How-to leverage Data Analytics to support Audit or Investigation Services

CHARLES COE

ASSISTANT INSPECTOR GENERAL, INFORMATION TECHNOLOGY AUDITS AND COMPUTER CRIME INVESTIGATIONS

U.S. DEPARTMENT OF EDUCATION
OFFICE OF INSPECTOR GENERAL

FEDERAL AUDIT EXECUTIVE COUNCIL
BI-MONTHLY MEETING AND TRAINING
AUGUST 22, 2012

Today's Agenda

- Part I – Foundation Cornerstones
  - Vision & Commitment
  - Strategy
  - Project management
- Part II – Data/Predictive Analytics & Risk Models
  - What is & What isn't
  - State & Local Risk Model
  - E-Fraud Risk Model

Part I - Foundation Cornerstones

- Vision and Commitment
- Strategy
- Project Management

Vision and Commitment

- Leadership is an essential ingredient
  - executive champion/sponsor who has an agreed upon vision of the value and direction of the implementation of a data analytic project
  - DON'T start without it!
- Set expectations and secure resources
  - don't expect to get a Cadillac on VW budget
  - no free lunch

Strategy

- Benchmark other organizations
  - who have been successful
  - do NOT forge your own path unless absolutely necessary!
- Take advantage of “Lessons Learned” from benchmark organizations
- Research legal & IT security requirements (SORN, CMA, C&A, etc …)
- Determine skill sets
  - needed “over time”
  - establish effective interview and selection methodology – no, you can’t fudge this one!

Project Management

- Identify your customer requirements & needs (Audit & Investigations)
  - Audit – traditional risk models reflecting how best to allocate limited audit resources that focus on critical operational areas
  - Investigations – data analytical models looking for known fraudulent patterns within key operational areas that have the highest return on investment of investigation resources

Project Management - continued

- Seek program experts to help develop your data analytics projects
- False positives are your enemy – these can be fatal!
- Deliver on time and don't miss milestones, otherwise you jeopardize losing your "Champion" supporters
- Be prepared for unintended consequences/results of creating risk models!

Part II - Predictive Analytics & Risk Models

20th Century - Standard Audit Practice

- Audit occurs significantly after transactions are completed.
- Rarely able to test all transactions in comprehensive fashion. Normal practice is using statistical sampling techniques.
- Therefore there may have been significant risk that errors could have occurred, but remain undetected.

21st Century - "***Predictive Analytics***" Techniques

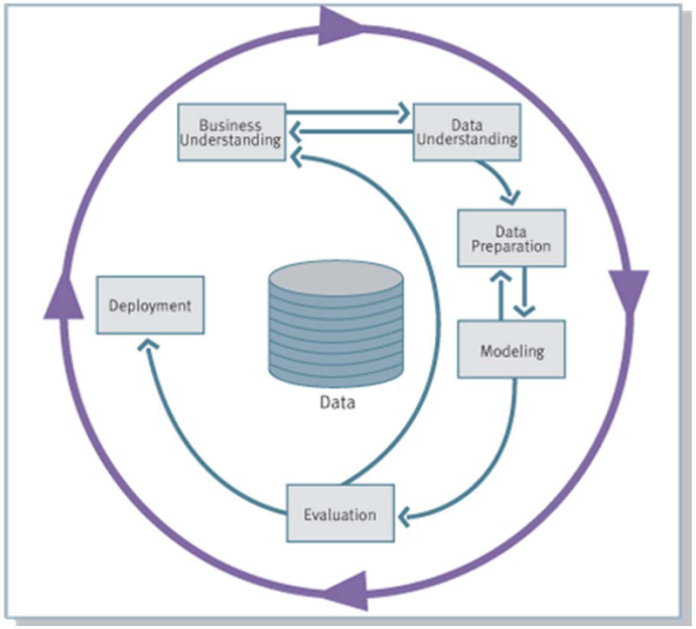
- Empowers Investigators and Auditors to leverage today's technology to predict with a high degree of probability, anomalies where fraudulent or inaccurate activity is likely using statistical and mathematical techniques.
- Makes the audit and investigative processes more efficient and effective.
- Ability to discover both **fraudulent anomalies** as well as **indicators of control deficiencies and emerging risk**.

