Taking Your Application Design To The Next Level With Data Mining

Peter Myers

Mentor – Solid Quality Mentors

Orange County SQL Server
and .NET User Groups – 8 December, 2009

PRESENTER

- Peter Myers
- Mentor and Trainer, Solid Quality Mentors
- BBus, MCP, MCITP (DBA, Dev, BI), MCT, MVP
- 12 years' experience designing, developing and supporting software solutions using Microsoft data and development platforms
- pmyers@solidq.com

WHO WE ARE

- ***Industry experts:***

  Growing, elite group of over 90 of the world's best technical experts who, as reflected by the high concentration of Microsoft MVP's and RD's in our ranks, achieve excellence in their industry by maintaining the highest credentials.
- ***Published authors:***
  Best technical reference books, Microsoft reference materials, industry white papers, technical magazine articles, and webcasts.
- ***Top technical speakers:***
  PASS Community Summit, Microsoft TechEd, The Microsoft BI Conference, SQL Server DevConnections, countless user groups, international conferences and events.
- **For more information visit *www.solidq.com***

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*

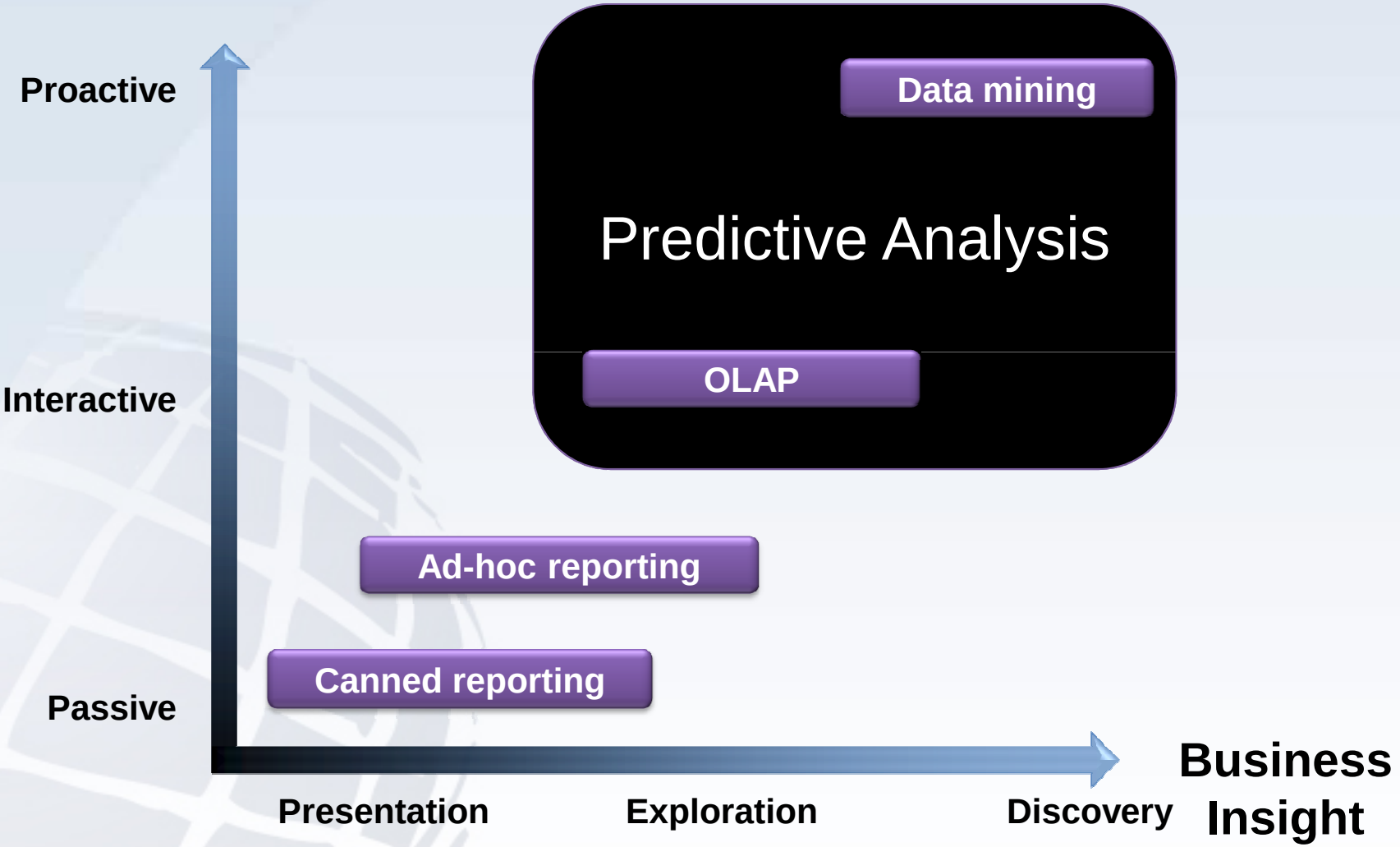
AGENDA

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

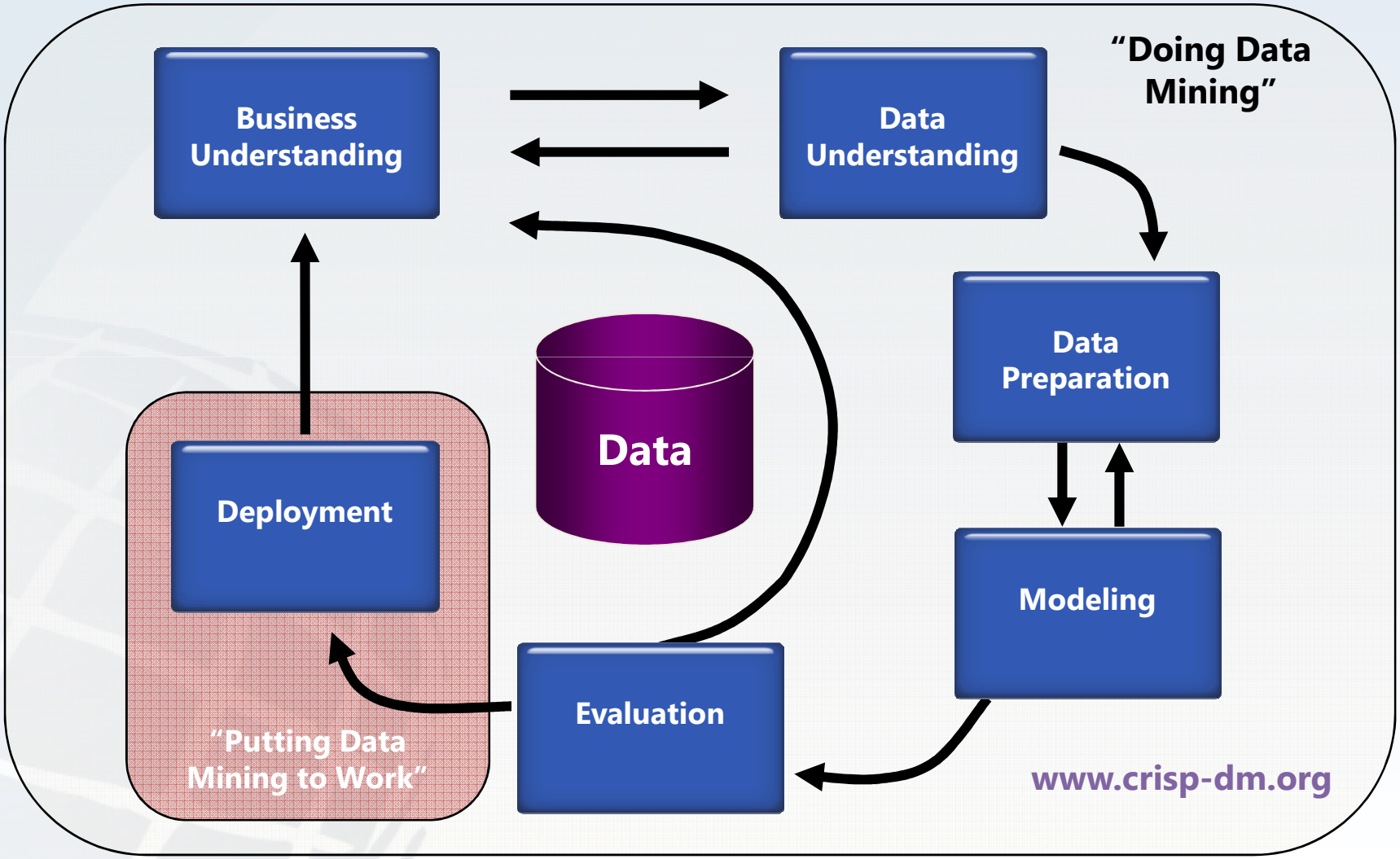
DATA MINING ENABLES PREDICTIVE ANALYSIS



BUSINESS SCENARIOS

- Identifying responsive customers/unresponsive customers (also known as churn analysis)
- Targeting promotions
- Detecting and preventing fraud
- Correcting data during ETL
- Forecasting sales and inventory
- Cross-selling

