FRAUD ANALYTICS AND DATA WAREHOUSING

Office of the Comptroller
State of Maryland

Andrew Schaufele; Director, Bureau of Revenue Estimates

Agenda

- Our Team
- Background
- Case with TBD Success
- Successful Case
- Lessons Learned

Team

- Deputy Comptroller
  - David Roose
- Bureau of Revenue Estimates
  - Project Management
- RAD – QRDT
  - Principal Users
- Teradata
  - Primary Contractor
- ASR Analytics
  - Primary Analytics Developer

Existing Processes: QRDT (Questionable Return Detection Team)

QRDT Processes – Pre Updates

- TY2013 Capacity: 130,000 returns
  - Stopped $23 million
- Myriad stand alone metrics
  - Ex: Refund/Withholding Ratio
- Aggregate Hit Rate: ~10% (90% False Positive)
  - Significant Hit Rate Variance Among Programs

TBD Case:
Census Model
(Geospatial Analysis)

Business Case & Proposed Solutions

- Fraud may be Concentrated in Specific Geographic Areas
- Smallest Geographic Area Available is Zip Code
- Idea:
  - Geocode Return to more Finite Areas
  - Review Returns from Areas with Anomalous Return Counts Relative to History/Census Data

Initial Solution

- Interactive Drill Down To Identify Fraudulent RefundsReport That Grouped Similar Anomalous Returns Together
- By Census Block
  - Zero Population Anomaly – Verify the block exists and is populated
  - Longitudinal Historical Filing Time – 95% confidence interval for when returns are filed
  - New Census Block Anomaly – Has Census population but no returns in prior year
- Additional Value Added (Discovered during development)
  - Out-of-State P.O. Box Frequency – By zipcode
  - Mail Forwarding Services – Identified service forwarding large amount of returns for “residents” of no income tax states

Revised Solution

- Interactive Drill Down Report That Grouped Similar Anomalous Returns Together
- Included All Components Of The Initial Solution, plus +
  - Analytical Model Anomaly (older model)
  - Analytical Model Geographic Anomaly – by block
  - Refund to Wage Anomaly – >25%
  - Income to Withholding Anomaly – >25%
  - Many others

Results Census

- Implemented 10/1/2014
- Not Used

Reasons for “TBD” Status

- Final Product not Ideal for Organizational Structure
  - End product required research of an analytical nature
    - After finding one anomalous return, would guide you to patterns
  - Did not “auto-suspend”
- Learning Curve
  - Should have used E-file Database
  - Daily Loads not Possible (anticipated at kickoff)
  - Current Year Data Loaded after Start of Filing Season

Successful Case:
Fraud Scoring Algorithm

Business Case

- Agency Focus -- Taxpayer Service
  - Reduce false positives while getting legitimate taxpayers their money in a timely manner
  - Balancing taxpayer service while protecting State and residents

Proposed Solution

- Idea: Utilize scoring algorithm to "Triage" all returns
  - Estimate probability return is fraudulent based off historical correlations
  - Increase fraud $ while increasing efficiency

