

您的潜力，我们的动力

Microsoft
微软(中国)有限公司

SQL2005 数据挖掘详解

Lesson 3

杨大川

CTO

北京迈思奇科技有限公司



MSDN Webcasts

讲师简介

- 杨大川 - 迈思奇科技有限公司CTO
 - 微软MVP.2004, 2005 (最有价值专家)
 - 曾任美国硅谷Annuncio公司首席工程师
 - 招商迪辰产品研发部总经理
 - 现兼任中科院客座教授
- Minesage :迈思奇科技有限公司
 - 微软数据分析/挖掘领域合作伙伴
 - 面向企业客户提供完整的数据分析与挖掘解决方案
 - 提供专业、高端的BI培训
 - www.minesage.com



Minesage

收听本次课程需具备的条件

- 本讲座难度属于高级
- 面向技术人员
- 建议收听过SQL2005 数据挖掘算法详解 Lesson 1（2005年12月21日）和SQL2005 数据挖掘算法详解 Lesson 2（2006年1月10日）

复习：前两讲的内容

- SQL2005数据挖掘概述
 - 贝叶斯 (Naive Bayes)
 - 决策树 (Decision Trees)
 - 神经网络 (Neural Networks)
 - 关联规则 (Association Rules)
 - 聚类分析 (Clustering)
 - 序列聚类分析 (Sequence Clustering)
 - 时序 (Time Series)
- 比较挖掘的准确度

什么是数据挖掘?

您的潜
Micro
微软(中国)



数据挖掘(**Data Mining**), 又称信息发掘 (**Knowledge Discovery**), 是用自动或半自动化的方法在数据中找到潜在的, 有价值的信息和规则.

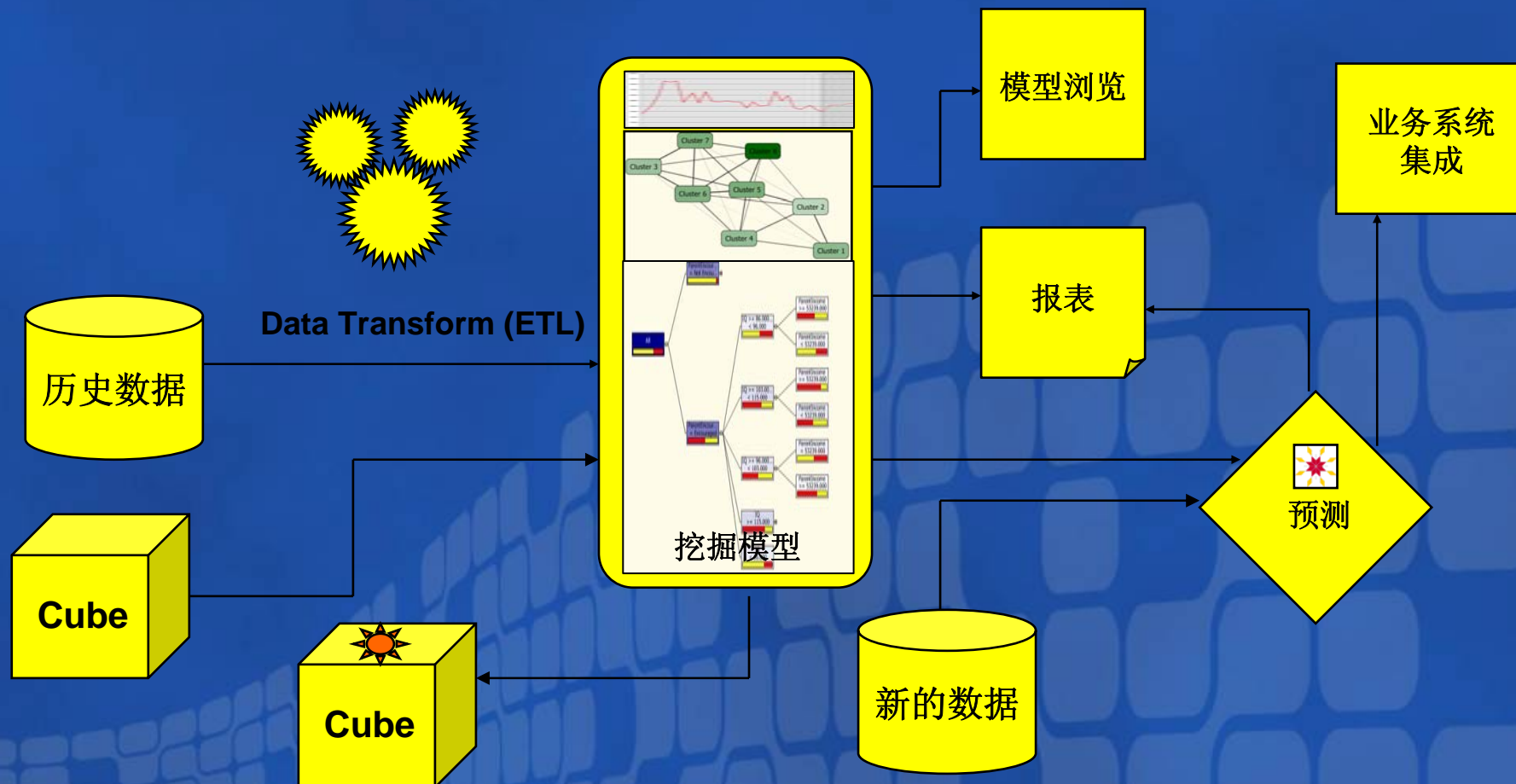
数据挖掘技术来源于数据库, 统计和人工智能.



