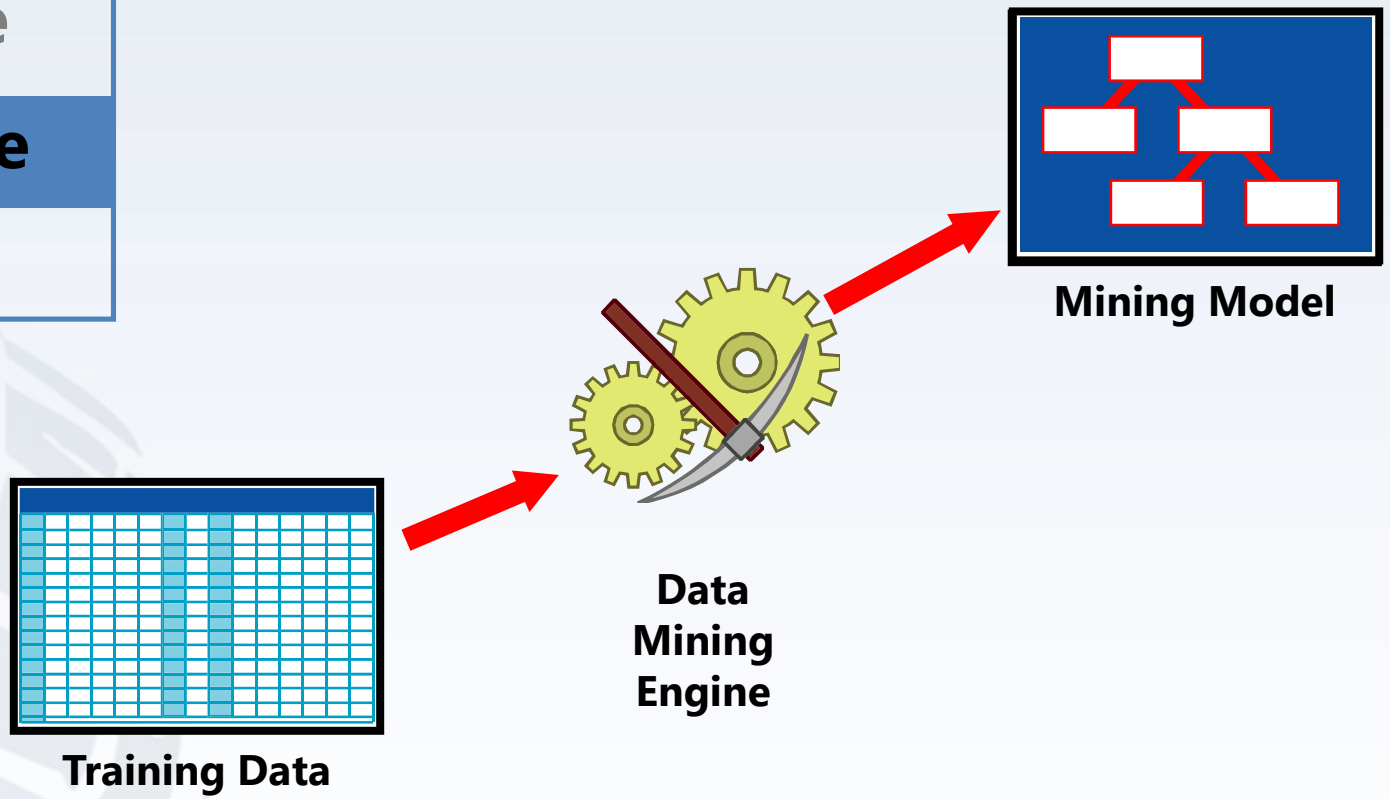
Query time



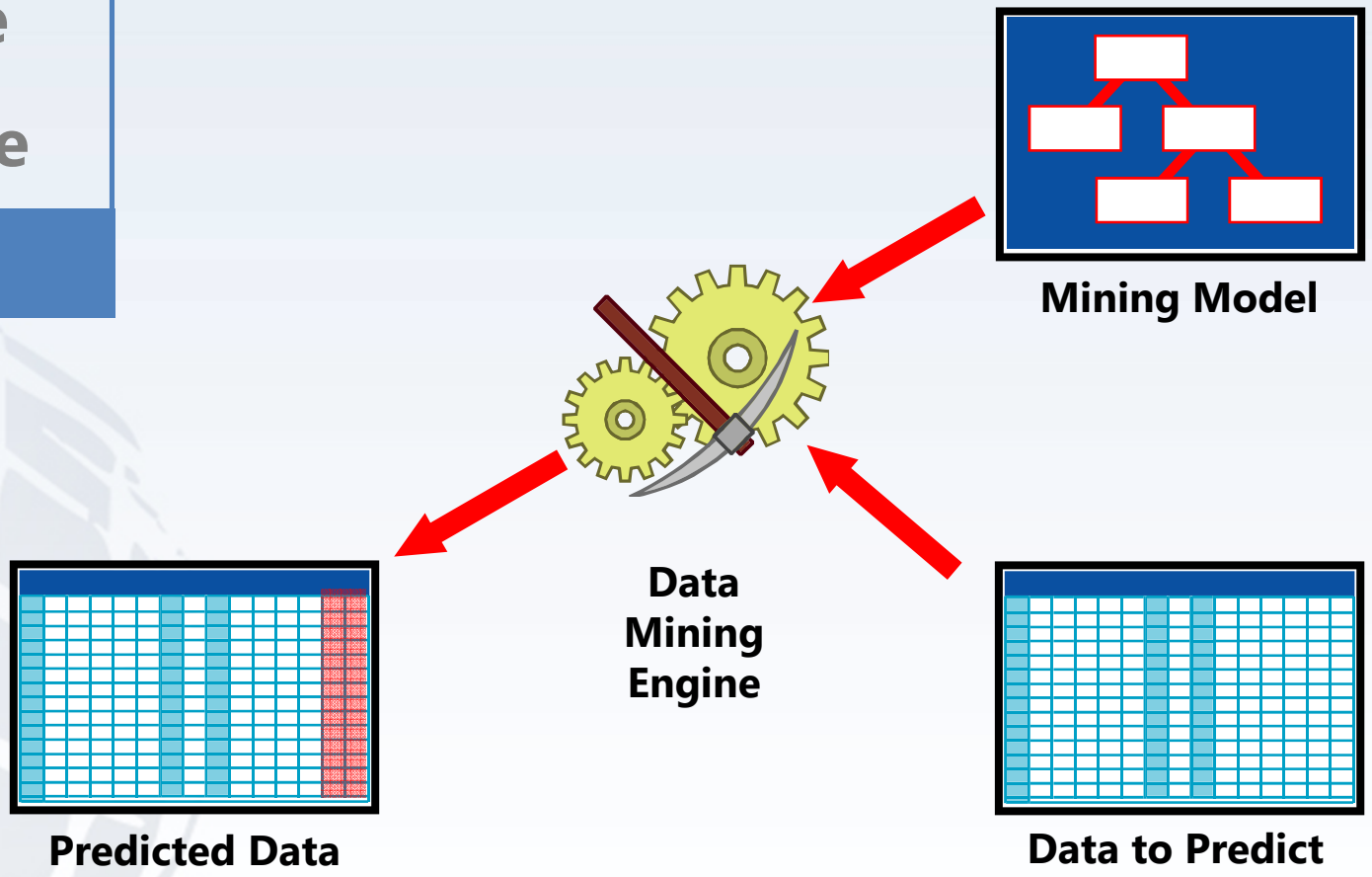
Mining Model

MODELING

- Design time
- **Process time**
- Query time



- Design time
- Process time
- **Query time**



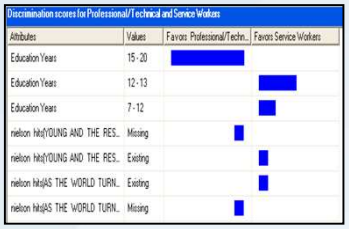
MODEL VALIDATION

- It is important that the model makes sense
  - Accuracy
    - Does it correlate and predict correctly?
  - Reliability
    - Does it work similarly for different test data?
  - Usefulness
    - Does it provide insight or only obvious trivialities?
- Commonly a holdout set of data is used to test model accuracy

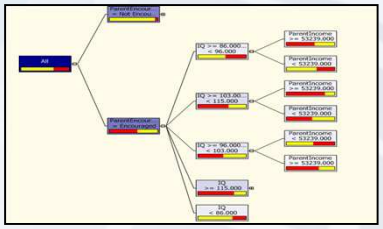
SQL SERVER 2008 DATA MINING

- Hides the complexity of an advanced technology
- Includes full suite of algorithms to automatically extract information from data
- Handles large volumes of data and complex data
- Data can be sourced from relational and OLAP databases
- Uses standard programming interfaces:
  - XMLA
  - DMX
- Delivers a complete framework for building and deploying intelligent applications

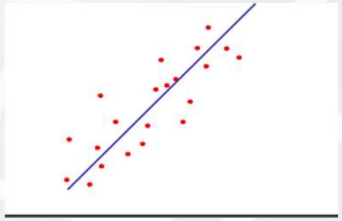
SQL SERVER 2008 ALGORITHMS



- Microsoft Naïve Bayes
  - Quick and approachable algorithm
  - Used for classification