DESCRIBING THE DATA MINING PROCESS



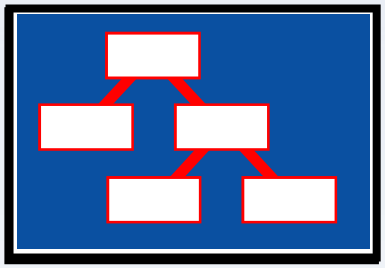
DATA PREPARATION

- 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

Design time

Process time

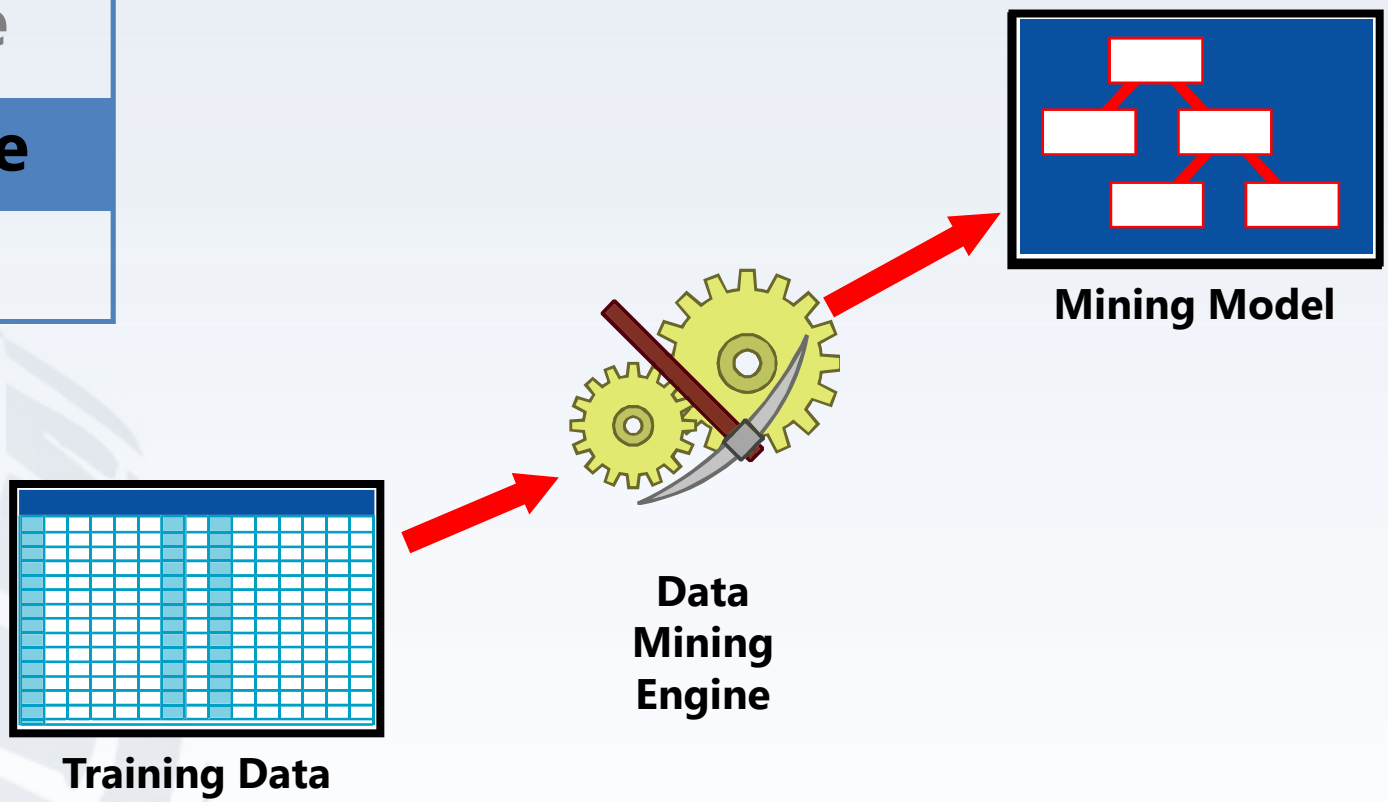
Query time



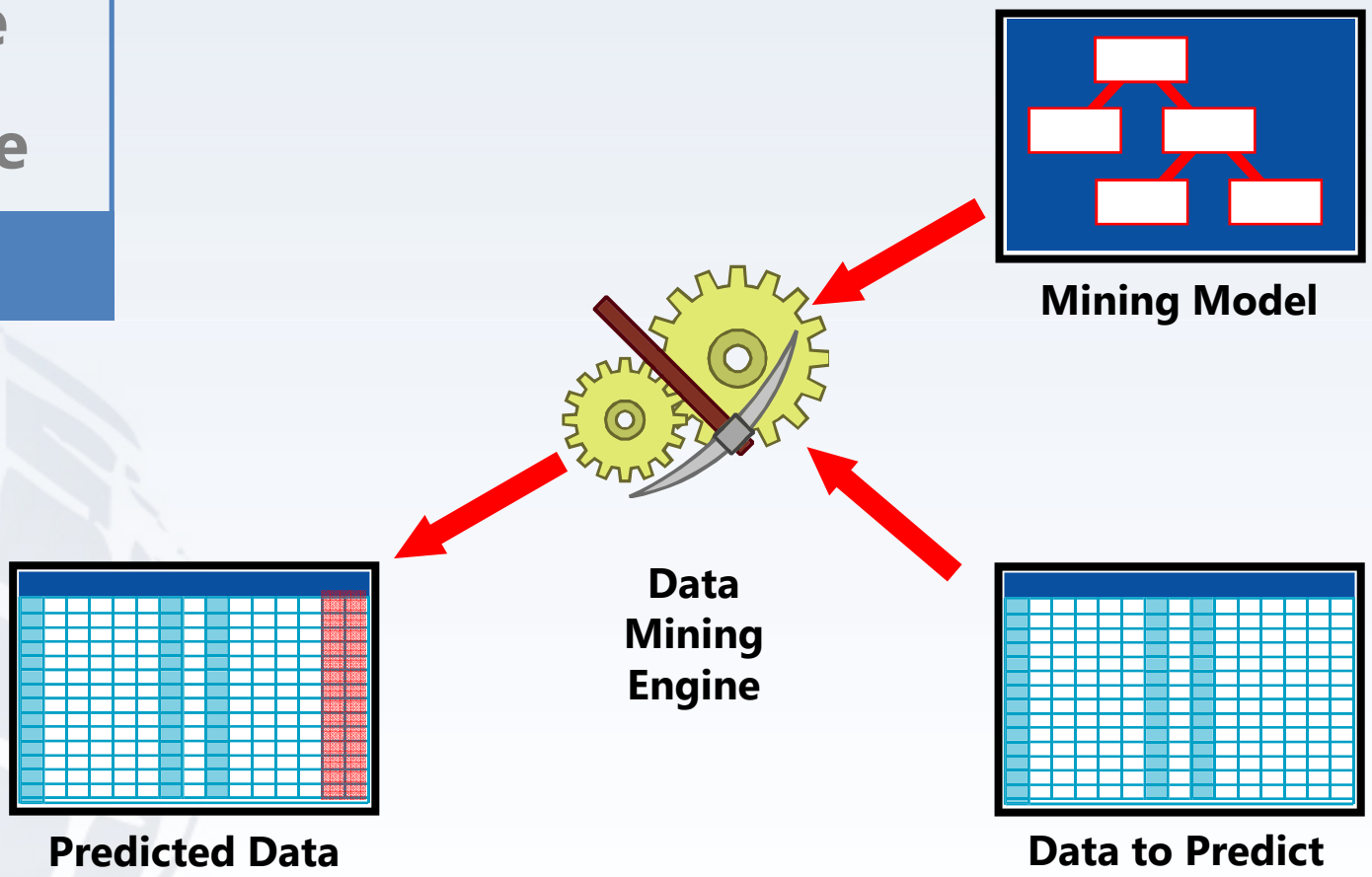