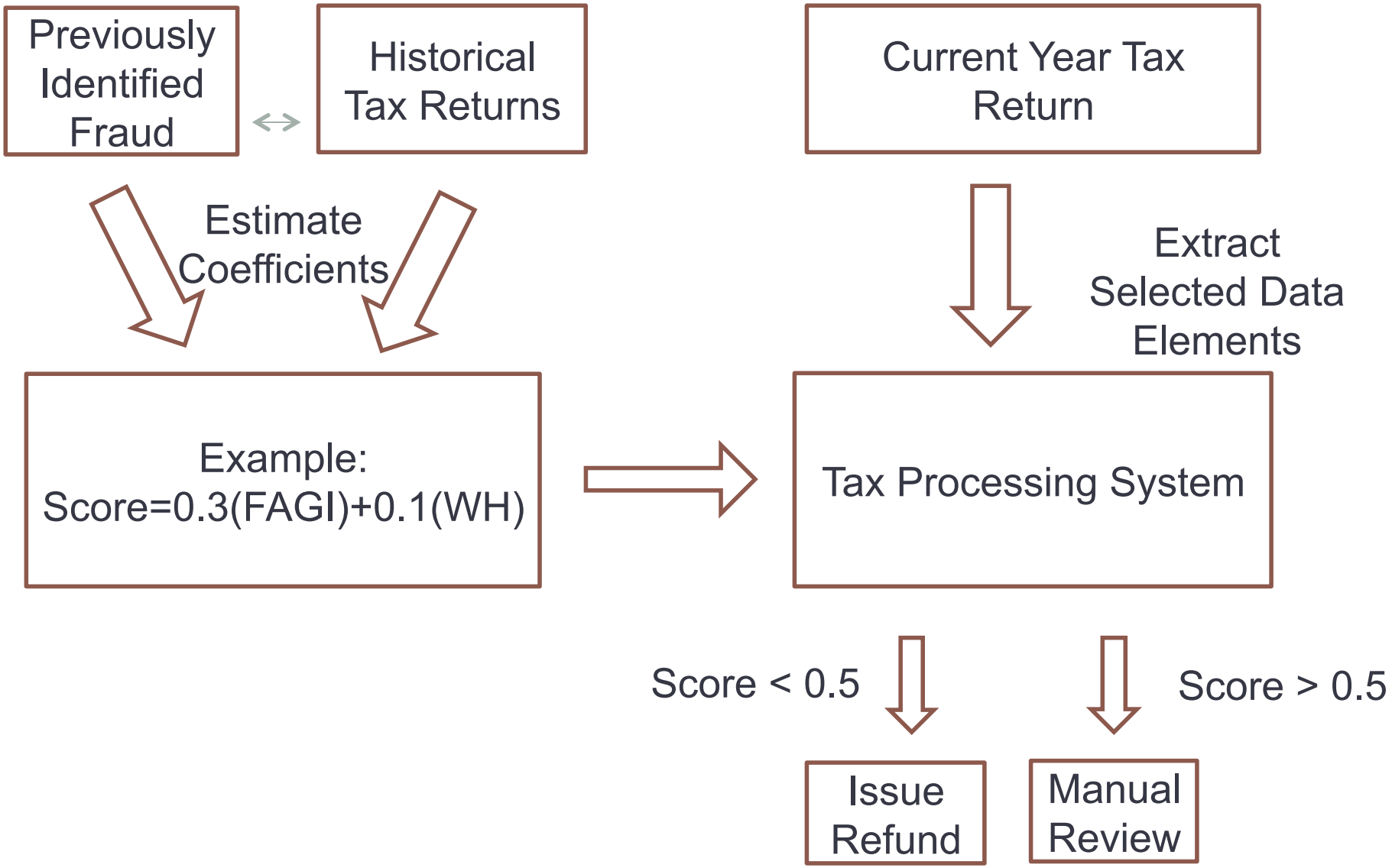
Implementation

- In 2011 deployed scoring algorithm directly in tax processing system
- Tried for three years with mixed results
- In 2015 re-estimated and deployed in data warehouse
- Results are preliminary but **Very Promising**

