

# **Wisconsin Dept. of Revenue**

## **Tax Compliance Infrastructure**

**Data Warehouse Project**

**Janna Baganz  
August 6, 2007**

# Project Beginnings

## Corporate Data Warehouse



### Upside:

- Contractor built with a quick project turnaround
- Successful audit projects

### Downside:

- Difficult maintenance efforts

# **Project Turning Point**

## **Developed Business Plan for Individual Data Warehouse**



### **Development Decision**

- **Expand current corporate project**
- **Contract out new project**
- **Build in-house**

# **Individual Data Warehouse**

## **Build Data Warehouse In-House**

- **Integrated team across IT and business units**
- **Business users control what data to load and surface from the data warehouse for reporting**
- **Phased approach to provide value along the way**
- **Keeps technical and business knowledge in-house**
- **SAS platform selected**

# **Individual Data Warehouse**

## **Phased Approach**

- **Phase 1 – WI tax returns, IRS individual tax returns, and electronically filed W-2 forms**
- **Phase 2 – IRS informational returns**
- **Phase 3 – Vehicle title, drivers license and registration information**
- **Phase 4 – Sales tax, corporation returns, IRS business tax data, unemployment insurance employer records**



# **Tax Compliance Infrastructure**

## **Advantages of the Data Warehouse**

- **Matching individuals and entities allowing reports across data sources**
- **Summarizing various data to the level of tax return information for reporting purposes**
- **Identifies and tracks relationships embedded in existing data**
- **Continually modifying based on user input**
- **Useful for many bureaus – Audit, Compliance, Research & Policy, etc.**

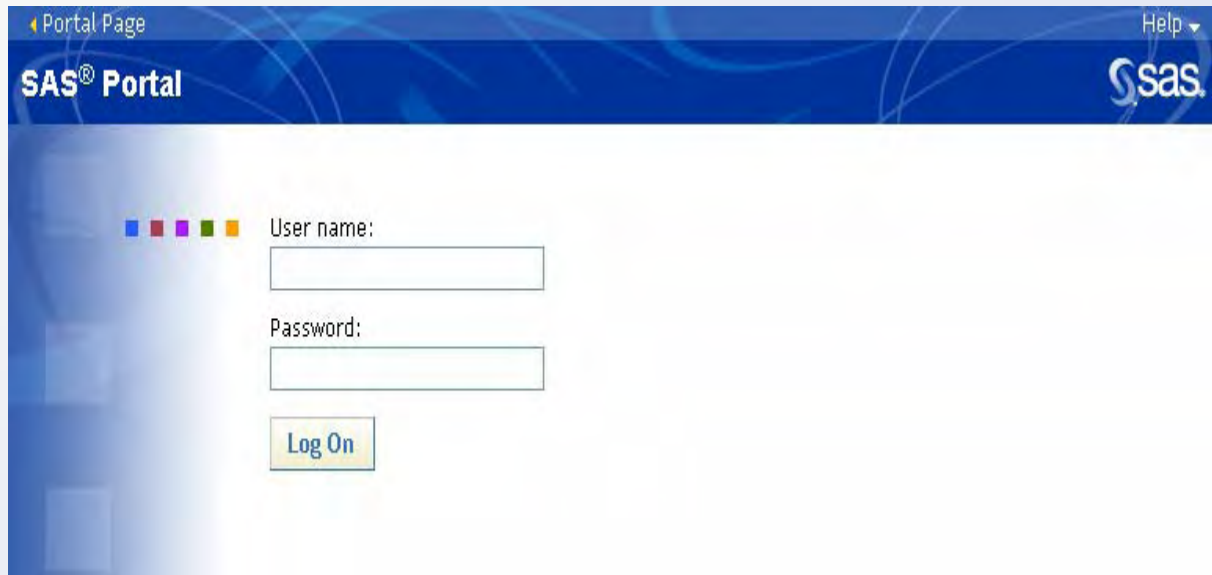
# **Data Warehouse Reports**

- **Income and Asset report**
- **Current Financials**
- **Address Report**
- **Multi-Year Compare**
- **Relationship Look-up**
- **Ad Hoc Capability**

# Data Warehouse Functionality

## Security Measures

- Secure user name and password to access
- Session timeout when inactive
- Software to track usage



The screenshot shows the SAS Portal login interface. At the top left, it says "Portal Page" and "SAS® Portal". At the top right, there is a "Help" dropdown menu and the SAS logo. The main content area features a login form with a "User name:" label and an input field, a "Password:" label and an input field, and a "Log On" button. To the left of the form, there are five small colored squares (blue, red, purple, green, yellow) and a vertical navigation bar with several square icons.

# Data Warehouse Functionality

## Security Measures

- Authorization disclosure
- Control of user access to data and system resources

The screenshot shows the SAS Portal Public Kiosk interface. The top navigation bar includes 'SAS® Portal' and 'sas.' logos, with tabs for 'Public Kiosk', 'Compliance', and 'Home'. The main content area displays a 'Warning' message:

**Warning**

This System contains State, Local, or Federal tax return information.

Unauthorized access, disclosure, printing, or publishing of federal tax return information is prohibited by the Internal Revenue Code Sections 8213, 7213A, 7431 and 18 USC Section 1905 and may result in criminal or civil action

Unauthorized access, disclosure, printing, or publishing of State and Local tax return information may also be prohibited by state or local laws.

Additional UI elements include 'Welcome Page' and 'Shared: PortalHomePage' labels, and 'Sticky: PortalHomePage' and 'Shared: PortalHomePage' status indicators.

# Data Warehouse Functionality

## Audit User Portal

The screenshot displays the Audit User Portal interface. At the top, there is a navigation bar with tabs for "Public Kiosk", "Audit", and "Home". The "Audit" tab is currently selected. In the top right corner, the user's session information is shown as "Sticky: SAS\_AUDIT\_BI\_CLIENT".

