The Washington Dept. of Revenue Data Mining Pilot Pilot Project:

A Retrospective Overview

2000 FTA Revenue Estimating and
Tax Research Conference
September 26, 2000

Stan Woodwell
Research Information Manager
Washington Dept. of Revenue

Data Warehousing/ Data Mining Study Team

Three Integrated Efforts

- building small data warehouse -- “Data Mart”
- testing query & analysis tools
- doing data mining pilot project

Data warehousing

A data warehouse is a copy of transaction data specifically structured for querying, analysis and reporting.

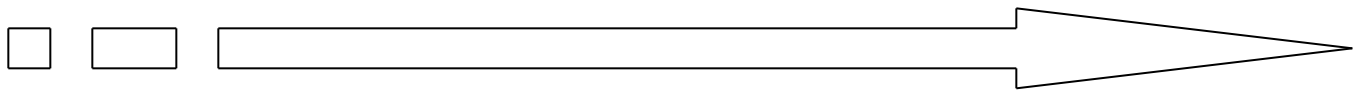
That is,

- on physically separate hardware
- organized differently, especially for querying, analysis & reporting

Data Mart/Query Software

- **Data Mart**
  - SQL Server - 50 Gig
  - NT Operating System
  - ODBC
- **Query Software -- COGNOS**

Data Mining Continuum



Query

Statistical Procedures

Decision Trees

Neural Networks

Data Mining
What's New, What's Not

- Neural Networks
- Decision Trees (Rule Induction)
  - represent
- "Artificial Intelligence"
- Data trains software
- Query Logic
- Statistical Procedu
  - Regression
  - Cluster Analysis
  - Association Rules (Affinity/Market Basket Analysis)

Driving What's New...

- **Incredible Increases in Computer Speed and Memory**
- **Software Utilizing Extremely Complex Iterative Processes Can Really Crank**

Dogbert the Consultant

“If you mine the data hard enough you can also find messages from God”

EXPECTATIONS

- Top Management
- Conferences
- Vendor Presentations

#1 DOR Priority

Committee Decisions
Data Mining

- **Selection of Pilot Project**
  "Proof of Concept"
- **Selection of Data Mining Software**
  for Pilot Project

Data Mining Software Selection

- NCR
- SAS
- SPSS
- IBM

- SPSS Clementine Miner
  - “In a cavern, in a canyon, excavating for a…”

Criteria for Pilot Project

- Doable
- Measurable
- Produces Efficiency withi
  Program
- Within Budget
- Divisional Resources
  Available
- Can be completed by End o
  June

Projects Considered for Pilot

- Enhancing Audit Selection
- More Sophisticated Audit Retail Profiling
- Expanded Active Non-Reporter Profiling
- Tax Discovery - Identifying Non-Filers
- Parallel Taxpayer Education Effort
- Examining Transactions for Fraud
- Controlled Experiment with Collections

Data Mining Pilot Project Audit Selection

Purpose

- Provide “Proof of Concept” for Advanced Dat Mining
- Demonstrate Enhanced Predictive Capabilit through Utilization of Sophisticated Softwar
- Lead to Development of More Productive Au Selection Criteria

Data Mining Pilot Project
Audit Selection

- **Design**
  - **“Quasi-Experimental”**
  - **Utilizes ODBC from Data Mart**
  - **Dependent Variable Audit Recovery**
  - **Build “Supervised” Model Using Known Resu from Audits Issued in 1997**
  - **Use Model to Predict Recovery for 1998 Aud**
  - **Compare Predictions with Actual 1998 Resu**

Data Mining Pilot Project Audit Selection

- **Process**
  - Divided Audit Recovery into 4 Bands
    - $1 - 1,000
    - $1,000 - 5,000
    - $5,000 - 10,000
    - Over $10,000
  - Divided 1997 Audit Sample into 2 Samples -- "Test" Sample and " Training Sample"
  - Built Models using Training sample, applied to Tes sample to test generalizability
  - Applied Best Models to Predict Recovery for 1998 Audits

SPSS Clementine Rule Set Example
Rule Induction modeling

```
Rules for 2:
     Rule #1 for 2:
          if  Tax_Due_Amount_1 <= 38618.2
          and  Taxable_Amount_2 > 521297.0
          and  Gross_Amount_3 <= 1683394
          and  Tax_Due_Amount_3 > 4506.39
          and  Total_Wages_Amount_3 <= 392286
          and  Average_Employee_Count_4 > 3
          and  taxlag1 > -3950.45
          and  dedlag1 > 3694.86
          and  dedlag3 > -17693.7
          and  emplag2 > -5
          and  emplag2 <= 1
          then -> 2

     Rule #2 for 2:
          if  Line_Code_Num_6 == T
          and  Taxable_Amount_3 > 457921.0
          and  Taxable_Amount_3 <= 9191570
          and  Total_Wages_Amount_1 <= 489394
          then -> 2

     Rule #3 for 2:
```