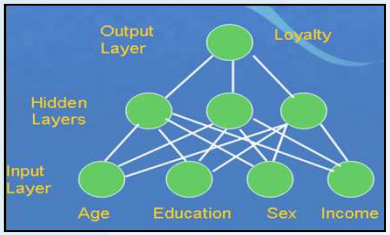


- Microsoft Decision Trees
  - Popular data mining technique
  - Used for classification, regression and association

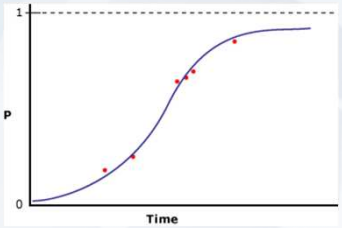


- Microsoft Linear Regression
  - Finds the best possible straight line through a series of points
  - Used for prediction analysis

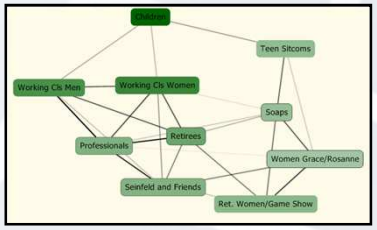
SQL SERVER 2008 ALGORITHMS



- Microsoft Neural Network
  - More sophisticated than Decision Trees and Naïve Bayes, this algorithm can explore extremely complex scenarios
  - Used for classification and regression tasks

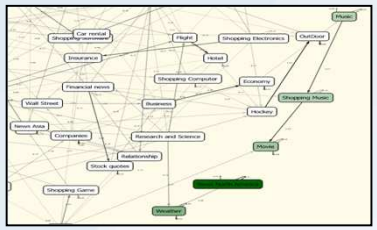


- Microsoft Logistic Regression
  - A particular case of the Neural Network algorithm

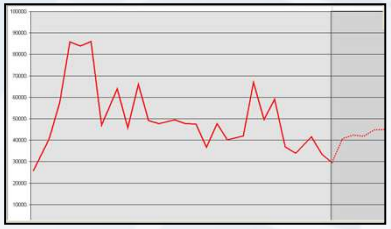


- Microsoft Clustering
  - Finds natural groupings inside data
  - Supports segmentation and anomaly detection tasks

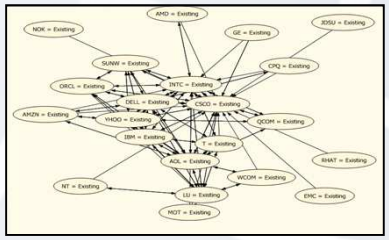
SQL SERVER 2008 ALGORITHMS



- Microsoft Sequence Clustering
  - Groups a sequence of discrete events into natural groups based on similarity



- Microsoft Time Series
  - Used to predict future values from a time series
  - Has been improved in SQL Server 2008 to produce more accurate long-term forecasts



- Microsoft Association Rules
  - Commonly supports market basket analysis to learn what products are purchased together

SQL SERVER 2008 ALGORITHMS

Classify	Estimate	Cluster	Forecast	Associate
• Decision Trees • Logistic Regression • Naïve Bayes • Neural Networks	• Decision Trees • Linear Regression • Logistic Regression • Neural Networks	• Clustering	• Time Series	• Association Rules • Decision Trees