The main content area is organized into several sections, each with a "Shared: PortalHomePage" link and an expand/collapse icon:

- Audit Reports**: Contains a "Click here to refresh collection." link and a list of report files:
  - Duplicate Refund Identification.srx Audit Report 4
  - Non-Filer Compare Wl Withholding to Wages on W-2's.srx Audit Report 1
  - Taxpayers Filing Status Different for Federal and Wl.srx Audit Report 6
  - Under-Reporters W-2 Wages Greater Than GR Total Wages.srx Audit Report 3
  - Wl Withholding Different than Claimed on Reports.srx Audit Report 5
- Audit Writeback**: Contains a "Click here to refresh collection." link and two report files:
  - SP Audit Writeback
  - SP Audit Writeback FEIN1
- Compliance Reports**: Contains a "Click here to refresh collection." link and one report file:
  - Compliance Current Financials.srx Compliance Current Financials
- Acceptance Testing Area**: Contains a "Click here to refresh collection." link and one report file:
  - Multi-Year Return Compare
- Web Report Studio**: Contains a "Click here to refresh collection." link and one link:
  - Link to Web Report Studio
- Data Warehouse Documentation**: Contains a "Click here to refresh collection." link and one link:
  - SAS How To Instructions
- Revenue Shared Reports**: Contains a "Click here to refresh collection." link and a list of report files:
  - SP Indv Asset Report
  - SP Indv Address Report
  - SP Busn Address Report
  - Dependent\_LookUp.srx
  - DW Relationship Look Up.srx
  - W2 Info by Business.srx
  - Vehicle History Report.srx
  - K-1 by Business 1065 1120s 1041.srx
  - Stocks and Bonds from 1099-Bs.srx
  - DW Data Element Description Search.srx
  - DW Data Element Search.srx
  - IRMF 1099 by Business Payee.srx

# Data Warehouse Functionality

## Compliance User Portal

Public Kiosk **Compliance** Home

*Sticky: SAS\_COMP\_BI\_CLIENT*

### Compliance Reports

*Shared: PortalHomePage*

[Click here to refresh collection.](#)

[Compliance Current Financials.srx](#)  
Compliance Current Financials

### Revenue Shared Reports

*Shared: PortalHomePage*

[Click here to refresh collection.](#)

- [SP Indv Asset Report](#)
- [SP Indv Address Report](#)
- [SP Busn Address Report](#)
- [Dependent\\_LookUp.srx](#)
- [DW Relationship Look Up.srx](#)
- [W2 Info by Business.srx](#)
- [Vehicle History Report.srx](#)
- [K-1 by Business 1065 1120s 1041.srx](#)
- [Stocks and Bonds from 1099-Bs.srx](#)
- [DW Data Element Description Search.srx](#)
- [DW Data Element Search.srx](#)
- [IRMF 1099 by Business Payee.srx](#)

### Data Warehouse Documentation

*Shared: PortalHomePage*

[Click here to refresh collection.](#)

[SAS How To Instructions](#)

### Development Reports

*Shared: PortalHomePage*

[Click here to refresh collection.](#)

- [Compliance Current Financials DEV.srx](#)  
Foundation/BIP Tree/ReportStudio/Shared/Reports/Development
- [SP Indv Asset Report](#)  
Foundation/BIP  
Tree/ReportStudio/Shared/Reports/StoredProcesses/Development

### Acceptance Testing Area

*Shared: PortalHomePage*

[Click here to refresh collection.](#)

- [Multi-Year Return Compare](#)
- [Multi-Year Schedule C Compare](#)
- [Multi-Year Schedule F Compare](#)

# Data Warehouse Functionality

This section contains representations of the reports available within the Data Warehouse and does not include actual taxpayer information.

## Execution Options

Income and Asset  
Report

Multi-Year Compare

Multi-Year Schedule C

Current Financials

## Taxpayer Search

Top N:

SSN:

XXX-XX-XXXX

Last Name:

First Name:

Drivers  
License:

# Data Warehouse Reports

SAS Web Report Studio • Compliance Current Financials



Report ▾

Edit Report

View Report

How Do I? ▾

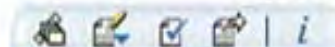
Refresh Data

## Current Address Info



Ssn	Name	Address	Address2	City	State	Zip	Zipext
000000008	John Smith	14 Quartz Ln		Bedrock	VA	00001	0003

## DOT Financial Info



Dot Date ▲	Dot Lender Id	Dot Lender Nm ▼	Dot Lender Address	Dot Lender City	Dot Lender State	Dot Lender Zip
03/30/2007	999999	ABC Lending Corp.	456 Main St	Bedrock	VA	00001

## W2 Employment Information



Empr Fein	Empr Name	Empr Name 2	W2 Source	W2 Yr
000000002	ABC Building Corp		W2 Electronic	2004

## IRMF 1099 Interest Info



Irmf Fein	Irmf Payer	Irmf Payer 2	Irmf Source	Irmf Yr
999999999	ABC Banking		IRMF 1099 Int	2004

# Data Warehouse Reports

Portal Welcome REVJ61 Help

SAS Web Report Viewer • Individual Address Multi Search Criteria

Report How Do ?

Address Report Search Page Search by SSN Search by Name - Current Address Search by Drivers License No. ALL Address Search Refresh Data



Wisconsin Department of Revenue  
revenue.wi.gov

Please select search tab to enter search criteria.