您的潜力, 我们的动力

Microsoft
微软(中国)有限公司

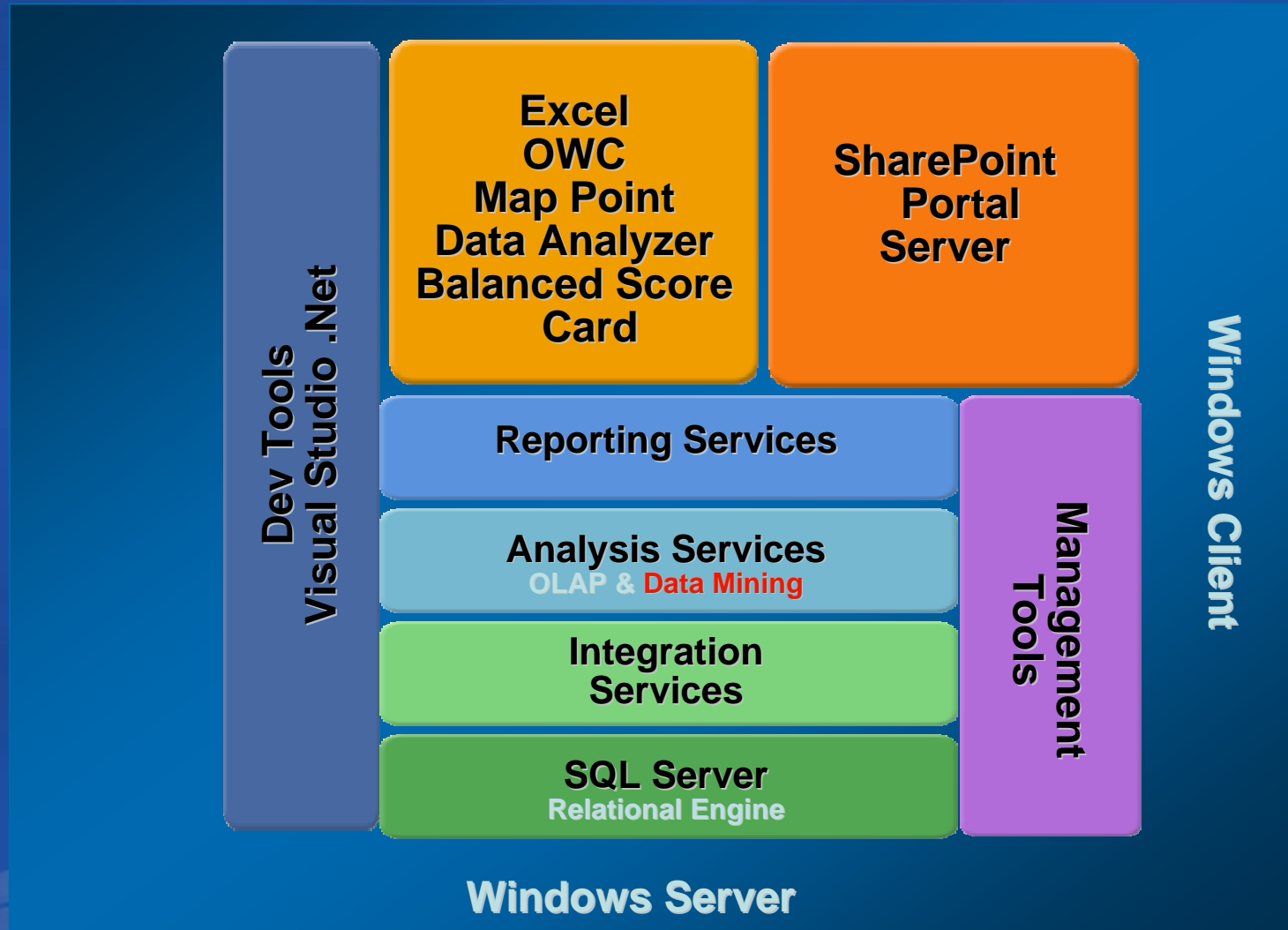
Data Mining Data Flow



数据挖掘 与 Microsoft 商务智能

您的潜力, 我们的动力

Microsoft
微软(中国)有限公司



API: DMX

CREATE MINING MODEL CreditRisk

(CustID LONG KEY,
Gender TEXT DISCRETE,
Income LONG CONTINUOUS,
Profession TEXT DISCRETE,
Risk TEXT DISCRETE PREDICT)

USING Microsoft_Decision_Trees

INSERT INTO CreditRisk

(CustId, Gender, Income, Profession,
Risk)

Select

CustomerID, Gender, Income,
Profession, Risk

From Customers

Select NewCustomers.CustomerID, CreditRisk.Risk,
PredictProbability(CreditRisk)

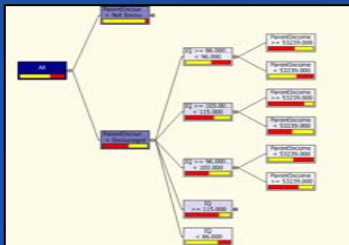
FROM CreditRisk **PREDICTION JOIN** NewCustomers

ON CreditRisk.Gender=NewCustomer.Gender

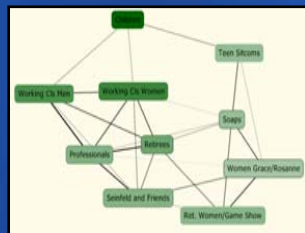
AND CreditRisk.Income=NewCustomer.Income

AND CreditRisk.Profession=NewCustomer.Profession

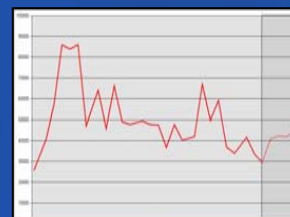
丰富的算法集合



决策树



聚类



时间序列

您的潜力. 我们的动力
Microsoft
 微软(中国)有限公司

Recommendation scores for Professionals/Technical and Service Workers			