First Attempt

Perform Scoring in Processing System

First In Service For TY2011

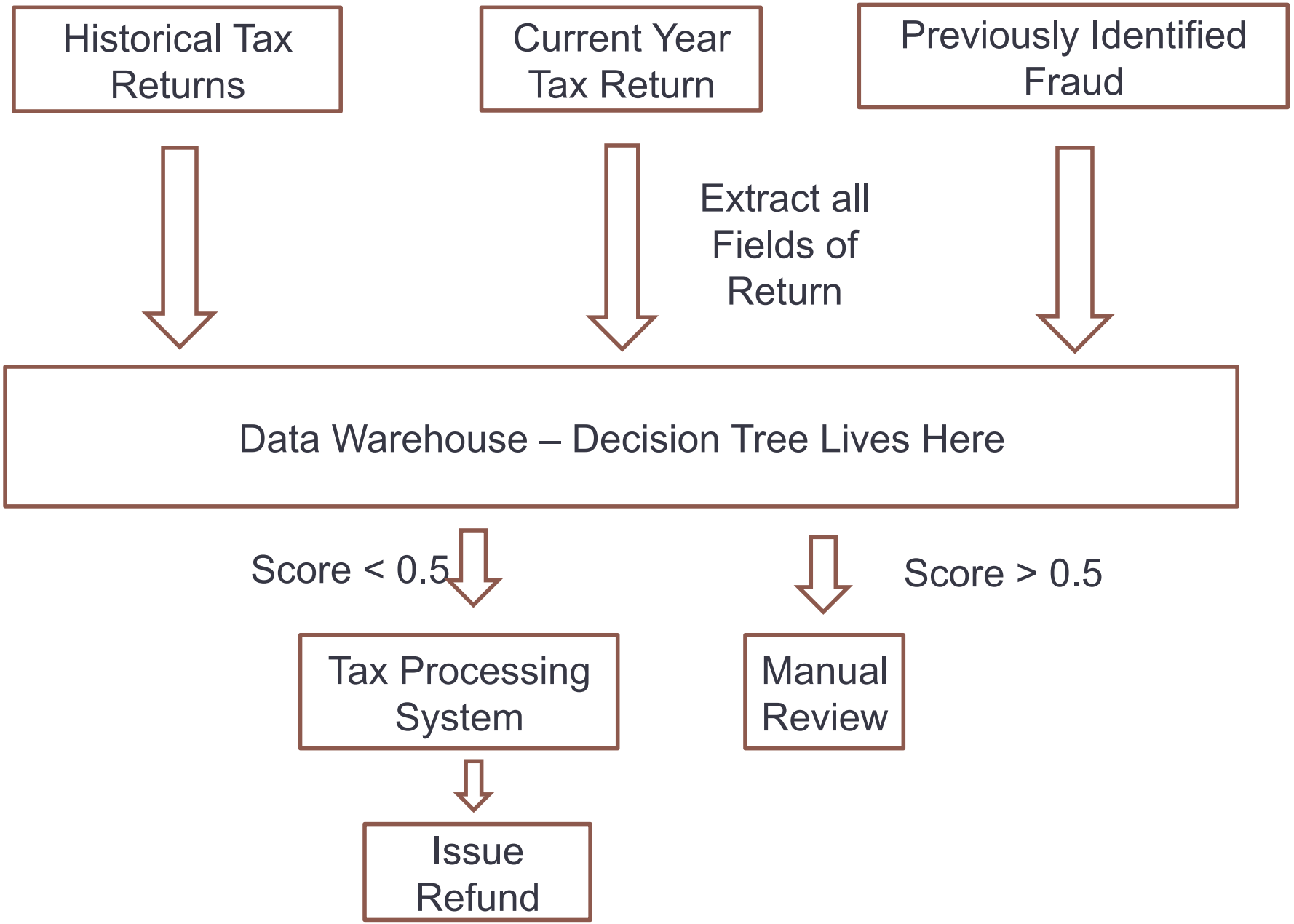

Previously Identified Fraud
Historical Tax Returns
Current Year Tax Return
Estimate Coefficients
Extract Selected Data Elements
Example: Score=0.3(FAGI)+0.1(WH)
Tax Processing System
Score < 0.5
Score > 0.5
Issue Refund
Manual Review

Results: Model #1

- Active for tax years 2011-2013
- 72,086 Returns ID'ed (24,000/ Year)
- 10.53% Returns ID'ed were fraud
- $14.4 Million in fraud ($4.8M/Year)

Second Attempt:

Perform Scoring in Data Warehouse


Historical Tax Returns
Current Year Tax Return
Previously Identified Fraud
Extract all Fields of Return
Data Warehouse – Decision Tree Lives Here
Score < 0.5
Score > 0.5
Tax Processing System
Manual Review
Issue Refund

Results: Model #2

- Deployed on 4/6/2015
  - Two months active
  - Not Peak Fraud Time
- 5,869 Returns ID'ed (Est. 33K/Year)
- 55.31% Returns ID'ed were fraud
- $7 Million in fraud (Est. $24M/Year)
- Auto-suspends!