SAS Web Report Viewer Individual Address Multi Search Criteria

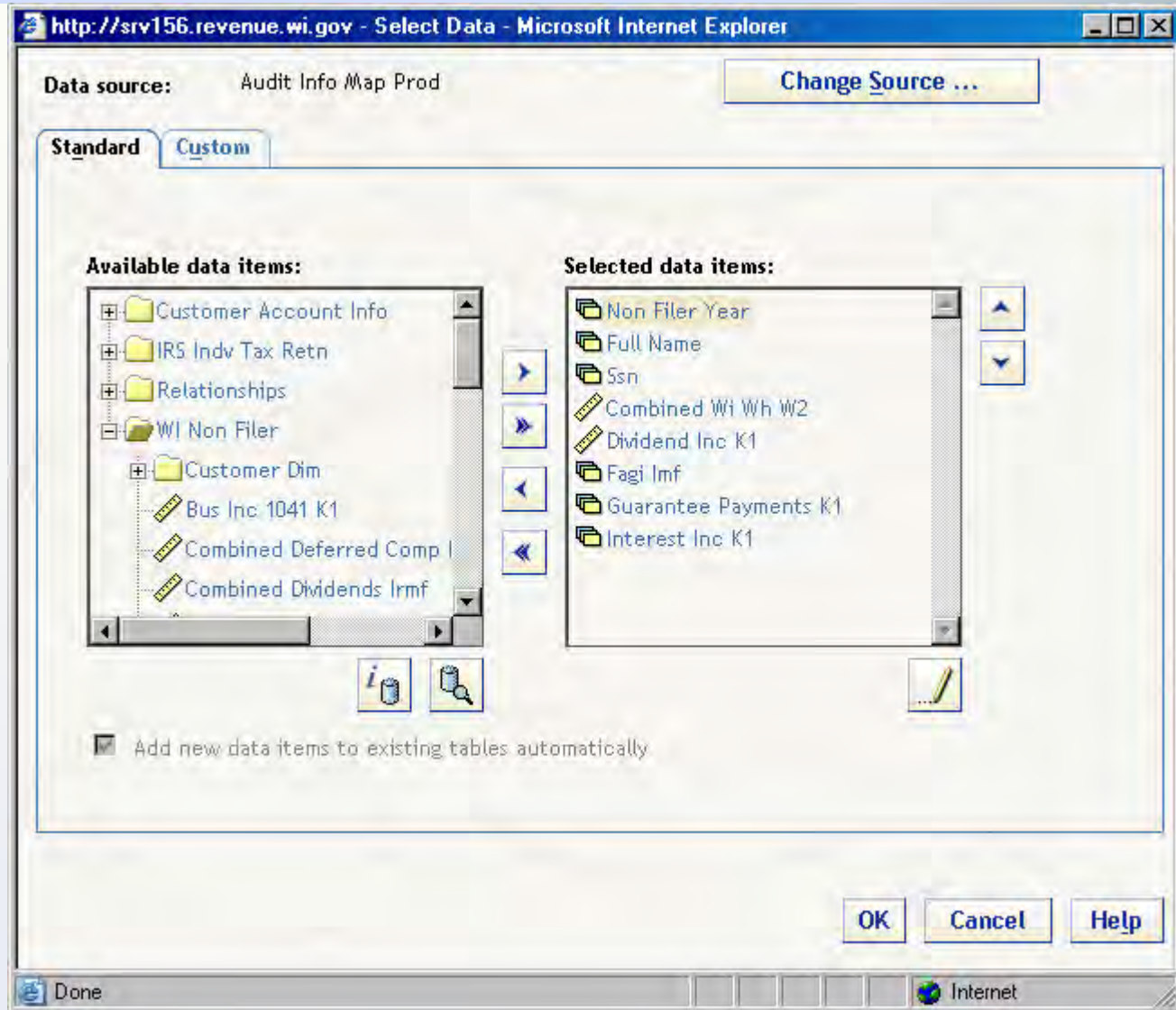
Report How Do ?

Please answer the prompts below and click the View Report button to continue.

* Please enter First Name:	<input data-bbox="1054 986 1245 1011" type="text" value="%"/>
* Please Enter Last Name:	<input data-bbox="1054 1068 1245 1092" type="text" value="%"/>
* Please Enter Street Address	<input data-bbox="1054 1149 1245 1173" type="text" value="%holton%"/>
* Please enter City:	<input data-bbox="1054 1230 1245 1255" type="text" value="milwaukee"/>
* Please Enter State:	<input data-bbox="1054 1312 1245 1336" type="text" value="wi"/>

View Report Reset to Defaults

# Ad Hoc Query Capability



# **WI Warehouse Success**

## **Workload and time savings**

- **Dynamic environment allowing for continuing evolution of useful reports and business projects across data sources**
- **Continue to work towards goal of performing better audits with better results**
- **Information Center supported by DW Staff**
- **Audit projects**
  - **Merger of warehouses will help identify non-filers**
  - **Project has identified \$19 million in assessments by comparing Form K-1 and Form 5S**
  - **Recycling surcharge project has already assessed \$400,000 and collected \$250,000**



**THE  
POWER  
TO KNOW®**

# Advanced Analytics for Identifying Tax Under- filers

---

Revathi Subramanian  
FTA, Kansas City  
August 6, 2007

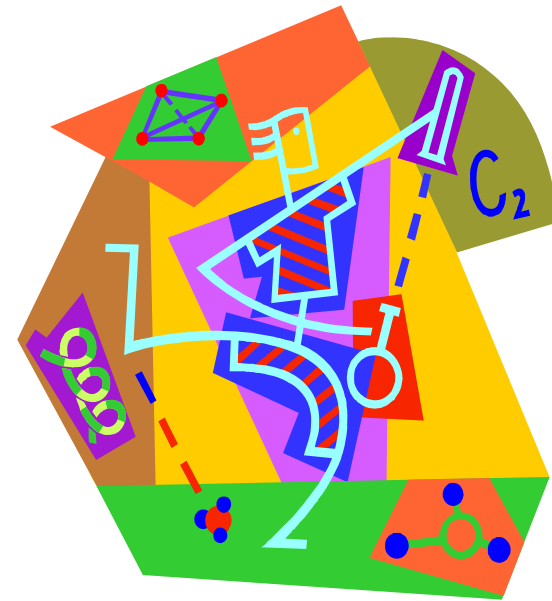
# Predictive Modeling

- Predictive modeling used very effectively to address high value risk problems
- Financial industry a great example of the benefits realized
- Billions of \$ saved by forecasting risk early enough so that effective action is taken