Attributes	Values	Favor Professionals/Techn.	Favor Service Workers
Education/Year	15-20		
Education/Year	12-13		
Education/Year	7-12		
relation hnd(YOUNG AND THE RES.	Missing		
relation hnd(YOUNG AND THE RES.	Existing		
relation hndAG THE WORLD TURN.	Existing		
relation hndAG THE WORLD TURN.	Missing		

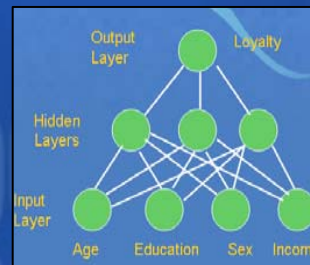
Naïve 贝叶斯



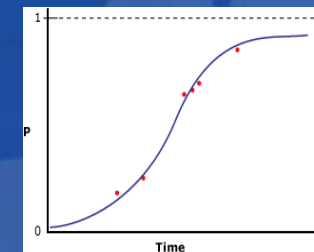
序列聚类



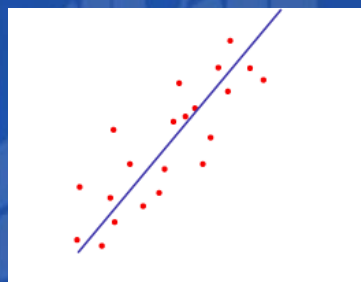
关联



神经网络



逻辑回归



线性回归

When upgrading to Microsoft® SQL Server™ 2000, you can upgrade servers in your org at a time; however, when servers are used for **Microsoft**, you must upgrade the Distrib. Publisher second, and then Subscribers. Upgrading servers one at a time following this is recommended when a large number of Publishers and Subscribers exist because you can **Microsoft** data even though servers are running different versions of SQL Server. You can publications and subscriptions with servers running instances of SQL Server 2000, and in subscriptions created in SQL Server 6.5 or SQL Server 7.0.