Mining Model

MODELING

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



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



MODEL VALIDATION

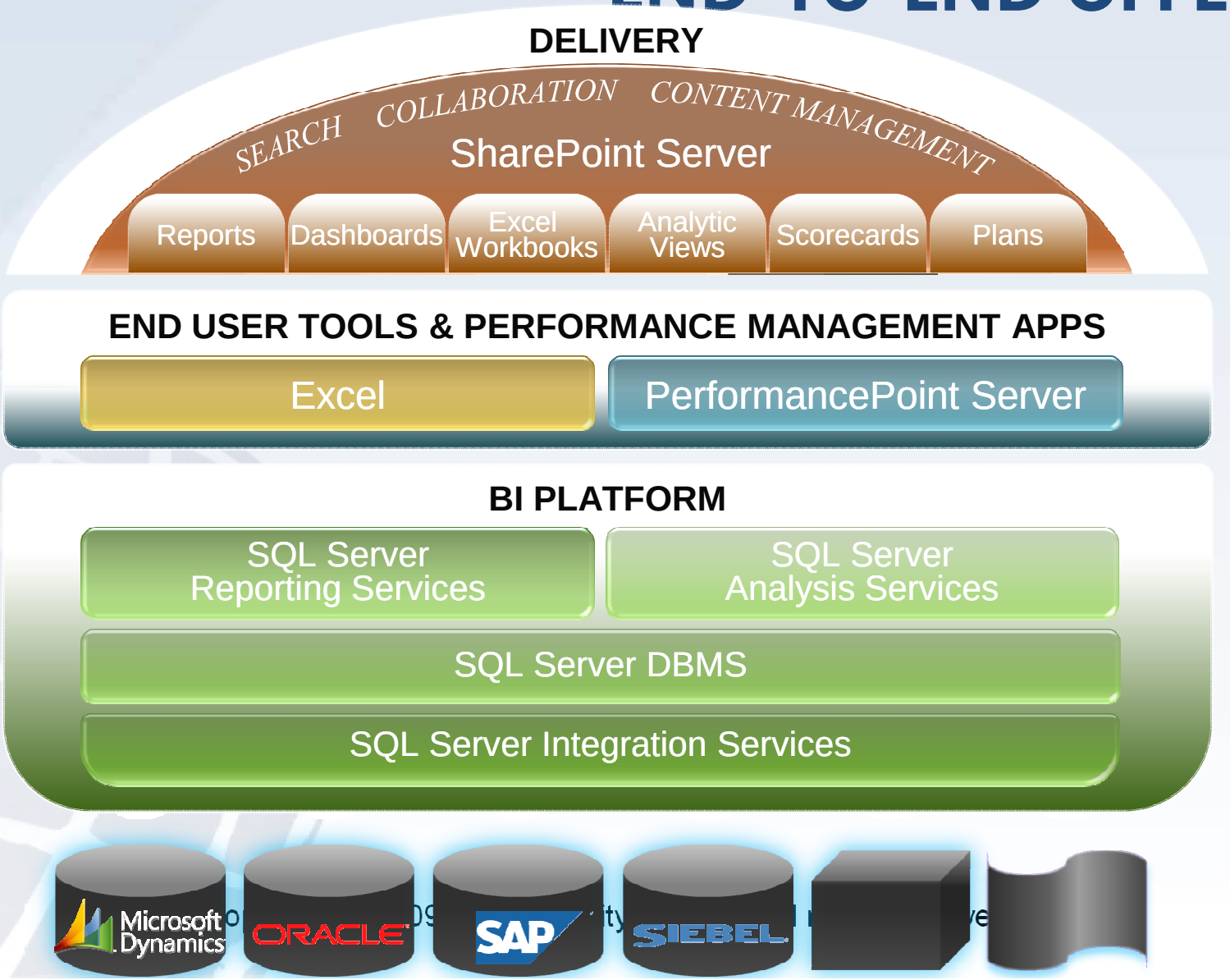
- It is important that the model makes sense
  - Accuracy
    - Does it correlate and predict correctly?
  - Reliability
    - Does it work similarly for different test data?
  - Usefulness
    - Does it provide insight or only obvious trivialities?
- Commonly a holdout set of data is used to test model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