What is NOT Data Mining?

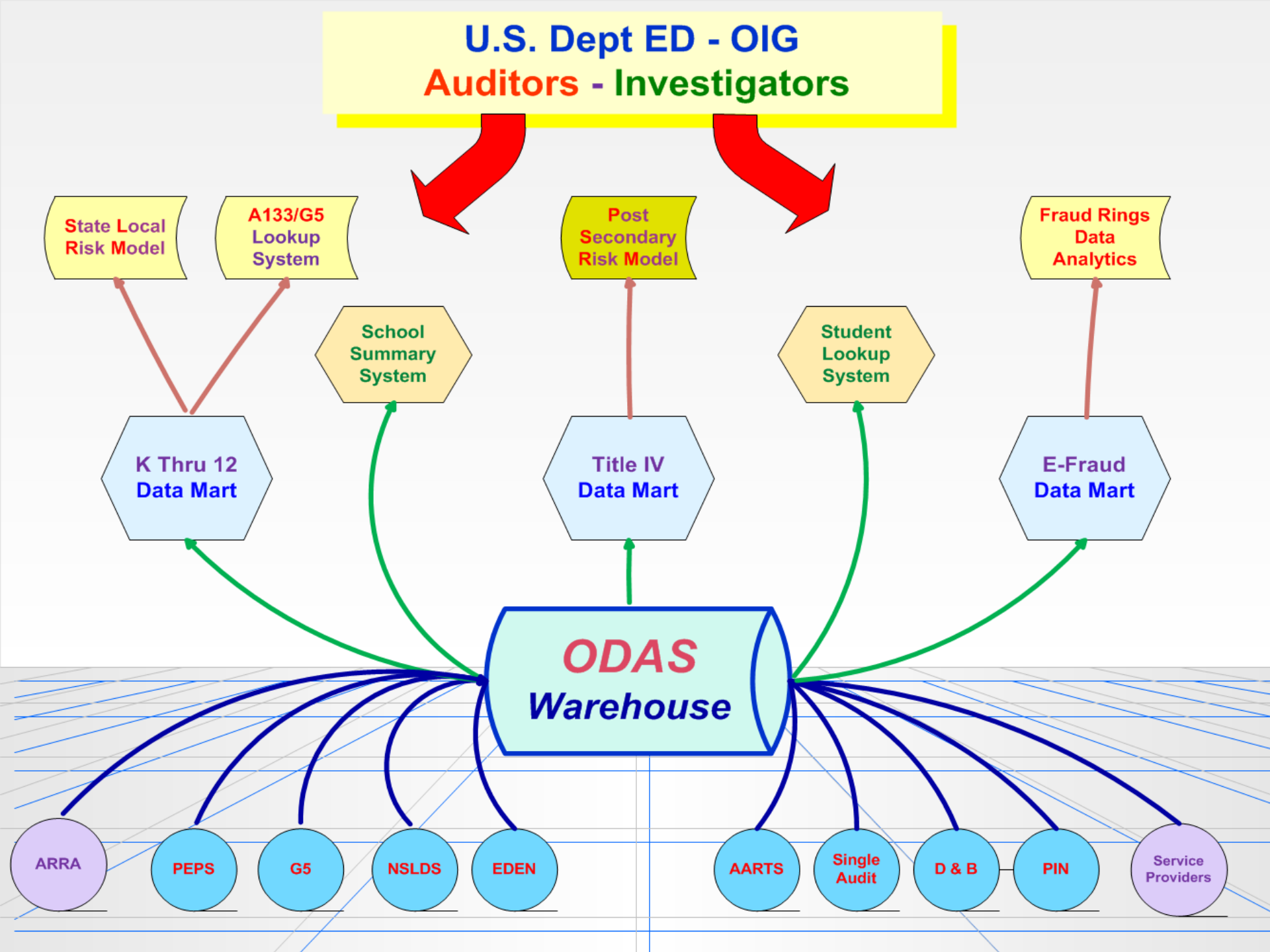
1. **Data Matching**
   - Do any of our current contractors match those on the debarred/excluded party list?
2. **Database Queries**
   - How many beneficiaries in our program are over 100 years old?
3. **Slicing & Dicing Data in Excel Spreadsheets**
   - Which contract has the highest dollar value?
4. **Visualization**
   - Who is connected to the suspicious contractor?