When using transactional **Microsoft**, you can upgrade Subscribers before the Publisher, using immediate updating with snapshot **Microsoft** or transactional **Microsoft**. There are upgrade recommendations in this topic under Upgrading and Immediate Updating.

You can upgrade **Microsoft** servers running SQL Server 6.5 or SQL Server 7.0 to SQL S the server is running SQL Server 6.5, you do not need to upgrade it to SQL Server 7.0 b upgrading to SQL Server 2000.

IMPORTANT When upgrading servers configured for **Microsoft** to SQL Server 2000, the compatibility level must be set to 70 (version 7.0 compatibility) or later. If you have a running in 65 (version 6.5) or an earlier compatibility level, temporarily change them during the upgrade process.

When the Publisher or Subscriber is running in 65 or an earlier compatibility level due to SQL Server 2000, error 15046 will be raised stating that the operation is supported Server version 7.0 or SQL Server 2000.

For more information about setting the backward compatibility level, see **SQL Server 2000: Books and Topics**.

If you are upgrading **Microsoft** on a failover cluster, you must uncluster the previous in before upgrading. Unclustering the previous installation means that you must delete all p remove **Microsoft**, and reconfigure it after upgrading to SQL Server 2000. This will not requirement when upgrading SQL Server 2000 to future releases.

文本挖掘

本讲座的主题

- 数据挖掘与SSIS的整合
 - SSIS
 - DMX
 - DDL
 - XMLA
- 数据挖掘二次开发
 - ADOMD.NET
 - AMO
 - ASP.NET

实际案例：找出潜在客户

- 已知条件
 - 现有的几万名客户信息
 - 哪些是“好的”，哪些是“不好的”
- 目标
 - 针对若干新的客户
 - 分出类别
 - 好的
 - 可能是好的
 - 坏的
 - 可能是坏的

设计并且训练挖掘模型

- 可以选择的模型包括
 - 贝叶斯 (Naive Bayes)
 - 决策树 (Decision Trees)
 - 神经网络 (Neural Networks)
 - 聚类分析(Clustering)
- 这个案例中, 我们推荐决策树