# Broad Classes of Problems

- Supervised Learning:
  - Target fully known
  - Learn from the examples and extrapolate in a robust way
  - Credit card fraud, bankruptcy good examples
- Semi-Supervised Learning:
  - Partial target known
  - Learn from the known targets as well as anomalous behavior to predict risk
  - Tax fraud, purchase card fraud good examples



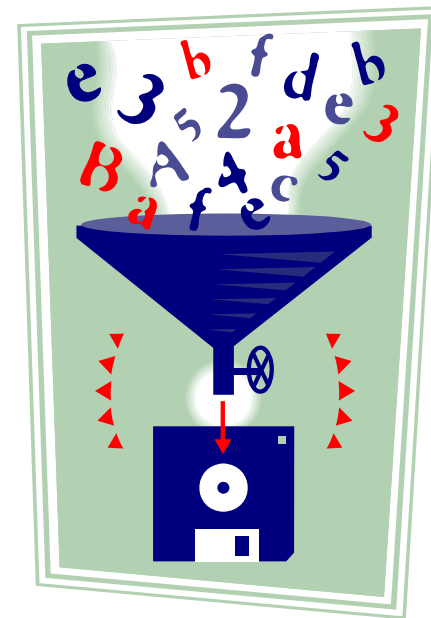
# Broad Classes of Problems

- Unsupervised Learning:
  - Target completely unknown
  - Learn from anomalous behavior and isolate cases
  - Insurance fraud, network intrusion good examples



# Converting Data to Information

- The world has spent significant \$ in aggregating data
- Data only as useful as the information that can be extracted from it
- Extracting **actionable** and **timely** information from data is a non-trivial exercise
- Advanced techniques almost a requirement for a number of risk / collection problems



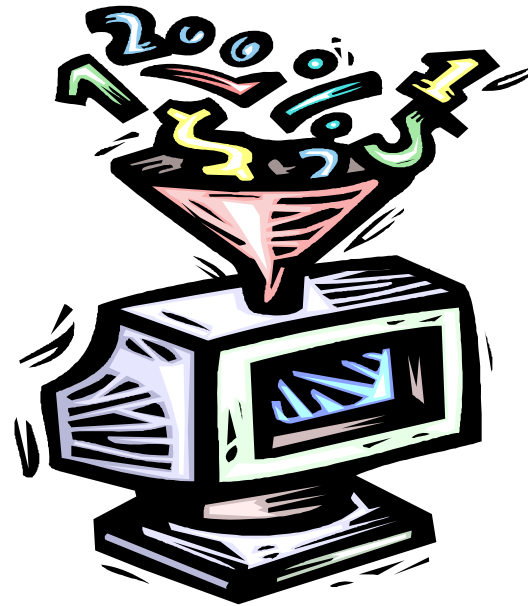
# Converting Information to Action

- The advanced models built on tax data used to create a score reflective of the likelihood of under filing on the account
- The scores can be used by tax departments to develop operational rules using SAS tools to take action
- The actions taken and the results tracked and reported on will make the process continuously better



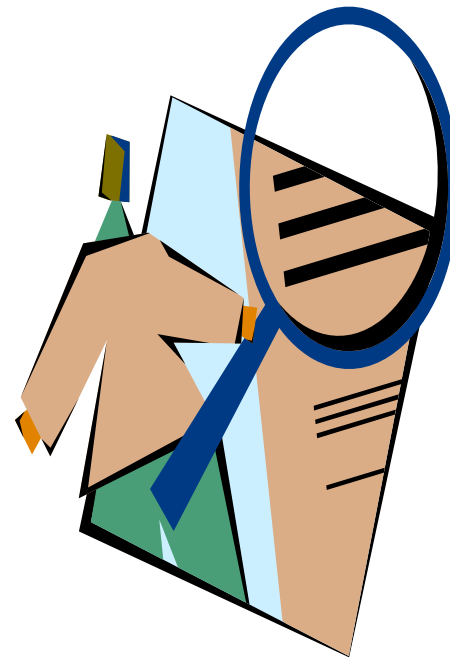
# Evolution of Detection Systems over Time

- Data Discovery & Reporting
- Correlation Analysis
- Predicting the outcome using rules
- Predicting the outcome using advanced predictive models



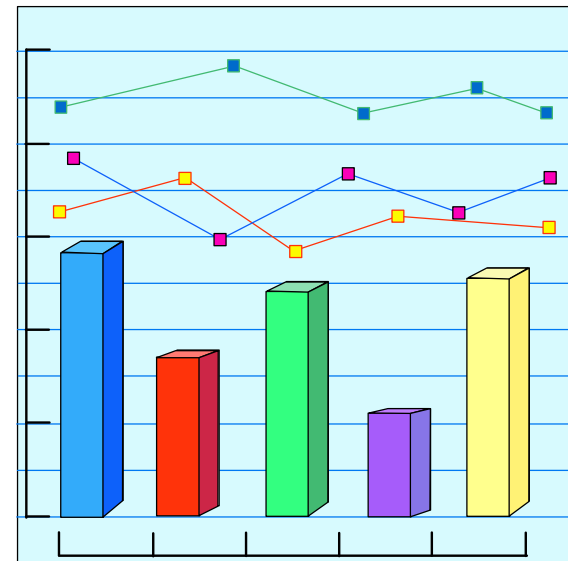
# Data Discovery

- Typically happens after the data has been aggregated
- Focus is on summarizing and understanding data
- Can be done using simple reports and univariate analyses
- Focus is to summarize what has happened



# Correlation Analysis

- Goes a step beyond summarizing data
- Focus is to understand how different pieces of the data relate to each other
- Simple relationships between attributes are discovered
- Correlation does not imply causal relationship