What IS Data Mining?

- Data Mining: Discovering **patterns in past data** that can be used to **predict the outcome of future cases.**
- Build **predictive models** with valuable business knowledge from SMEs
- Allows the **computer** to find the patterns and anomalies that humans are not able to find



Industry-Standard CRISP-DM Process


U.S. Dept ED - OIG
Auditors - Investigators
State Local Risk Model
A133/G5 Lookup System
Post Secondary Risk Model
Fraud Rings Data Analytics
School Summary System
Student Lookup System
K Thru 12 Data Mart
Title IV Data Mart
E-Fraud Data Mart
ODAS Warehouse
ARRA
PEPS
G5
NSLDS
EDEN
AARTS
Single Audit
D & B
PIN
Service Providers

State & Local Education Agencies Risk Model (SLRM)

- Identify those State and Local Educational Agencies at the highest to lowest level of risk.

- The SLRM risk model will provide audit and investigation management with a continuous auditing[analysis] functionality thereby enhancing audit planning and investigation resource management.

Methodology of SLRM

- To assemble similar size Local Education Agencies (LEA) into groups and rank them based on weighted scores assigned to selected risk factors.
  - Groups, Risk Factors, Scores, and Weights were agreed to and determined by the SLRM Project Team
    - LEAs split into six groups based on student population.
    - Risk Factors from five primary sources of data.
    - Risk factor data transformed into scores ranging from zero to 100.
    - Scores weighted by multiplying by 1, 2, 3, 4, or 5.
  - Ranked on a scoring system within each group
    - Highest score represents the highest risk LEA in group
    - Highest possible score is 2700 points in each group
- To rank States by combining group rankings of LEAs together for each state.

Risk Factors were Derived From Five Primary Sources of Data

- **NCES –** [Performance] *National Center for Education Statistics*
- **G5 –** [Administrative] ED Grant System
- **Dun and Bradstreet –** [Financial] Financial risks such as Federal Debt Indicator, Payment history, Debarment…
- **ARRA -**Funds Received that was not reported or exceeds Sub-award Amount…
- **A-133 Single Audit -** [Audit] Federal Audit Clearinghouse

Grouping of LEAs

LEA Groups	LEA Type	LEA Count	Student Count
1	Non-Charter	27	**>=100,000**
2	Non-Charter	361	**>=20,000** <100,000
3	Non-Charter	499	**>=10,000** <20,000
4	Non-Charter	3,831	**>=2,000** <10,000
5	Non-Charter	2,542	**>=1,000** <2,000
6	Charter	2,356	N/A
Rest	Non-Charter	8,790	<1,000

Risk Factors and Highest Possible Scores

- Risk Factors and highest possible score for each are broken down as follows:
  - Administrative – 500 points
  - Financial – 500 points
  - Single A-133 Audit – 300 points
  - ARRA – 300 points
  - Met Adequate Yearly Progress – 300 points
  - Charter Schools – 300 points
  - Dropouts – 200 points
  - Graduation Rates – 200 points
  - Discipline Incidents – 100 points