您的潜力. 我们的动力

Microsoft[®]
微软(中国)有限公司

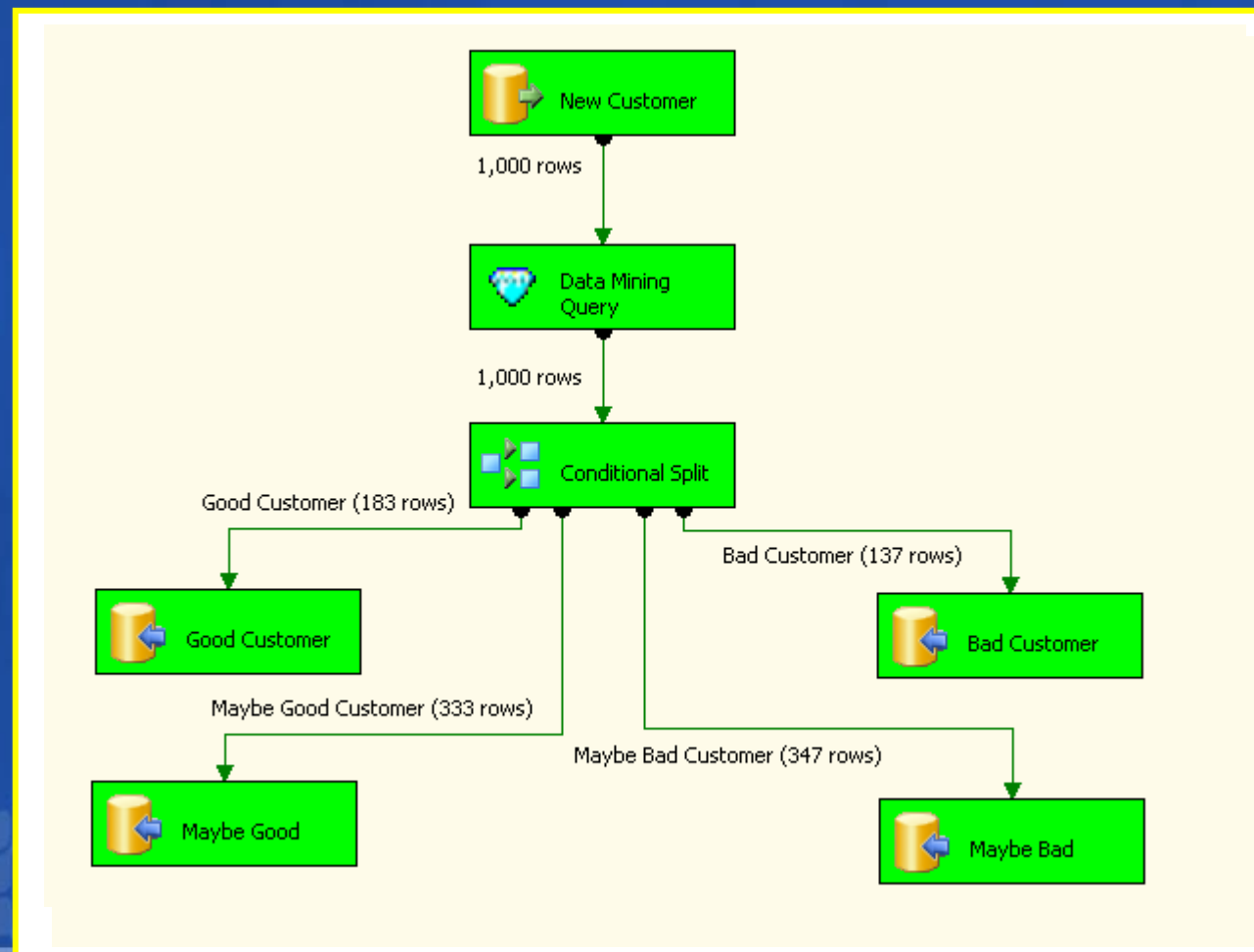
Demo 1

- 创建模型
- 训练模型

预测的DMX

```
SELECT FLATTENED
    t.[CustomerKey],
    [TM Decision Tree].[Bike Buyer],
    (PredictProbability([TM Decision Tree].[Bike Buyer])) as [Prob]
From [TM Decision Tree] PREDICTION JOIN
    @InputRowset AS t
ON [TM Decision Tree].[Marital Status] = t.[MaritalStatus] AND
    [TM Decision Tree].[Gender] = t.[Gender] AND
    [TM Decision Tree].[Yearly Income] = t.[YearlyIncome] AND
    [TM Decision Tree].[Total Children] = t.[TotalChildren] AND
    [TM Decision Tree].[Number Children At Home] = t.[NumberChildrenAtHome] AND
    [TM Decision Tree].[House Owner Flag] = t.[HouseOwnerFlag] AND
    [TM Decision Tree].[Number Cars Owned] = t.[NumberCarsOwned] AND
    [TM Decision Tree].[Commute Distance] = t.[CommuteDistance] AND
    [TM Decision Tree].[Region] = t.[Region] AND [TM Decision Tree].[Age] = t.[Age]
```

搭建全部的SSIS流程



您的潜力. 我们的动力

Microsoft[®]
微软(中国)有限公司

Demo 2

- 自动进行客户分类

您的潜力. 我们的动力

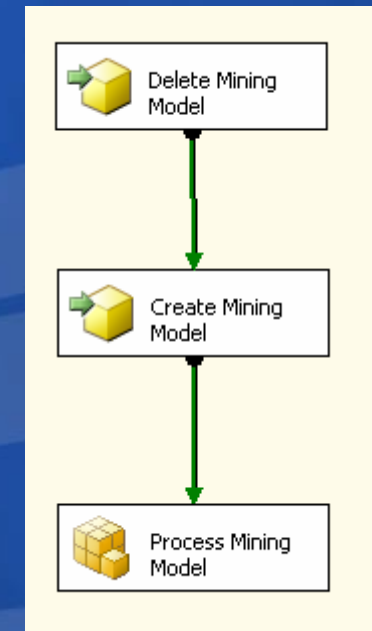
Microsoft
微软(中国)有限公司

如何自动建立、训练模型

- 什么是DDL
- DDL 与 XMLA

Demo 3

- 使用 DDL 创建自动流程



您的潜力. 我们的动力

Microsoft
微软(中国)有限公司

创建模型的几种方法

- DDL
- DMX
- AMO

您的潜力. 我们的动力

Microsoft
微软(中国)有限公司

数据挖掘的二次开发

- 背景知识介绍
 - ADO
 - ADO.NET
 - ADOMD
 - **ADOMD.NET**
 - DSO
 - **AMO**

案例：电影销售

- 某连锁店
 - 会员制
 - 销售电影DVD
- 目标
 - 搭建一个网站, 根据用户的购物车, 选择推荐产品
- 挖掘模型
 - 关联规则 (推荐)
 - 决策树