# Simple Example

- Study structures that caught fire over the last year
- # of fire trucks will have high correlation with size of fire
- Cannot decide to reduce the number of fire trucks sent in the next fire to reduce size of fire!
- Correlation does not imply causality!
- Establishing causality requires domain expertise, out-of-time validations, advanced analytical variables and models



# Prediction Using Simple Rules

- Correlation analysis naturally evolves into simple rules
- Rules are easy to understand and implement, can catch obvious cases
- Rules typically do not utilize the full breadth and depth of data
- Rigid cut offs do not help with precision / rank ordering
- As rules multiply, too many entities are classified as “targets”
- There is an explosion in the number of cases to examine



# Prediction Using Simple Rules (Contd.)

- Provide an automated system to support the risk management
  - Better quality and quantity of risk referrals
  - Consistent, objective risk referrals
  - Reduction in risk loss
- Combination of multiple tools and technologies
- An expert rule system generates automated referrals
  - Using Predictive indicators/red flags
  - May use decision trees or similar methods to identify patterns of fraudulent behavior.



# Audit Selection Project Overview

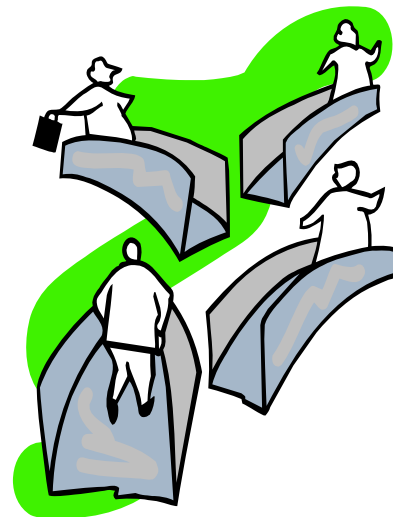
- SAS set out to create scores that rank order tax-returns such that the highest scoring returns are highly likely to be under-filers
- WI return, IMF, IRTF files received for 2000 - 2005
- A small set of past audits and results received
- Based on the known examples and anomaly detection techniques, SAS produced a list of preliminary audit candidates
- WI DOR safeguarded taxpayer confidential information by the following:
  - All personal identifiable information was cleansed from all data sources prior to submission to SAS
  - SAS entered into non-disclosure agreement with WI DOR to ensure data security
  - IRS approved usage under general tax administration rules

# Sampling for Preliminary Audit List

- Step 1: Identify the top scoring x% returns
- Step 2: Eliminate returns already in the existing audit file
- Step 3: Sample returns from the remaining records such that:
  - Many of the returns are top scoring high likelihood audit candidates
  - Many of the returns are borderline between highly likely and somewhat likely audit candidates
  - Somewhat likely audit candidates picked solely for the purpose of improving analytical models!
- SAS looked at the new audit results and incorporated the new known cases to retrain the models

# Problem Complexity

- WDoR audit selection scenario falls under semi-supervised learning
- Need sophisticated algorithms to effectively reduce false positive rate
- The SAS team has addressed similar problems in different industries



# Problem Definition

- Examine tax returns to identify under payment of taxes by individuals
- Identify under-filers with highest payback potential
- Use all available data

# Risk and Magnitude Scores

- Identify individual tax returns with highest payback potential
- Three sets of models
- Audit risk replacement score (ARR)
  - Value between 50 and 999
  - Supervised models built using past audits from 2000 – 2004 as target
- Audit risk augmentation score (ARA)
  - Value between 50 and 999
  - Unsupervised models built to identify outliers

# Risk and Magnitude Scores (Contd.)

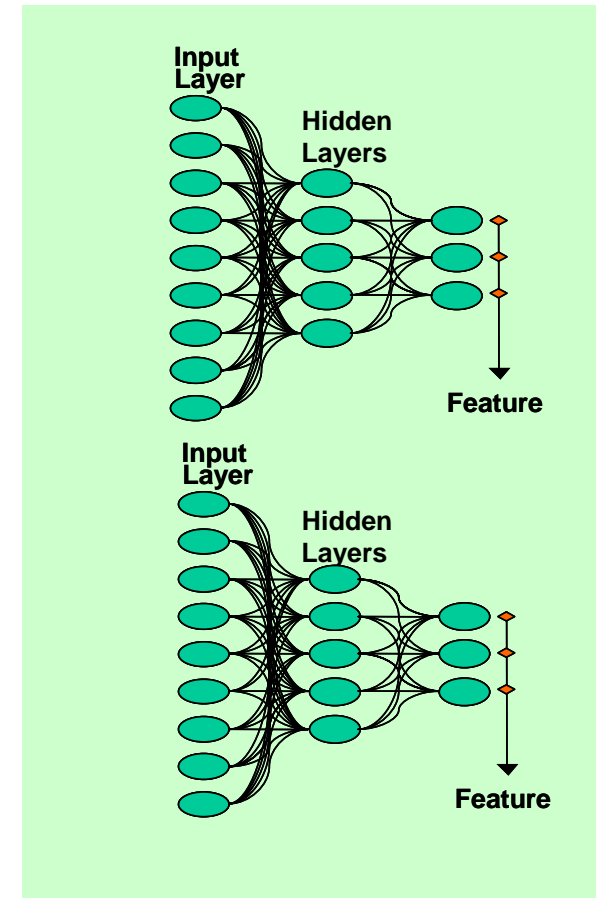
- Audit magnitude score (AM)
  - Take non-negative values
  - Rank order tax returns with the highest assessment magnitude getting the highest scores
- Combine the scores to rank order tax returns with the highest payback potential

# Semi-Supervised Learning Strategies

- A very small percentage of the records are tagged
- Strategy 1
  - Build two models using disparate features
  - Score all the untagged records using the models
  - Present records with highest discrepancy between two models to domain expert for tagging
- Strategy 2
  - Build a single model
  - Present records with highest uncertainty to domain expert for tagging