Results: Model #2

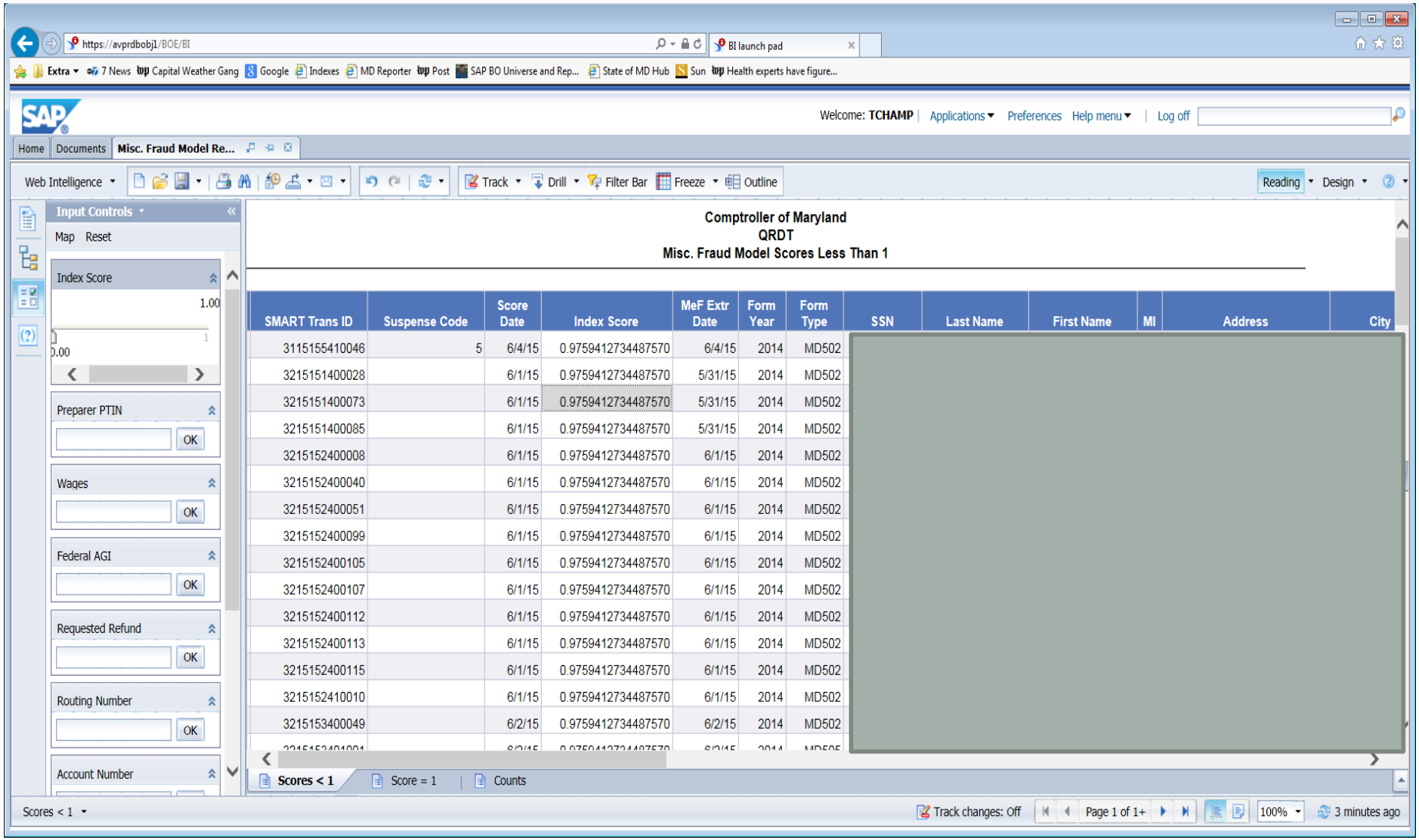
COMPTROLLER of MARYLAND
Serving the People