您的潜力. 我们的动力

Microsoft
微软(中国)有限公司

使用AMO 创建模型

- 创建模型
 - MiningStructure
 - MiningModel
- 处理（训练）模型

Demo 4

- 使用 AMO 创建并且处理模型

Form1

Create Model Process Model

最大项集 0

最小支持度 0.01

最小概率 0.3

预测的MDX

SELECT

Predict ([Asso Mining].[Movies],
\$AdjustedProbability, 5)

From

[Asso Mining]

NATURAL PREDICTION JOIN

(SELECT (SELECT 'Men in Black' AS [Movie]) AS
[Movies]) AS input

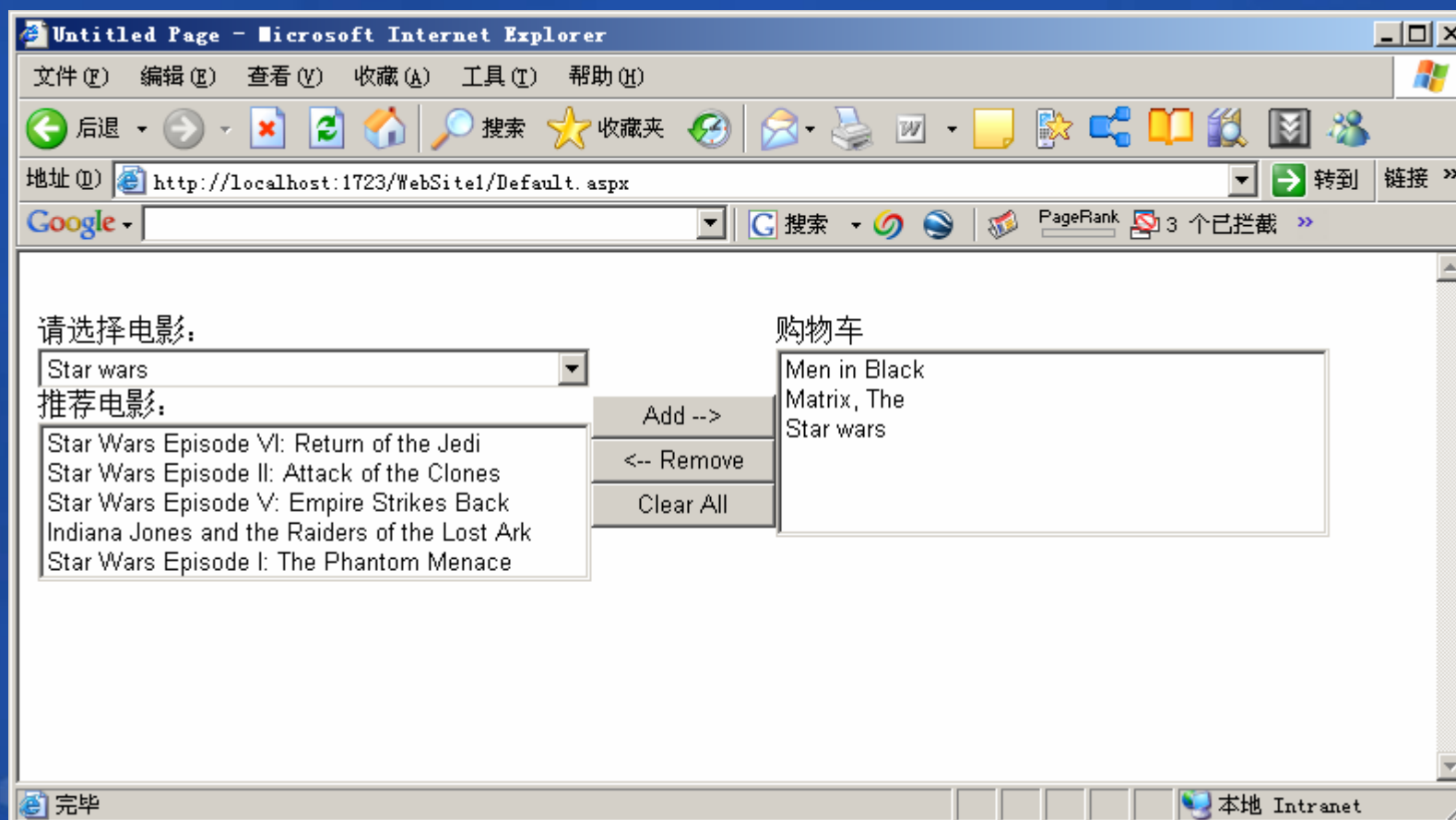
使用ADOMD.NET进行预测

```
AdomdConnection conn = new AdomdConnection("Data
Source = localhost\\sql2005; Catalog = MovieSample;
Integrated Security = SSPI");
conn.Open();
AdomdCommand cmd = conn.CreateCommand();
cmd.CommandText = generateDMX();
AdomdDataReader dr = cmd.ExecuteReader();
while (dr.Read())
{
    suggestListBox.Items.Add(dr.GetString(0));
}
conn.Close();
```

