MINNESOTA · REVENUE

The Minnesota DOR's Experience:
Using Data Warehousing Techniques to Increase Tax Compliance

CSC

MINNESOTA · REVENUE

Agenda

- Background
- Data Cleanse and Match
- Data Mining
- Lessons Learned
- Q & A

CSC

2

MINNESOTA · REVENUE

Minnesota DOR

- Serve population of about 5.2 million
- 1287 employees
  - Approximately 2/3 in direct compliance & collections functions
- Filers
  - 150,000 active Sales Tax filers
  - 164,000 active Withholding filers
  - 50,000 active Corporate filers
  - 3,600,0000 Income Tax filers
- $17.5 billion in tax collections
  - No motor vehicle, very small property tax
- Collections arm brought in $170 million in delinquent taxes & $21.6 million in non-tax debt
  - No Child Support

CSC

3

MINNESOTA · REVENUE

Minnesota DOR Strategic Plan

Vision

- Everyone pays the right amount of tax
- Information is timely, accurate and convenient
- Employees have necessary skills, tools, and resources
- Revenue system works well, in policy and operation

Mission

"Make the revenue system work well"

Strategies

- Focus compliance efforts on those who deliberately evade the tax laws, not on those who make an effort to "get it right"
- Measure the effectiveness and cost of activities, and shift resources to those that demonstrate the greatest success in achieving our mission.

CSC

4

MINNESOTA · REVENUE

MN DOR Warehousing

- History
  - 1994 – Sales Tax Reengineering
  - 1996 – Collections System Warehouse
  - 2003 – Income Tax Reengineering
- Problem
  - 3 Database platforms
  - No Integration
  - No Data Warehouse Strategy
  - No formal organization inside DOR
  - Spotty expertise
- Solution
  - Develop and Implement a Roadmap for an Integrated Data Warehouse

CSC

5

MINNESOTA · REVENUE

Roadmap Elements

- Create Data Warehouse Steering Team
- Staff full-time Data Warehouse positions
- Form Data Warehouse Coordination Team
- Acquire Extract, Transform and Load Tool
- Utilize Data Quality & Match techniques and technologies
- Research Business Intelligence tools
- Migrate to One Integrated Warehouse
- Add new data sources
- Pilot Data Mining projects using University of Minnesota graduate students
- Measure value of Data Warehouse sources

CSC

6

MINNESOTA · REVENUE

What we use our warehouse for

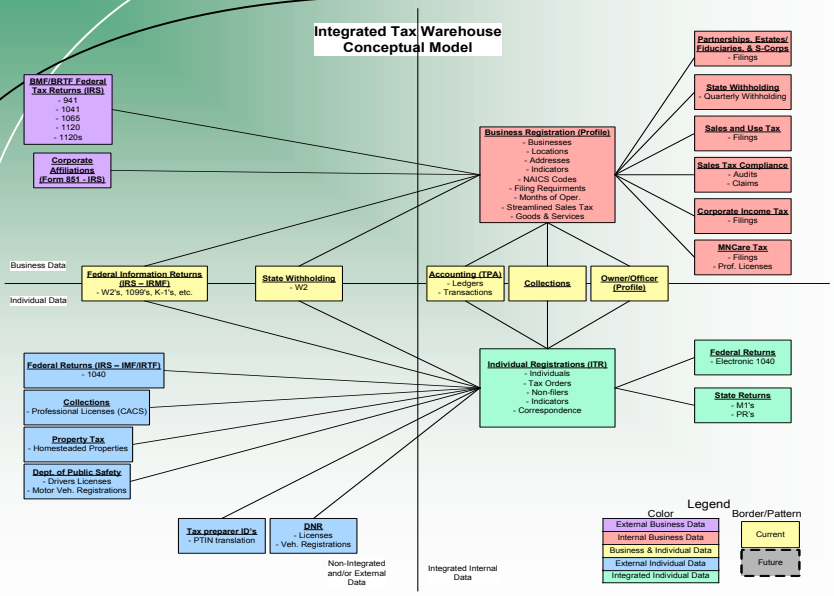
- Ad Hoc Queries
  - Compliance
  - Collections
  - Research
- Reporting
- Data Mining

- "Show me the money"
  - We have committed to collect an additional $300+ million each biennium over the last 5 legislative sessions

CSC

7

MINNESOTA · REVENUE



CSC

8

MINNESOTA · REVENUE

Data Quality and Matching Problems

- To find new audit cases, new sources of data, usually external, need to be used
- We usually have little control over the quality of the data we receive from these external sources
- Have to ensure that one taxpayers information in existing data sources gets linked accurately to that taxpayers information in the external data source
- SSN/FEIN – The traditional matching criteria
  - Fewer and fewer external sources have SSN/FEIN
  - Have to allow for missing or incorrect SSN/FEIN
- Given the volume of data, human intervention and manual data correction is not possible
- How have we overcome these obstacles?