SQL SERVER™ 2008 DATA MINING

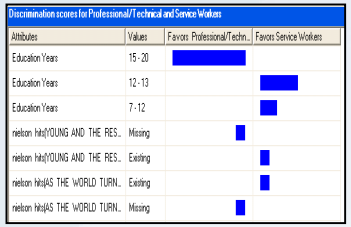
- Hides the complexity of an advanced technology
- Includes full suite of algorithms to automatically extract information from data
- Handles large volumes of data and complex data
- Data can be sourced from relational and OLAP databases
- Uses standard programming interfaces:
  - XMLA
  - DMX
- Delivers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

SOLID QUALITY MENTORS

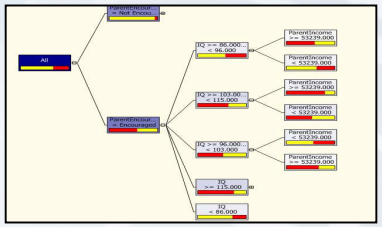
INTEGRATED END-TO-END OFFERING



SQL SERVER™ 2008 ALGORITHMS



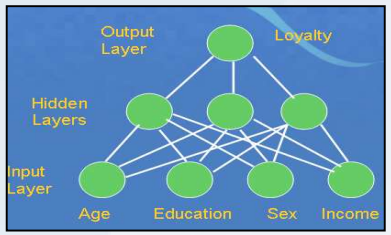
- Microsoft Naïve Bayes
  - Quick and approachable algorithm
  - Used for classification



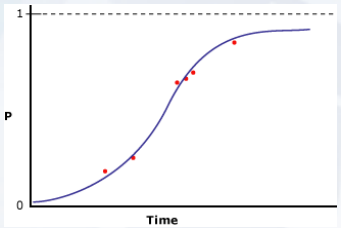
- Microsoft Decision Trees
  - Popular data mining technique
  - Used for classification, regression and association

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

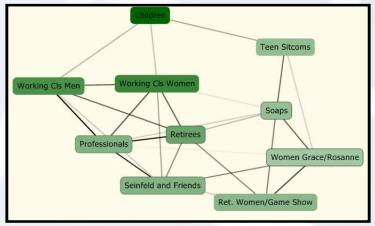
SQL SERVER™ 2008 ALGORITHMS



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

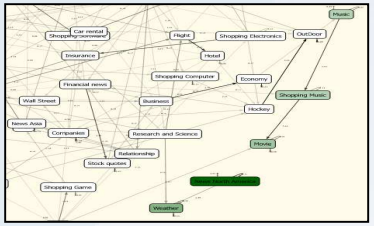


- Microsoft Logistic Regression
  - A particular case of the Neural Network algorithm

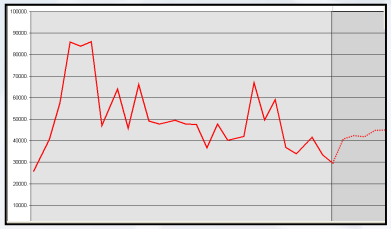


- Microsoft Clustering
  - Finds natural groupings inside data
  - Supports segmentation and anomaly detection tasks