您的潜力. 我们的动力

Microsoft
微软(中国)有限公司

使用 ASP.NET 搭建应用程序



Demo 5

您的潜力. 我们的动力

Microsoft
微软(中国)有限公司

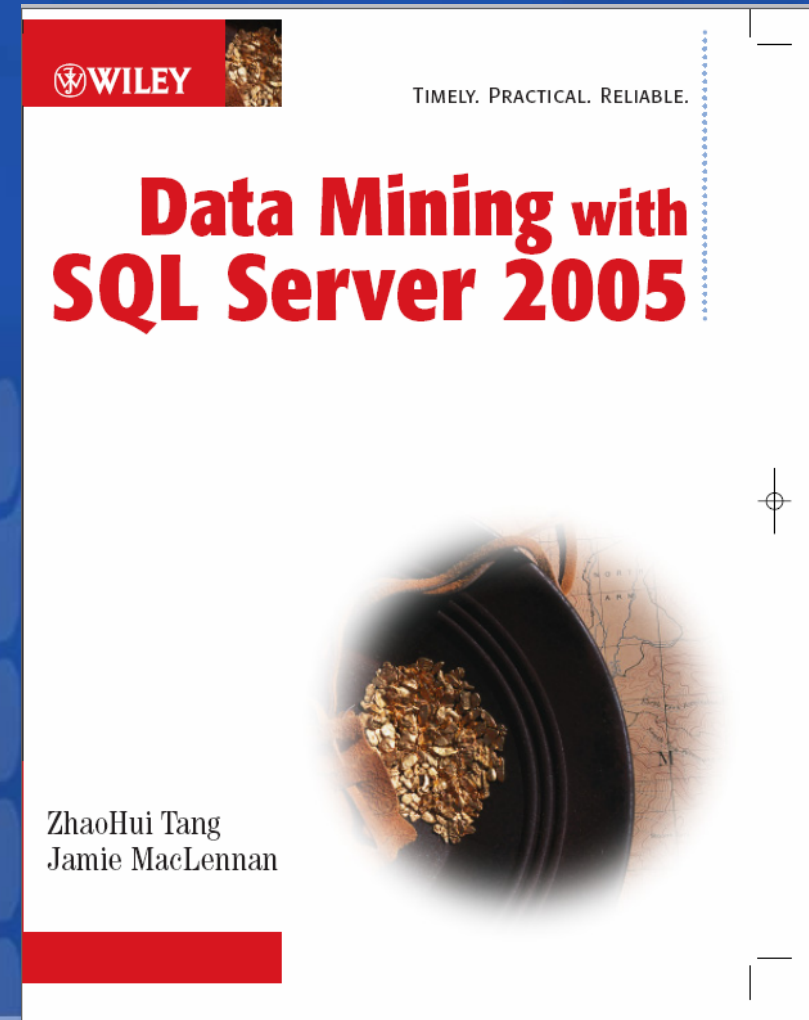
- 智能促销网站的演示

更多信息...