Comptroller of Maryland
TRACE Analytics – Miscellaneous Fraud

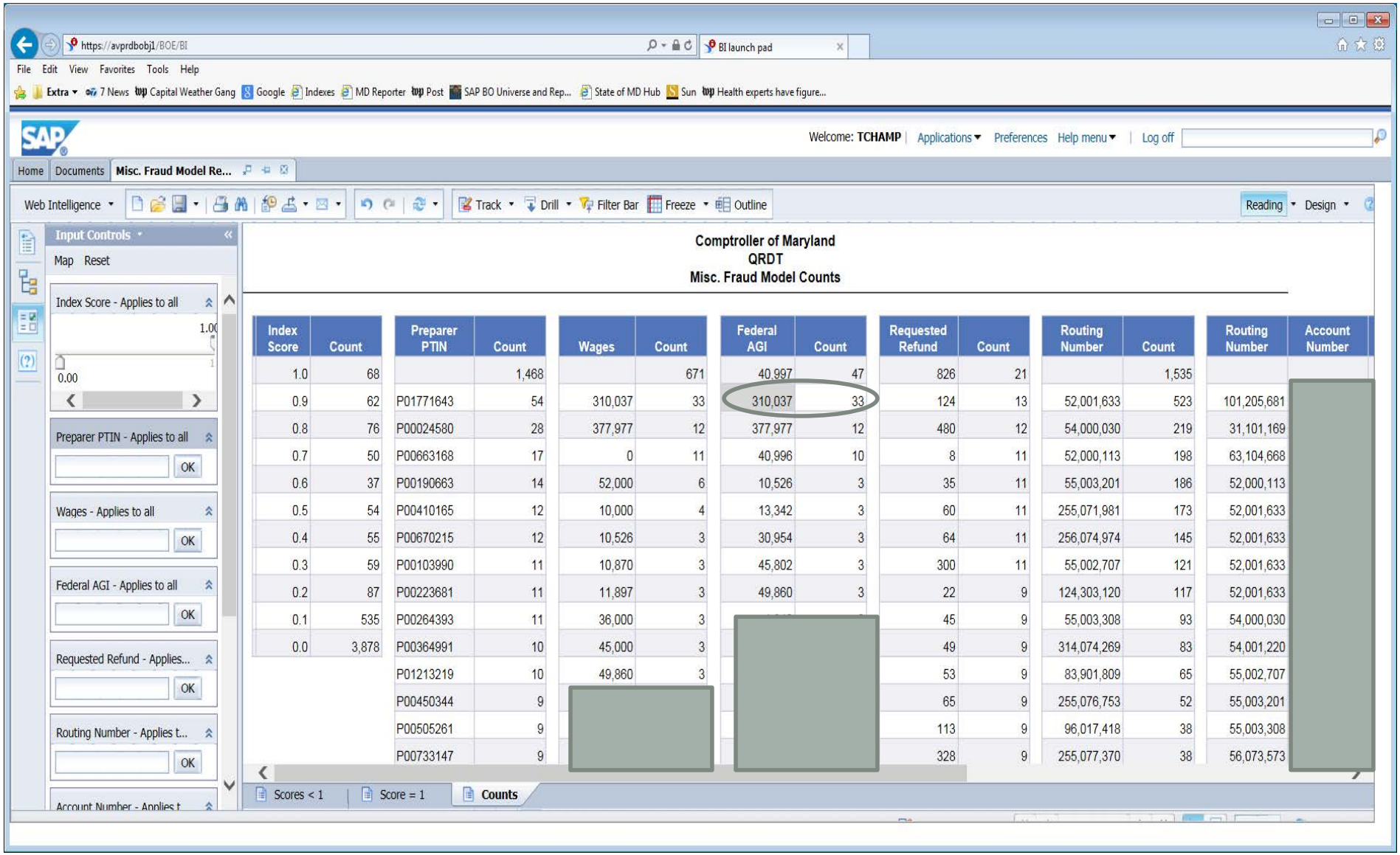
Table 3: Tax Year 2013 Model Results

	Returns	Stopped By Legacy QRDT Processes	NOT Stopped By Legacy Processes	Known Fraud	Known Fraud Rate	Adjusted Fraud Rate	Percentage of Known Fraud Dollars Captured	Fraud Dollars Recovered (Millions)	Projected Unrecovered Fraud (Millions)	Total Dollar Potential (Millions)
Scores >=0.80	13,260	12,167	1,093	9,377	71%	77%	71%	$16.4	$1.3	$17.7
Scores >=0.60	23,686	17,782	5,904	11,047	47%	62%	79%	$18.8	$5.3	$24.1
Scores >=0.5	31,057	21,393	9,664	11,569	37%	54%	82%	$19.4	$7.3	$26.7
Current Capacity	110,784	42,121	68,663	12,911	12%	31%	92%	$21.1	$24	$45.2