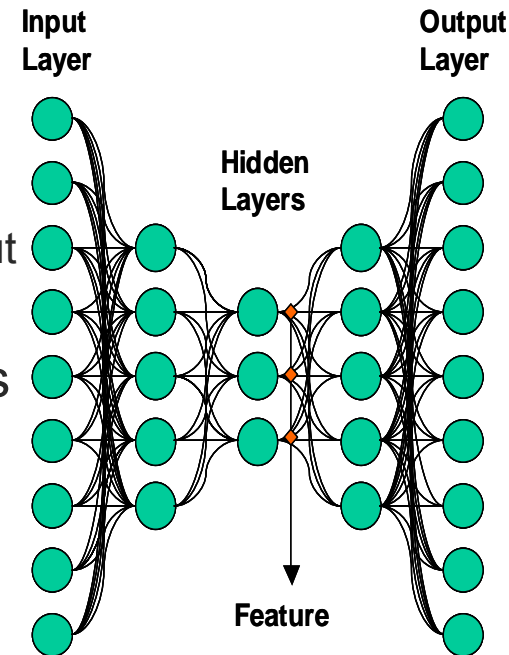
# Back-Propagation Neural Networks (BNN)

- A powerful nonlinear learning technology
- An extension of well known linear methods such as multivariate regression
- BNN can be described in a two-step algorithm
  - inputs are first fed forward through the neural network and the error in predicting the audit magnitude computed.
  - The prediction error is then back-propagated to change the network weights.
  - This procedure continues until the network weights reach a steady state