CSC

9

MINNESOTA · REVENUE

Use of Data Quality & Match Techniques

- Using Data Cleansing and Matching Software
- What does this software do?
  - Parsing - breaks names into individual components
  - Standardization - Uses dictionaries present in software to accomplish this
    - Out of the box
      - Business Names
      - Individual Names (first, mid, last)
      - Address
    - Customizable
      - Modify out of the box dictionaries
      - Create new dictionaries from scratch
  - Advanced algorithms for matching
    - The software will indicate “John Doe Enterprise” in one source and “J D Enterprise” in another source are very similar

CSC

10

MINNESOTA · REVENUE

Individual Income Tax

- Match Driver's Licenses Data to Individual Registration System
  - Used to establish residency
  - Used to determine value of vehicles registered
- Calculated match scores using four match attributes
  - Had to meet threshold for at least three of the matching attributes
  - Published scores to warehouse for users to view

Matching Attribute	Scoring Algorithm	Notes
Full Name	Bigram	• Parsed and standardized first • Scored with and w/o last name
Address	Exact Match	• Parsed and standardized first
SSN	Hamming	
Date of Birth	Hamming	

CSC

11

MINNESOTA · REVENUE

Sales Tax

- Match Schedule C's to Business Registration System
  - Looking for non-filers and under-filers
  - Taxpayers are not consistent in reporting demographic information
- Calculated match scores for each of six match components
  - Built weighted rules to generate an overall match score between 0 and 100%.
  - Published six match scores & overall scores for each match to warehouse for users
  - Allowed for one Schedule C to match to multiple Business Registrations & one Business Registration to match to multiple Schedule C's
    - Identified the "best" match (the one with the highest score)

CSC

12

MINNESOTA · REVENUE

Sales Tax (con't)

Matching Attribute	Scoring Algorithm	Notes
Business Name	Bigram	• Used dictionaries to remove "noise" words (i.e. Corp, Inc, etc.)
Address	Exact Match	• Standardized first • Used both Schedule C & 1040
FEIN	Hamming	
Industry Classification	Exact Match	• Used only first two digits
# of Active Owners	Exact Match	• If one active owner, then match
Organization Type	Exact Match	• If registered as Sole Prop, then match

CSC

13

MINNESOTA · REVENUE

Use of Data Quality & Match Techniques

- Lessons Learned
  - The Good
    - Software can identify one taxpayers information in multiple sources, even when information is slightly different
    - Matches can be made with much more accuracy than basic programming techniques can provide
    - Process can be tuned to require little to no human intervention
    - Able to match and use data that previously had to be discarded
    - Business staff that are familiar with data speed process
  - The Not So Good
    - It cannot eliminate all mistakes in matching. Incorrect matches will still occur
    - Marginal matches may need to discarded if human intervention is not possible or not cost effective
    - It needs to be configured and set up, which may require individuals with specific skills

CSC

14

MINNESOTA · REVENUE

Pilot Data Mining

- Worked with Professor Jaideep Srivastava of the University of Minnesota
  - A well known expert in the field of data mining
- Over the course of 14 months we had the opportunity to do data mining pilot projects with the help of 5 University of Minnesota PhD candidates working under the professors tutelage.
- Used the data mining techniques for :
  - Individual Income tax,
  - Sales and Use tax,
  - Corporate tax and the Partnership/Estate/Fiduciary/S-Corp area taxes.