Output 1 - Worklist

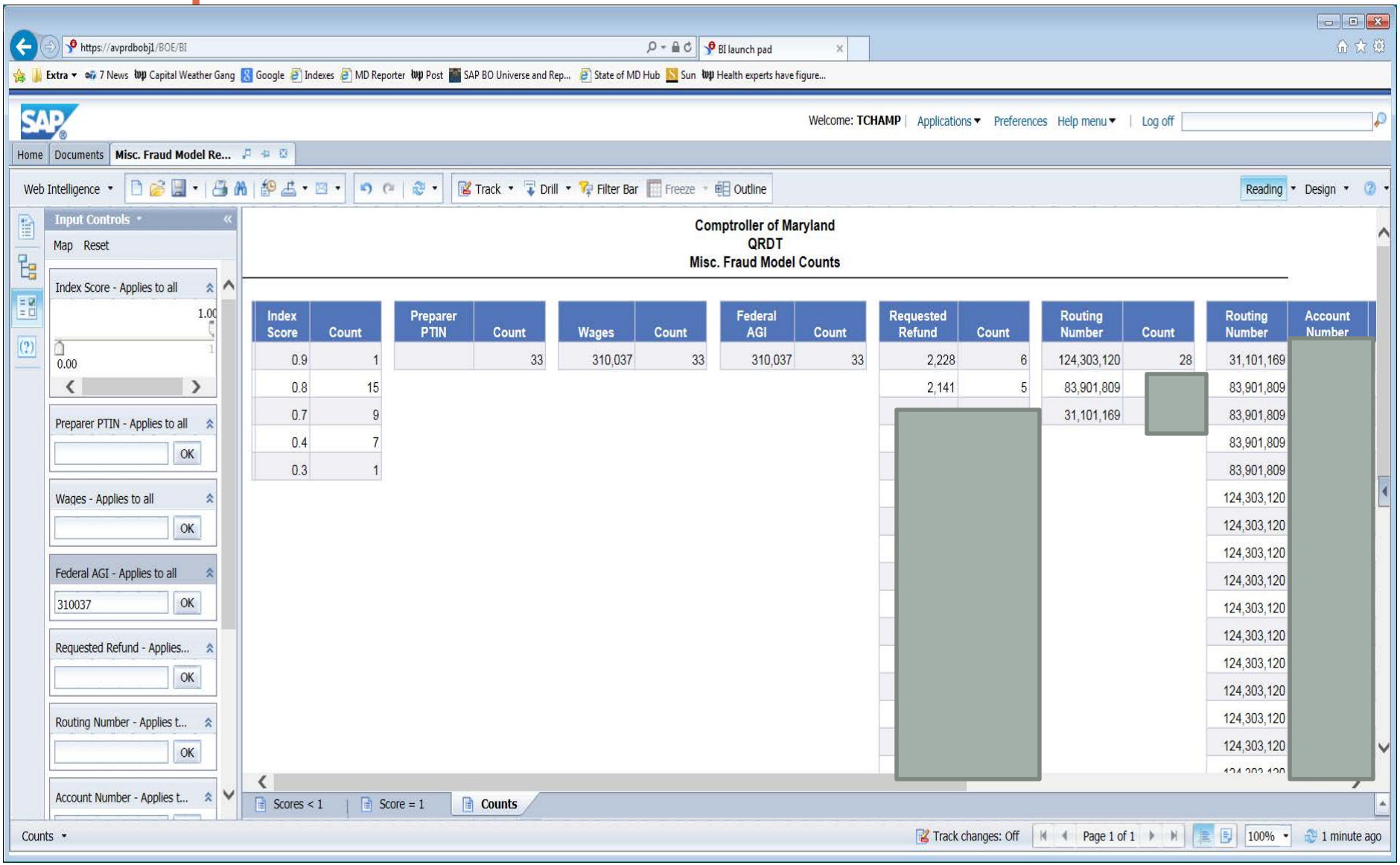


Output 2a – In House Enhancements

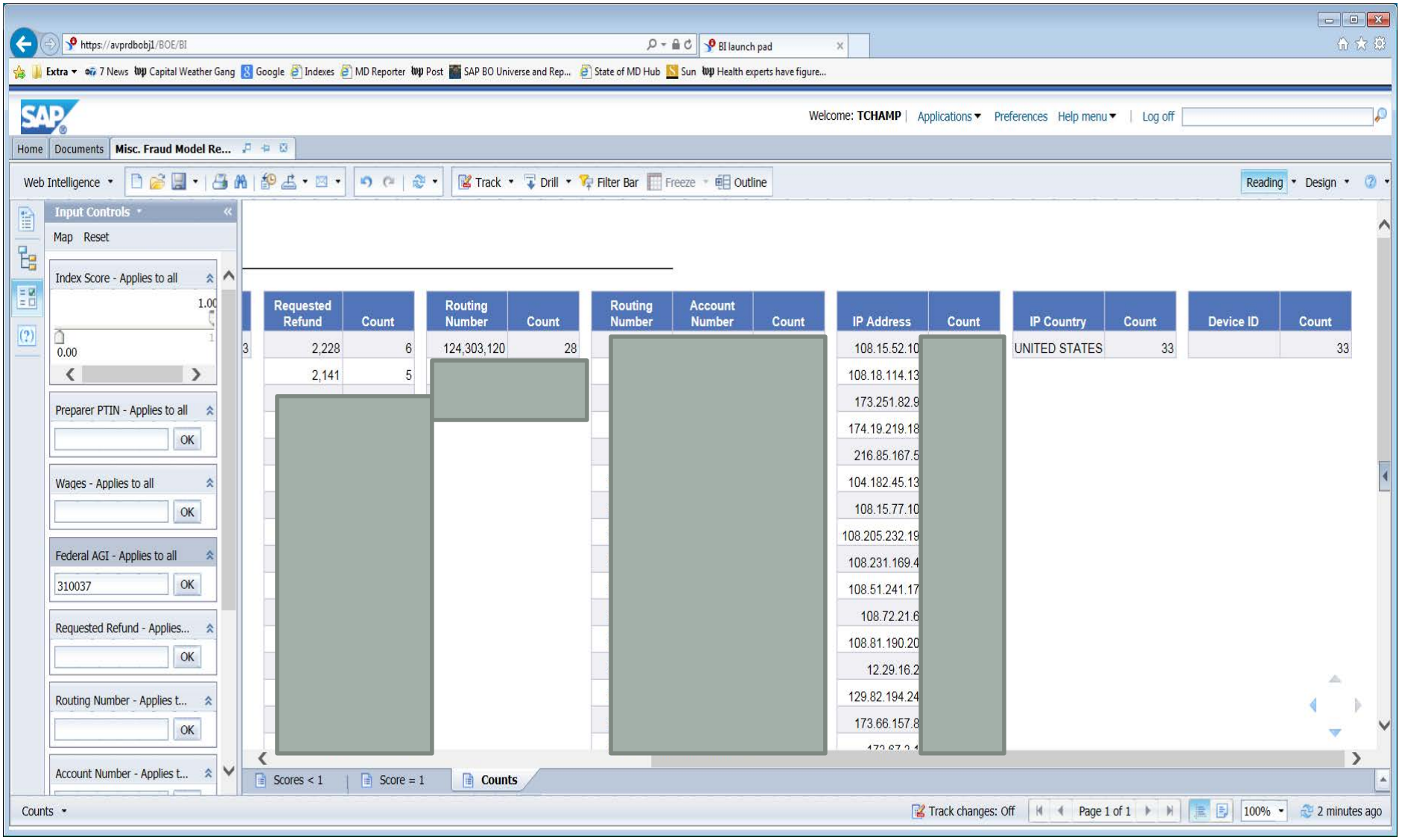


Index Score	Count	Preparer PTIN	Count	Wages	Count	Federal AGI	Count	Requested Refund	Count	Routing Number	Count	Routing Number	Account Number
1.0	68		1,468		671	40,997	47	826	21		1,535		
0.9	62	P01771643	54	310,037	33	310,037	33	124	13	52,001,633	523	101,205,681	
0.8	76	P00024580	28	377,977	12	377,977	12	480	12	54,000,030	219	31,101,169	
0.7	50	P00663168	17	0	11	40,996	10	8	11	52,000,113	198	63,104,668	
0.6	37	P00190663	14	52,000	6	10,526	3	35	11	55,003,201	186	52,000,113	
0.5	54	P00410165	12	10,000	4	13,342	3	60	11	255,071,981	173	52,001,633	
0.4	55	P00670215	12	10,526	3	30,954	3	64	11	256,074,974	145	52,001,633	
0.3	59	P00103990	11	10,870	3	45,802	3	300	11	55,002,707	121	52,001,633	
0.2	87	P00223681	11	11,897	3	49,860	3	22	9	124,303,120	117	52,001,633	
0.1	535	P00264393	11	36,000	3			45	9	55,003,308	93	54,000,030	
0.0	3,878	P00364991	10	45,000	3			49	9	314,074,269	83	54,001,220	
		P01213219	10	49,860	3			53	9	83,901,809	65	55,002,707	
		P00450344	9					65	9	255,076,753	52	55,003,201	
		P00505261	9					113	9	96,017,418	38	55,003,308	
		P00733147	9					328	9	255,077,370	38	56,073,573	

Output 2b – In House Enhancements



Output 2c – In House Enhancements



Annual Results Comparison
Legacy Scoring vs. Updated Scoring

Processing System:

- 24K Returns ID'ed
- 10.53% Positive ID
- $5M Refund Fraud

Data Warehouse:

- 33K Returns ID'ed
- 55.31% Positive ID
- $24M Refund Fraud

Annual Results Comparison
The BIG Picture

Legacy QRDT Processes

- 110K Reviewed
- ~ 10% Positive ID
- $23M Refund Fraud

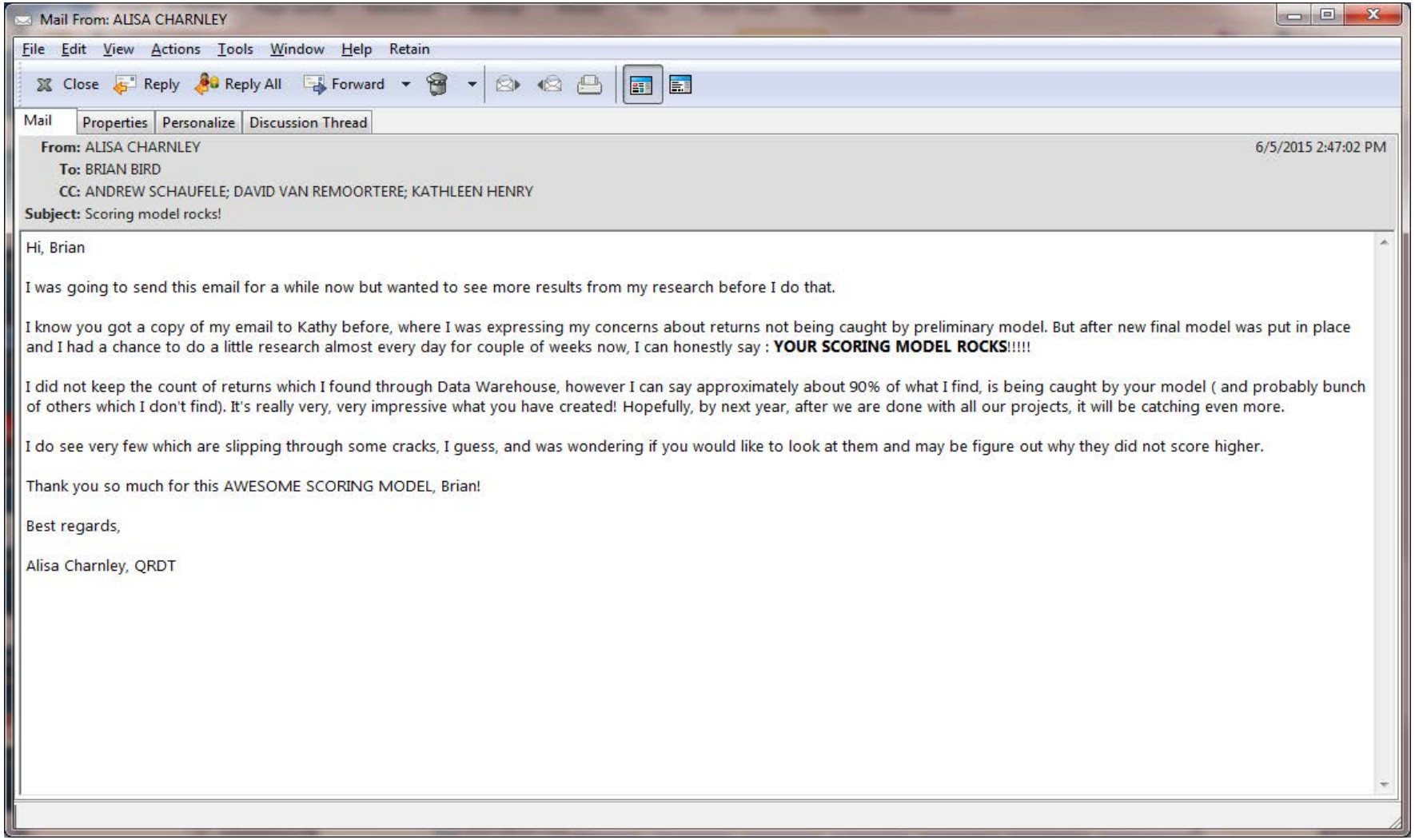
Data Warehouse:

- 33K Returns ID'ed
- 55.31% Positive ID
- $24M Refund Fraud

QRDT Reaction

- Very Dedicated and **Passionate** Group
- Initial Reaction: Excitement with Apprehension
- “Shut Down” several legacy stand alone programs
  - Initial concern about letting some fraud through
  - With coaching, understood that yes, the model isn’t perfect, but with it you will stop more fraud!!!!

QRDT Reaction



Mail From: ALISA CHARNLEY

File Edit View Actions Tools Window Help Retain

Close Reply Reply All Forward

Mail | Properties | Personalize | Discussion Thread

From: ALISA CHARNLEY 6/5/2015 2:47:02 PM
To: BRIAN BIRD
CC: ANDREW SCHAUFELE; DAVID VAN REMOORTERE; KATHLEEN HENRY
Subject: Scoring model rocks!

Hi, Brian

I was going to send this email for a while now but wanted to see more results from my research before I do that.

I know you got a copy of my email to Kathy before, where I was expressing my concerns about returns not being caught by preliminary model. But after new final model was put in place and I had a chance to do a little research almost every day for couple of weeks now, I can honestly say : **YOUR SCORING MODEL ROCKS**!!!!!

I did not keep the count of returns which I found through Data Warehouse, however I can say approximately about 90% of what I find, is being caught by your model (and probably bunch of others which I don't find). It's really very, very impressive what you have created! Hopefully, by next year, after we are done with all our projects, it will be catching even more.

I do see very few which are slipping through some cracks, I guess, and was wondering if you would like to look at them and may be figure out why they did not score higher.

Thank you so much for this AWESOME SCORING MODEL, Brian!

Best regards,

Alisa Charnley, QRDT

Data Warehouse Enhancements

- Use all fields from return in scoring model
- More advanced statistical procedures
  - Decision tree
- Data in same location as scoring allows for more dynamic model
  - Historical filing comparisons
  - Adapt for fraud identified during year

Other Positive Contributing Factors

- Experience of Contractors/State Personnel
  - Data Sources
  - Desired End Product
  - Processing System Constraints

Takeaway Points

- Fraud Analytics in DW can be Successful
- Success of one Single Project Uncertain
  - Final Product Characteristics
  - Data/Organization Differs from State to State
- Need Consistent Support from Management
- Lessons Learned from each Project Transferable to Subsequent Projects

Contact Info

- Andrew Schaufele
- Comptroller of Maryland
  - Director, Bureau of Revenue Estimates
  - 410.260.7450
  - aschaufele@comp.state.md.us