您的潜力. 我们的动力

Microsoft
微软(中国)有限公司

- DM URL
 - www.SqlServerDataMining.com
- 新闻组:
 - Microsoft.Public.SQLServer.DataMining

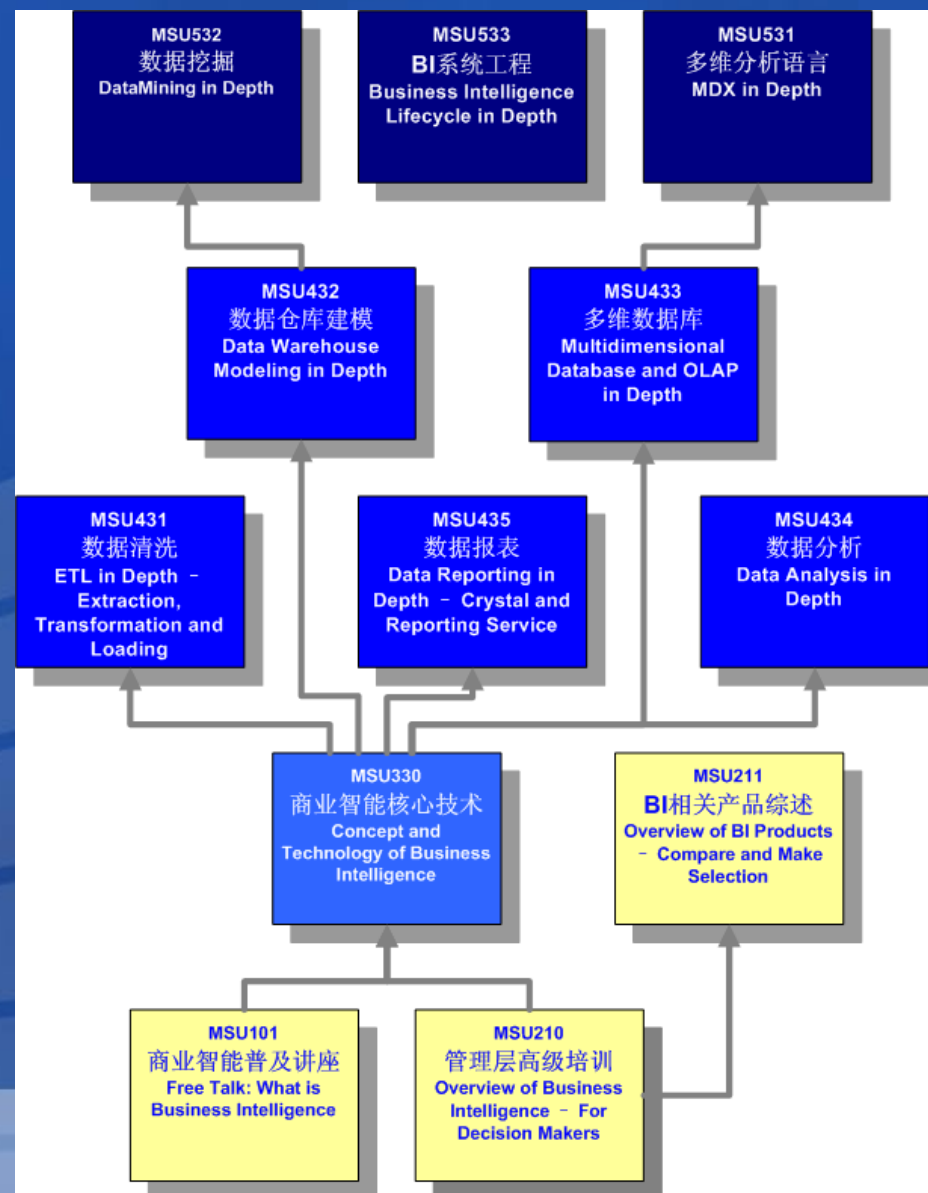


进一步的培训机会

- 迈思奇：专业的BI培训机构，全国各地长期开设各类培训
- www.minesage.com
- 最近一期培训时间：
– 2006/3/25 – 3/27


您的潜力，我们的动力

Microsoft®
微软(中国)有限公司





Question & Answer

如需提出问题，请单击“提问”按钮并在随后显示的浮动面板中输入问题内容。一旦完成问题输入后，请单击“提问”按钮。

 **问题和解答 (无问题)** ▲ ×

在此会议中尚未解答任何问题。

要向演示者提问，请在此处键入问

提问(A)

删除(D)

问题管理器(Q)

您的潜力，我们的动力

Microsoft®
微软(中国)有限公司

Microsoft®

msdn


MSDN Webcasts