Data Used for Modeling

Data NOT Used for Modeling

Washington Combined Excise Tax Return

- Sales Tax -- 1 line code, 1 rate
- Business and Occupation Tax and
- Public Utility Tax -- 22 line codes, different activities, different rates
- 27 Deduction Codes associated with line codes

Unable to Use

- line code amounts
- deduction type amounts
- deduction type by line amounts

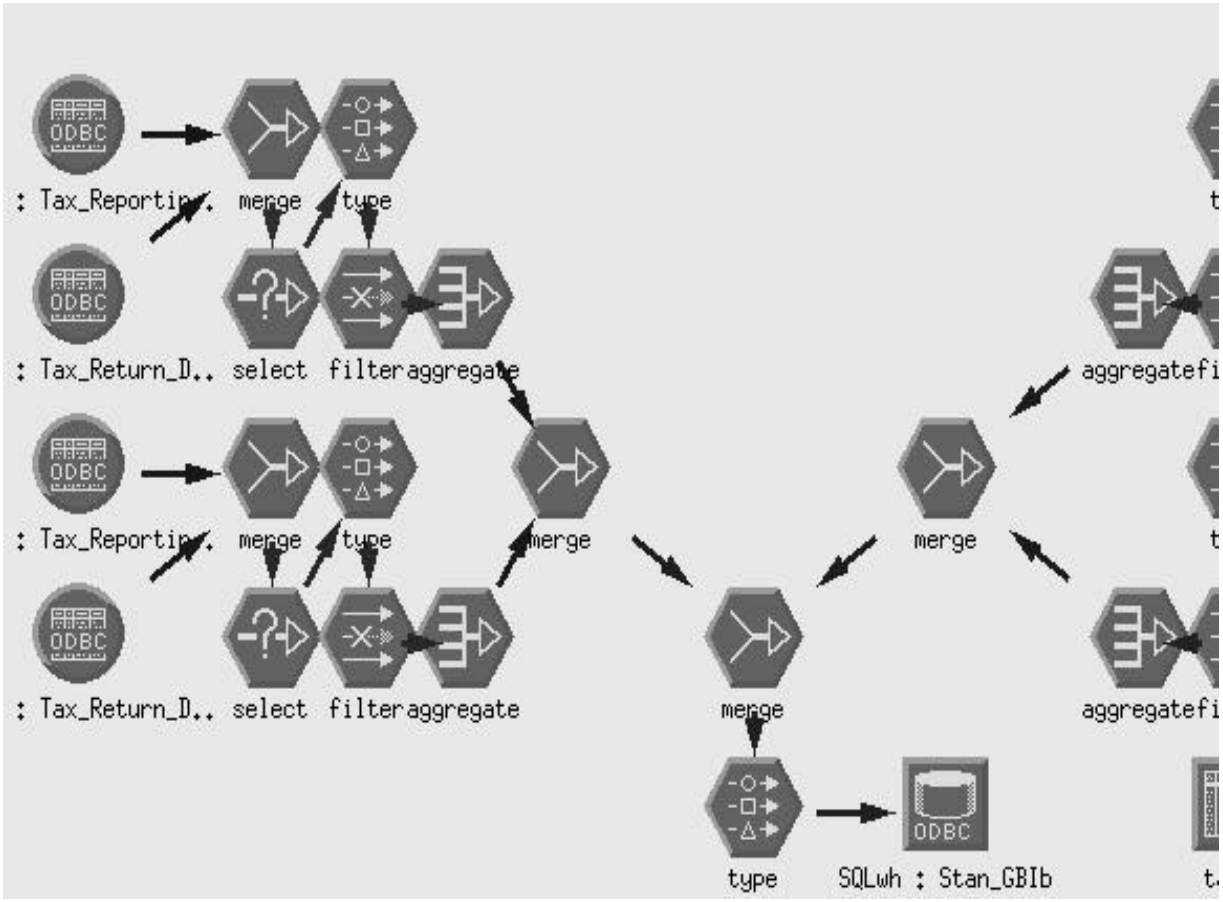
Major Problems

- Data Structures
- Missing/Imperfect Data
- Modeling Overspecification

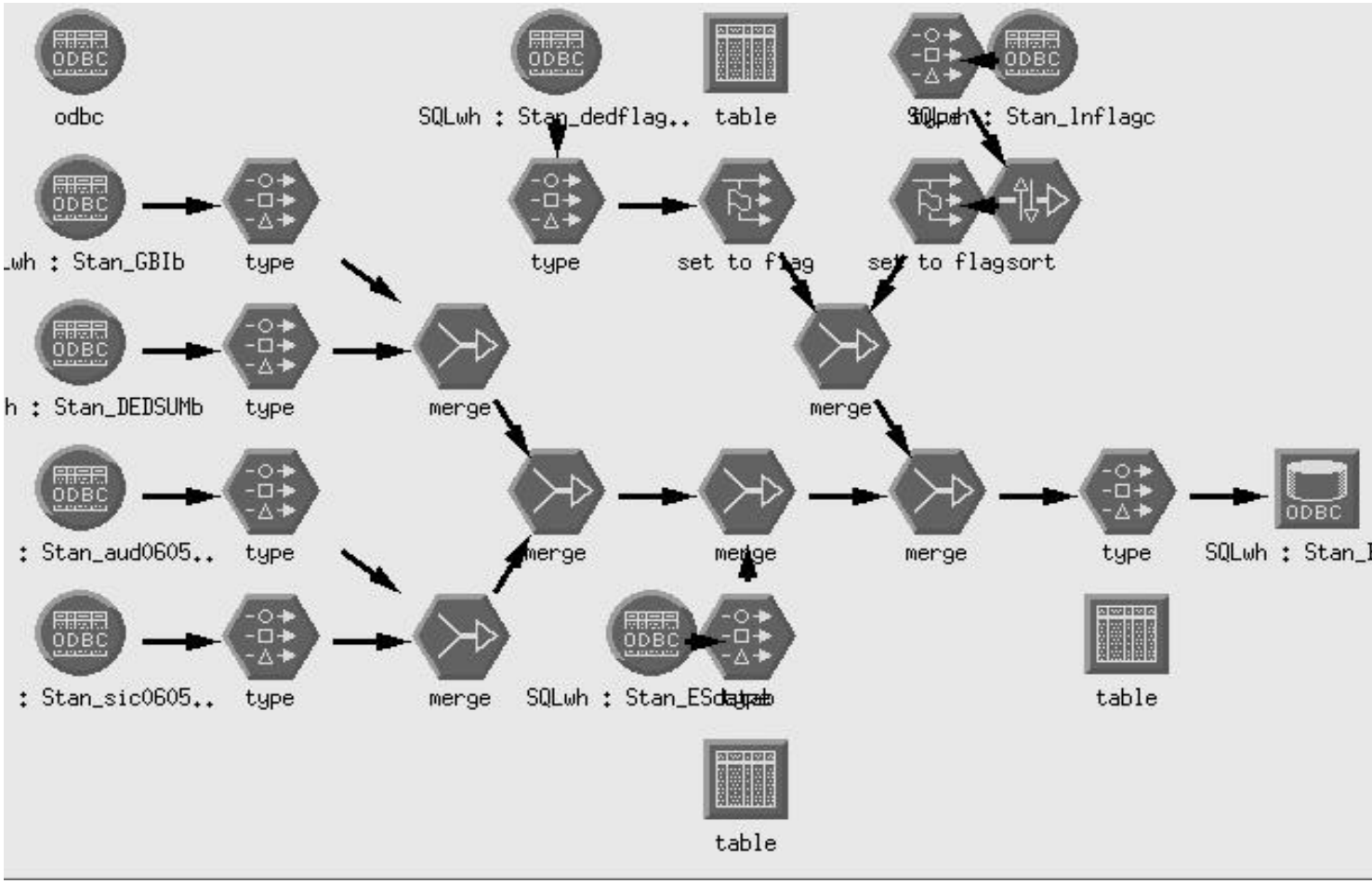
Data Structures

- Query Software
  - “Star Structures”
  - relational data base/hub tables
  - myriad tables connected by multiple keys
- Mining Software
  - “flat file”
  - single record containing everything for each taxpayer

SPSS Clementine Merge Stream



SPSS Clementine Merge Stream



Overspecification

- Model “too close to data”
- No problem generating rules to “predict” training sample with extreme accuracy
- Model’s predictive rules did not generaliz particularly well to test sample

Results--Predicting 1998 Audit Recovery Band

“Results Positive but Modest…”

Conclusions/Lessons Learned

- Due to a number of limiting factors, the predictive power of the pilot model was positive but modest

- As a learning experience the pilot was an unquestioned success. A great deal of technical knowledge was acquired within the Department i a very short period of time. Some of the major lessons learned are as follows:

Conclusions/Lessons Learned

- Optimal data structures for query software are definitely not optimal for mining software—a "two-tiered" approach to data warehousing will frequently be necessary.

- The major part of data mining (possibly 85 to 95%) is data preparation and data cleansing.

- Optimal use of mining software requires "perfect" data, structured with fillers for missing records and missing fields.

Conclusions/Lessons Learned

- Despite the power of the modeling software, modeling is still a complicated process of structural design, analysis and experimentation.

- While training is essential and limited use of outside consultants may be beneficial, the Department does have the technical capacity to d data mining in-house.

Conclusions/Lessons Learned

- Data Mining is not a "magic bullet." It requires a highly focused and structured approach. It is highly technical and resource intensive.

- For appropriate applications, sophisticated Data Mining could be an extremely valuable and cost effective strategy for the Department .

Into the Realm of Budget Process.

- $$$
- FTE's
- Internal Politics
  - Mining vs. Querying
- External Politics (Gov's Office, Legislature
  - Government Intrusivenss
  - Politically Correct Terms
- ??????????