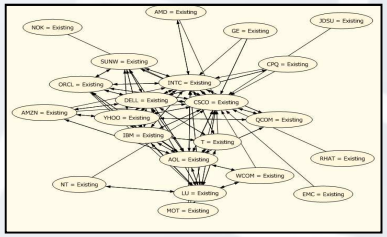
SQL SERVER™ 2008 ALGORITHMS



- 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

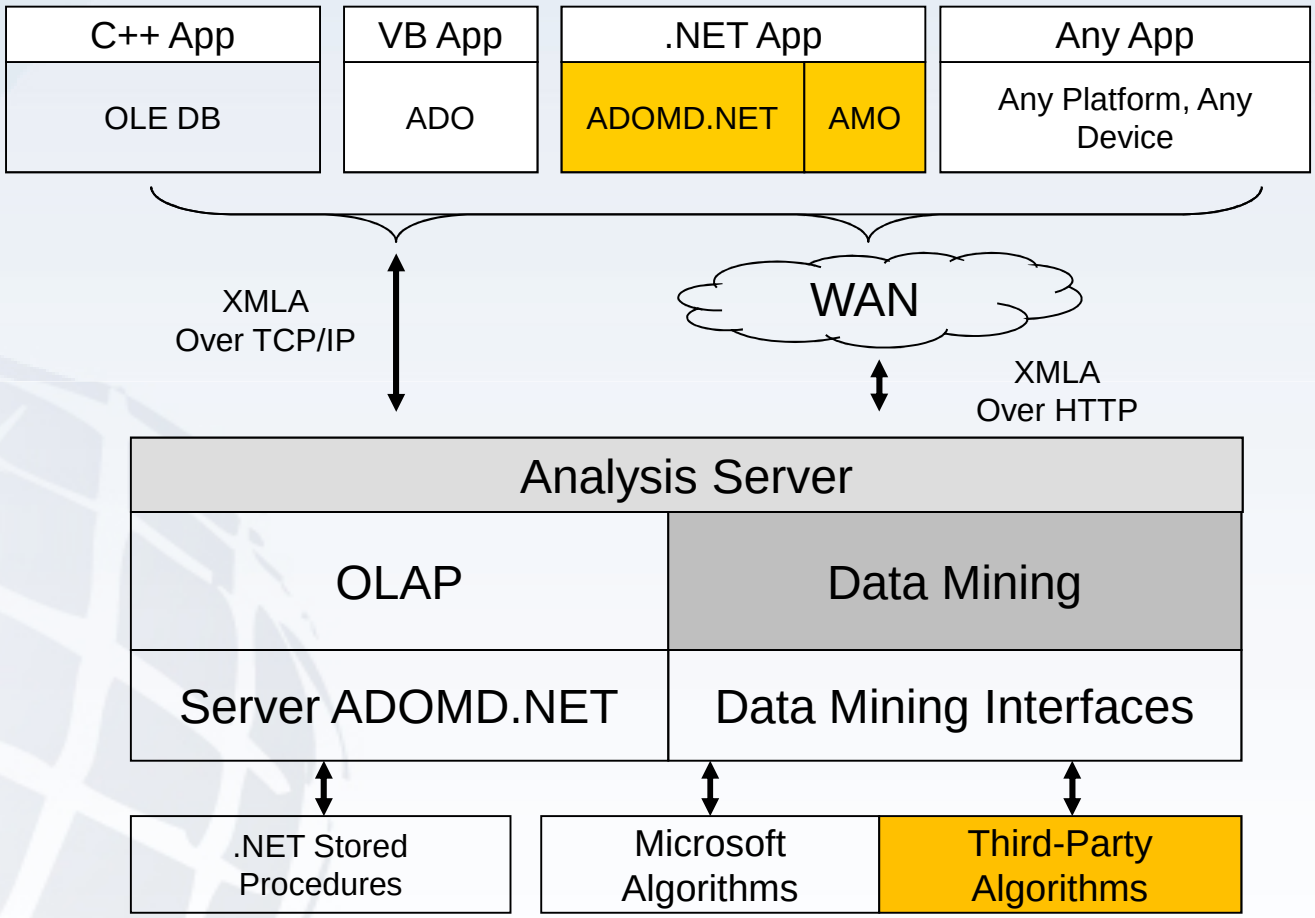
SQL SERVER™ 2008 ALGORITHMS

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

- 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

DEMONSTRATIONS

1. Creating, Training, Testing and Querying Mining Models with BIDS
2. Embedding Visualizations Into a Windows Forms Application
3. Embedding a Data Mining Report Into a Windows Forms Application
4. Enhancing an E-Commerce Site with Market Basket Analysis
5. Automating Data Validation With Data Mining

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
- Data Mining for SQL Server 2008, by ZhaoHui Tang and Jamie MacLennan