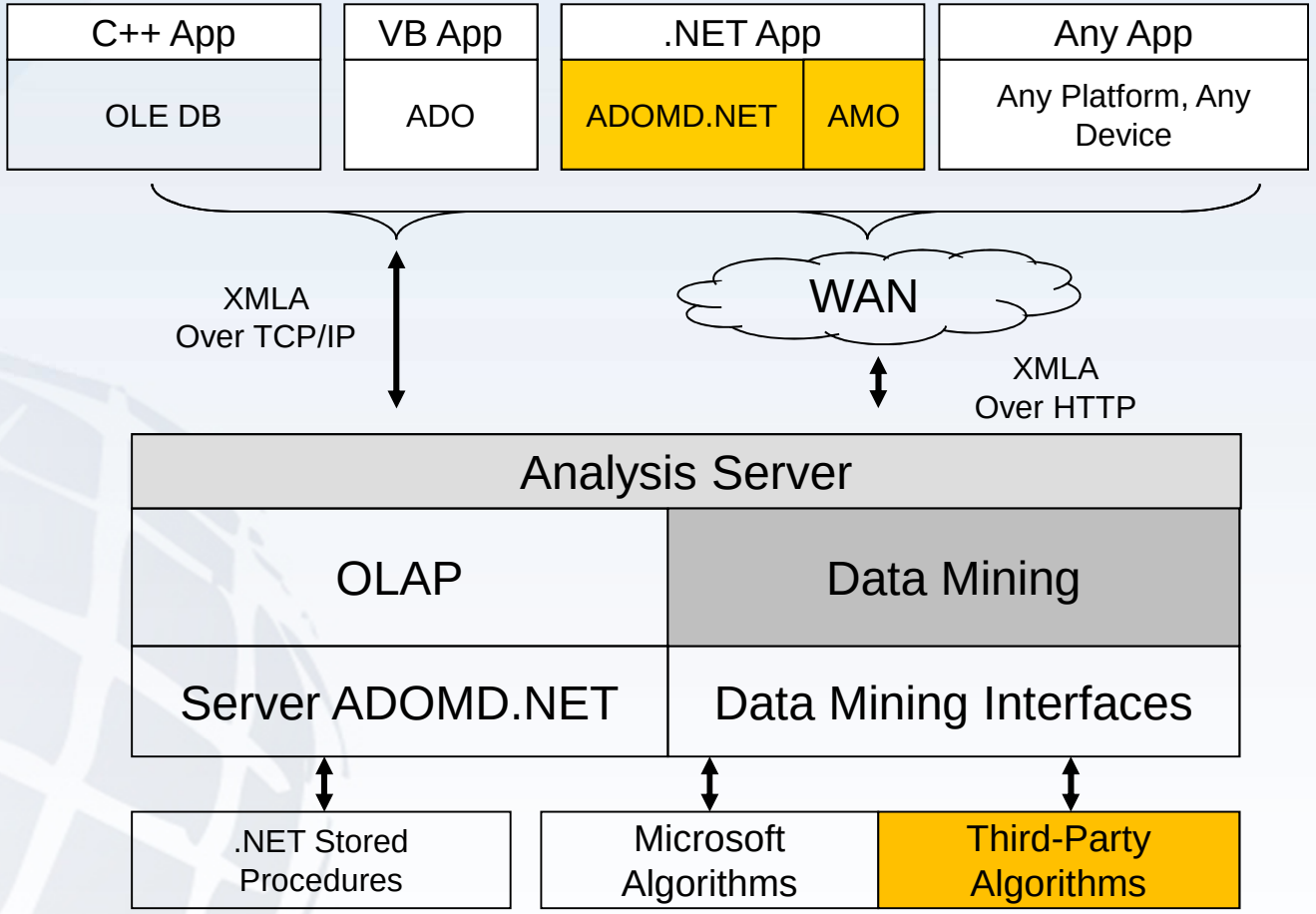
DATA MINING VISUALIZATION

- In contrast to OLTP and OLAP queries, data mining queries typically extract information that the user is not aware of
- Appreciate that end users do not typically query data mining models directly
- Visualizations can effectively present data discoveries
- SQL Server 2008 provides algorithm-specific visualizations that can:
  - Test and explore models in BIDS
  - Be embedded into Web and Windows Forms applications
- Developers can construct and plug-in custom data mining viewers

DATA MINING PROGRAMMABILITY

- Analysis Services APIs
- Embedding data mining into applications
- Extending the data mining capabilities

ANALYSIS SERVICES APIs



ANALYSIS SERVICES APIs

- AMO (Analysis Management Objects)
  - Administer database objects
  - Apply security
  - Manage processing
- ADOMD.NET
  - Connect to SSAS databases
  - Retrieve and manipulate data
- Server ADOMD.NET
  - Extend DMX by using .NET stored procedures

EMBEDDING DATA MINING INTO APPLICATIONS

- Embed data mining into intelligent applications:
  - Help validate or repair user entry
  - Integrate predictions
    - Targeted advertising
    - "Those that bought this book also purchased these books"
  - Embed custom visualizations into Windows Forms applications to allow users to explore and understand model patterns

EXTENDING THE DATA MINING CAPABILITIES

- Develop .NET stored procedures
- Enhance the Visual Studio data mining tools
- Develop plug-in algorithms and viewers

DEMONSTRATIONS

1. Creating, Training, Testing and Querying Mining Models with BIDS
2. Automating Data Validation With Data Mining
3. Enhancing an E-Commerce Site with Market Basket Analysis

RESOURCES

- www.microsoft.com/sql/technologies/dm
  - Links to technical resources, case studies, news, and reviews
- www.sqlserverdatamining.com
  - Site designed and maintained by the SQL Server Data Mining team
  - Includes: Live samples, tutorials, webcasts, tips and tricks, and FAQ
- www.predixionsoftware.com
  - Self-service predictive analytics in the cloud
  - Their Excel add-in picks up where the Microsoft add-ins have left off
- Data Mining for SQL Server 2008, by ZhaoHui Tang and Jamie MacLennan