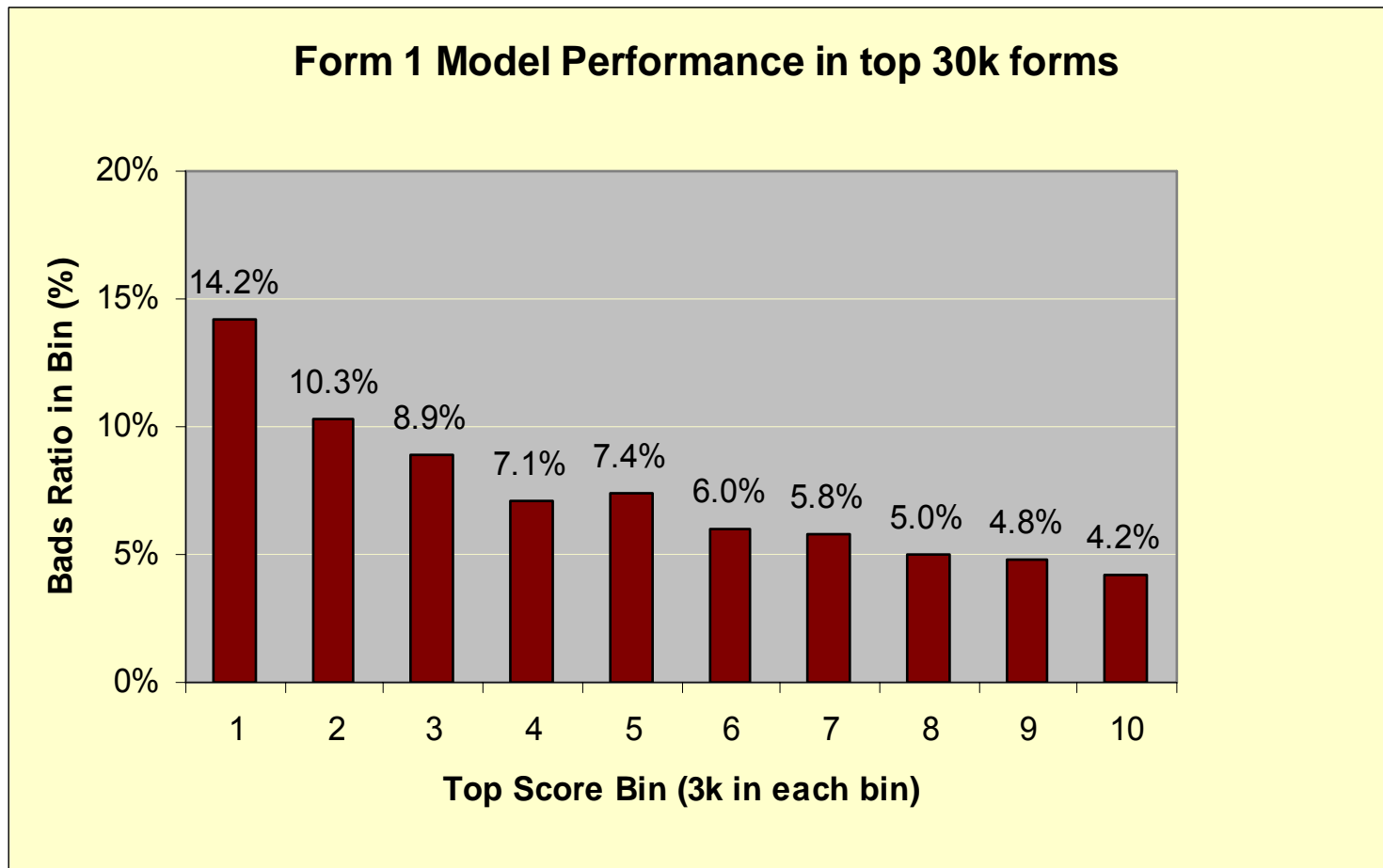


# Autoregressive Neural Networks (ANN)

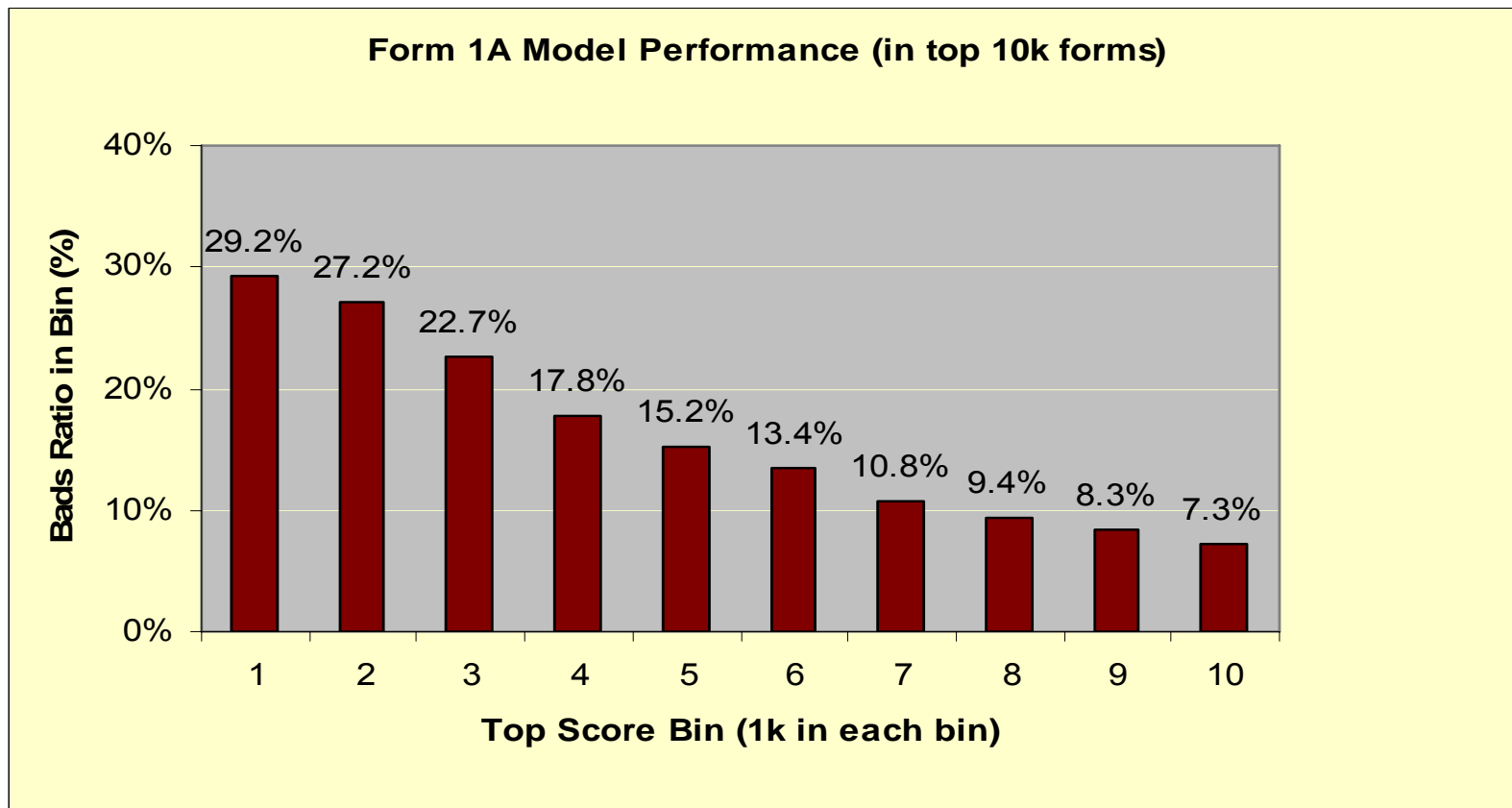
- ANNs are nonlinear extensions of well known linear methods (PCA) for outlier detection.
- Simple two-step algorithm:
  - Input is first **compressed** to a lower dimensional manifold first - analogous to “zipping” a Word doc.
  - Intermediate data is then **uncompressed** back at output – analogous to “unzipping” your Word doc.
- The error in reconstructing a data point from its compressed form is measure of anomaly:
  - “Normal” activity will have low reconstruction error.
  - “Abnormal” activity will have high error.
- Advantages:
  - Applicable to any data distribution
  - Multilayer structure handles complex, nonlinear interactions