CSC

15

MINNESOTA · REVENUE

Sales & Use Pilot

- Goal
  - Find productive smaller cases for entry-level auditors and to reduce the number of no change audits.
  - We defined a productive audit as an audit resulting in an assessment of at least $500 per year, or $1500 per case.
- Outcome
  - Four different audits models produced, one each for medium and small use tax audits and one each for medium & small use tax audits
  - Technique Used - Naïve Bayes with Multi-boosting
    - Produced a score between 0 and 1 for every potential audit candidate

CSC

16

MINNESOTA · REVENUE

Sales & Use Pilot (con't)

- Lessons Learned
  - Mining may first uncover what is already known
    - Had to remove Fortune 500 and Top 200 per audit region
  - May need multiple models to best capture taxpayer behavior
    - Had to create separate medium and small business models for both sales & use tax
  - Don't count anything out: we used every element from other business returns that we could, including Corp, S-Corp, Partnership, ….
  - Don't be discouraged at preliminary results; more refinement will eventually get you where you want to be.
  - Our goal was to find productive smaller cases for entry-level auditors and to reduce our number of no change audits. We defined a productive audit as an audit resulting in an assessment of at least $500 per year, or $1500 per case.
    - Data mining identified a large number of small businesses we would otherwise have chosen only by chance.
  - Predicting an amount of likely payment is possible, but a much more difficult problem

CSC

17

MINNESOTA · REVENUE

Income Tax Schedule C Loss Project

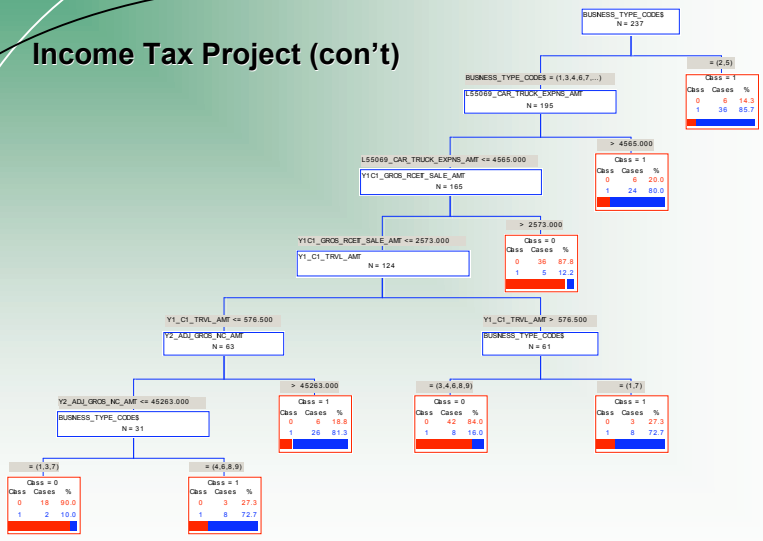
- Goal
  - Attempt to identify taxpayers who are incorrectly using a Schedule C (i.e. a business) to reduce their taxable income by participating in activities that are not consistent with a business that is trying to produce revenue.
    - Also known as the "hobby-loss" project.
  - Determine if data mining could reduce the number of unsuccessful audits
- Outcome
  - Built model using CART decision tree algorithm
- Lessons Learned
  - Getting buy-in from non-IT staff can be challenging.
    - Ran model against audit list created under old process
  - Defining a 'successful' audit is important, but can be difficult.
    - Defined success for this project as a audit that generated at least a $1,000 assessment.

CSC

18

MINNESOTA · REVENUE

Income Tax Project (con't)



19

MINNESOTA · REVENUE

C Corp and S Corp Income and Expense Audit Pilot

- **Goal**
  - Identify new audit leads for C-Corp and S-Corp businesses.
- **Outcome**
  - Data mining in our experience did not work out as we had planned.
- **Lessons Learned**
  - Need historical results - Income and Expense audits for C Corps have only been done for about a year. S Corps have been doing them longer, but they often have other issues that may cause problems with the coding of the audit. Because there are often more than one issue, the amount of the assessment would not be a true indicator of the amount assessed for the income and expense audit. Without this historical data it was difficult to get useful results.
  - Another problem we ran into was the number of audits done are so small in comparison to the number of returns, it is difficult to find patterns.
  - Another issue we had for C Corps, not all the fields are captured, so it limits what information we can mine.
  - Currently we are working to keep better records in one central location so in the future we would have some historical data to use.

20

MINNESOTA · REVENUE

Data Mining Skills – Lessons Learned

Methodology Step	Difficulty Acquiring Skill	Skill Required
Business Understanding	Medium	DOR needs to be able to better discern what tax compliance problems can be addressed by data mining and which ones cannot.
Data Understanding	Low	Basic statistical analysis (i.e. correlations, data profiling, sampling, etc.) of data to identify problems or basic relationships that will affect the later steps.
Data Preparation	Medium	Certain data mining techniques require data normalization techniques that use statistical procedures to modify data prior to the modeling phase.
Modeling	High	For a single data mining software package the tool provides 10 or more different data mining modeling algorithms, each algorithm requires the tuning of 10 or more parameters. Data mining techniques are evolving constantly and keeping up with changes will require additional time investment. Some modeling packages require programming skills and do not have graphical user interfaces to simplify the modeling process.
Evaluation	High	The ability to evaluate the statistical results produced by any of the modeling tool. Even more important and critical is having the knowledge and experience to know what to do next when any given model does not generate useful results.
Deployment	Low	If a given model is going to be run on a regular basis (every second to once a week), then the models need to be tied into an operational process.

CSC

21

MINNESOTA · REVENUE

Results Pilot Data Mining projects using University of Minnesota students

- We are still testing the models we developed to see how they perform compared to our old methods of audit selection.
  - We still need to complete a number of audits based on the data mining findings.
  - Not actively working with the U of M at this time.
- Some models turned out to be better at telling us who not to audit rather point out good audit candidates.
- There appears to be promise in the use of data mining for tax compliance, but there are still many unanswered questions on where and when this tool can be cost effective and sustainable.

CSC

22

MINNESOTA · REVENUE

Sales & Use Pilot Results

Results in % for All Categories	Pre Data Mining Avg Success Rate	Data Mining Predicted Success Rate	Actual Success Rate
Sales	29%	38%	37%
Use	39%	56%	51%

CSC

23

MINNESOTA · REVENUE

Sales & Use Pilot Results

Results in $ for All Categories	Pre Data Mining Avg Dollars	Data Mining Predicted Dollars	Actual Dollars
Sales	$6,497	$11,976	$8,186
Use	$5,019	$8,623	$10,829

CSC

24

MINNESOTA · REVENUE

Sales & Use Pilot Results

Results of 414 Audits (includes P&I)	Overall Total Assessed	Overall Average Assessed
Large Sales & Use	$1,399,436	$19,437
Small Sales & Use	$72,605	$2,504
Large Sales	$6,229,248	$23,776
Small Sales	$101,895	$1,998
Combined Totals	$7,803,184	$18,848

CSC

25

MINNESOTA · REVENUE

Schedule C Loss (Hobby) Pilot

255 Schedule C Loss Audit Results	Actual W/O Data Mining	Actual with Data Mining
Success Rate	76%	83%
Dollars Assessed	$3,606	$3,917

CSC

26

MINNESOTA · REVENUE

Lessons Learned – U of M Pilot

- Pilot Data Mining projects using University of Minnesota students
  - The Good
    - Professor Jaideep Srivastava
    - Very bright, inquisitive people
    - Exposure to cutting edge tools and techniques
    - The Cost – but its wasn't cheap
  - The Not So Good
    - Some students better suited than others
    - They are students, not professionals
    - Difficult to retain students for more than 6 months
    - It takes time to perform the audits suggested by data mining – students gone by the time results are in

CSC

27

MINNESOTA · REVENUE

Lessons Learned - General

- There are many "best" ways of using your data warehouse
  - Which one is best for you?
- Need buy in and participation from agency leadership
- Need individuals with depth in tax knowledge to partner with technology staff to be successful
  - Some of our best auditors are retiring we need to tap into this knowledge and encapsulate it inside of applications and systems now before it leaves
- The skills needed to pursue these goals are many and varied, acquiring and maintaining these skills while managing costs was and continues to be a challenge

CSC

28

MINNESOTA · REVENUE

Lessons Learned - General

- You need to be in it for the long haul
  - May need to change your data collection processes
  - Time to perform the audits to validate/refine your queries and models
  - Train staff and build relationships
- It is not just new data sources
  - Determine the cost of data sources and check to see if they are providing adequate value
  - Look at new ways to use data
  - Enhance collaboration
    - Not just Business and IT, but cross divisions and units

CSC

29

MINNESOTA · REVENUE

Questions

CSC

30

MINNESOTA · REVENUE

Contact Info

MINNESOTA · REVENUE

- Greg Tschida
  - Office of the CIO
  - Minnesota Department of Revenue
  - greg.tschida@state.mn.us
  - 651-556-6207

CSC

- Eric Bjorklund
  - Principal Consultant
  - Computer Sciences Corporation (CSC)
  - ebjorklu@csc.com
  - 763-567-6543

CSC

31