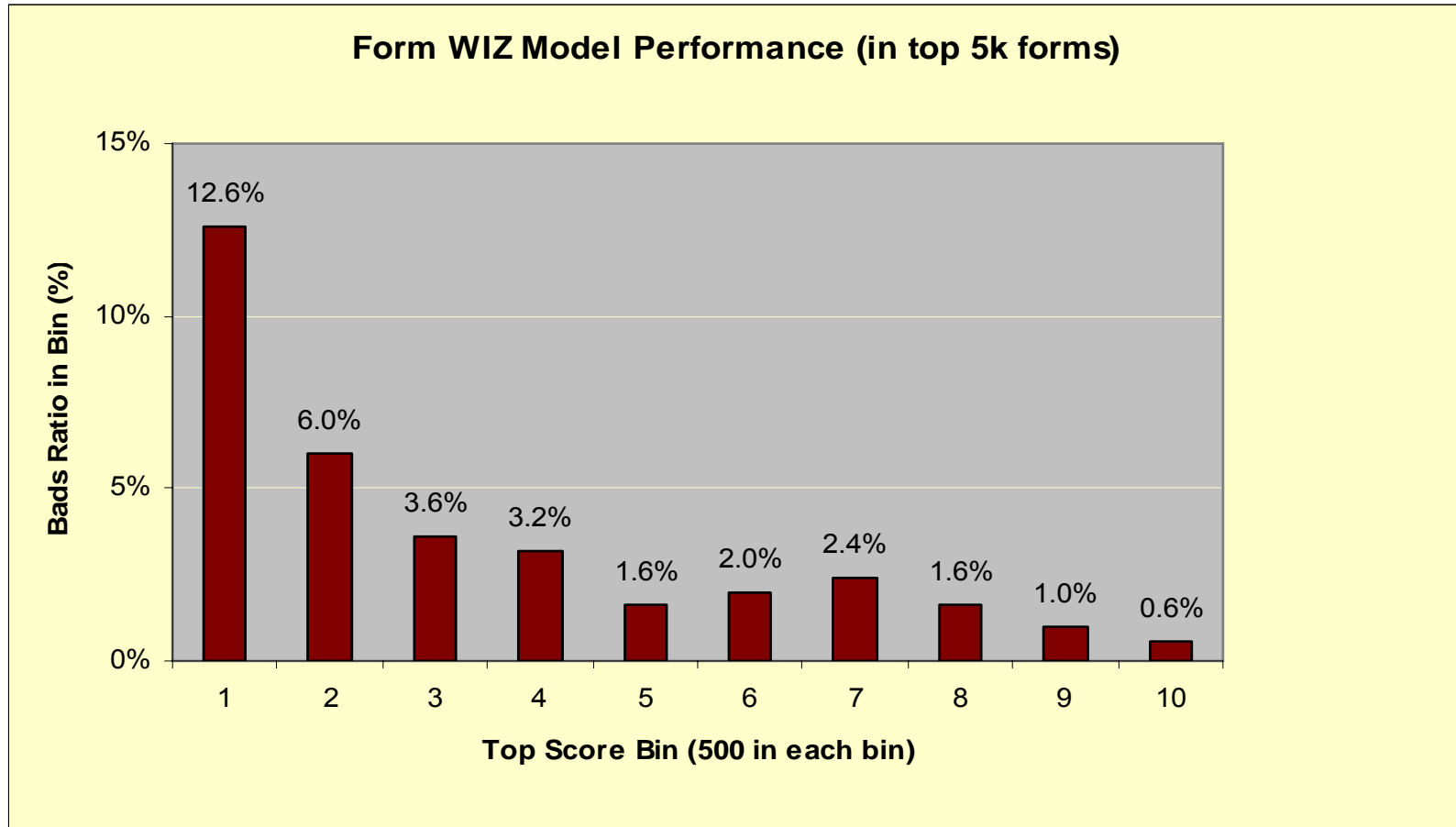
# ARR Score Performance: Form 1



# ARR Score Performance: Form 1A



# ARR Score Performance: Form WI-Z

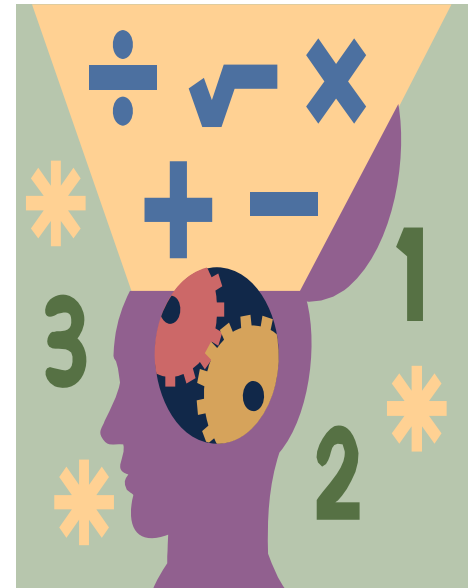


# Reason Code Examples

	Example 1	Example 2	Example 3	Example 4	Example 5
Cust_sk	3858876	914890	213843	529738	1966383
Year	2002, 2003	2002	2003	2002	2001
Reason 1	tot_itdd_amt_irtf_r	almy_paid_amt_irtf_r	wi_txbl_incm_amt_gr_r	Map_schd_c_grs_incm	Almy_rcvd_amt_irtf_r
Reason 2	bal_txbl_incm_r	rt_subt_ftpi_r	tot_itdd_amt_irtf_r	rt_sttax_tpi_r	TXBL_IRA_DTR BN_AMT_IRTF
Reason 3	bal_agi_r	othr_subt_amt_gr_r	rt_grsTax_ftpi_r	Tot_idd_amt_irtf_r	Bal_schC_prft

# Using Risk Scores to Generate Audit Cases

- Two alternatives
  - Incorporate scores into existing processes
  - Feed scores into expert rule system as one of the inputs to generate audit cases
- Advantages of using the risk scores
  - Fewer rules are needed as the risk score is an overall assessment of risk by considering many factors
  - Simplify the rule writing process
  - Generate fewer false-positives
  - Models are easier to build and maintain than expert rule system with hundreds of rules
  - Risk score also provide a rank order to prioritize the investigations



# Incorporate Scores into Existing Processes

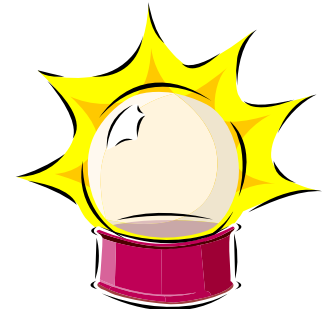
- Easy to incorporate scores into existing processes
  - Rank order accounts flagged by existing processes using the ARA, ARR scores to work on the most likely under-filers first
  - Rank order accounts flagged by existing processes using the AM score to work on the most valuable under-filers first
  - Use AM score to decide between office audits versus field audits
  - Use scores to assign cases between experienced versus novice auditors
- Using scores in existing environment will increase familiarity
  - Facilitate transition to a score based rules system
  - Increase understanding of scores as well as reason codes

# Detection Using Predictive Scoring

- A More Comprehensive solution — facilitates both detection and investigation capability
- Constant monitoring — monitor every entity for signs of fraud using many data sources
- Sophisticated scoring models and early alert engine — for state-of-the-art multi-stage detection capability
- Flexible to create referrals — integrate and support client specific usage strategies
- Advanced analytical tools — to assist investigation and reporting

# Why Do Both Rules and Scoring?

- Rules allow the input of client specific intellectual property and operation constraints
- Rules allow tracking and adjustments for new or short term risk patterns
- Models pick up **non-obvious risk patterns** and behaviors
- The combination of rules and scores provides better detection rate and better quality referral
- Business implication - with the same amount of resource,
  - Catching more risk activity
  - Catching them earlier
- Faster way to get a good ROI





**THE  
POWER  
TO KNOW®**

Thank you!