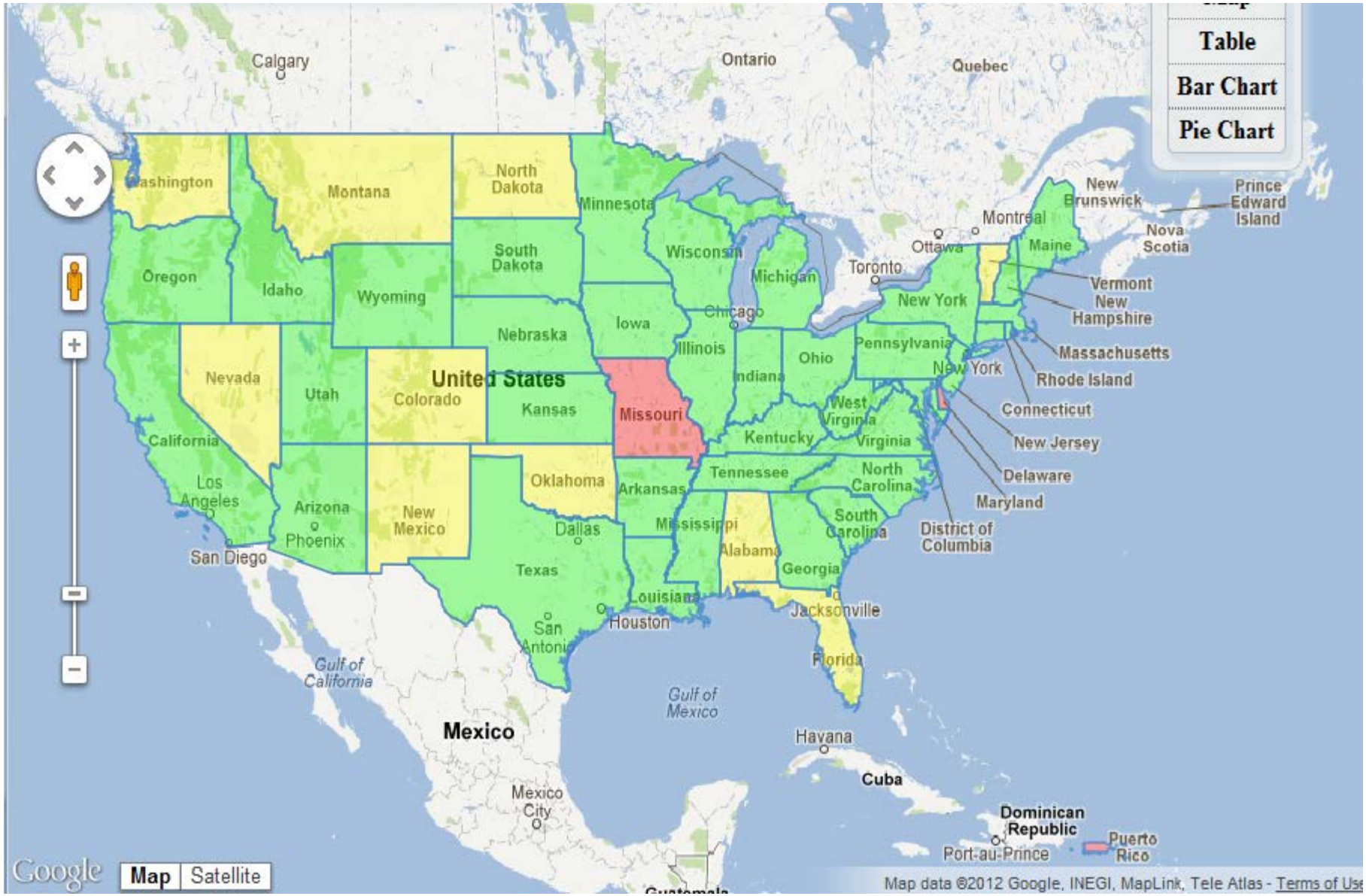
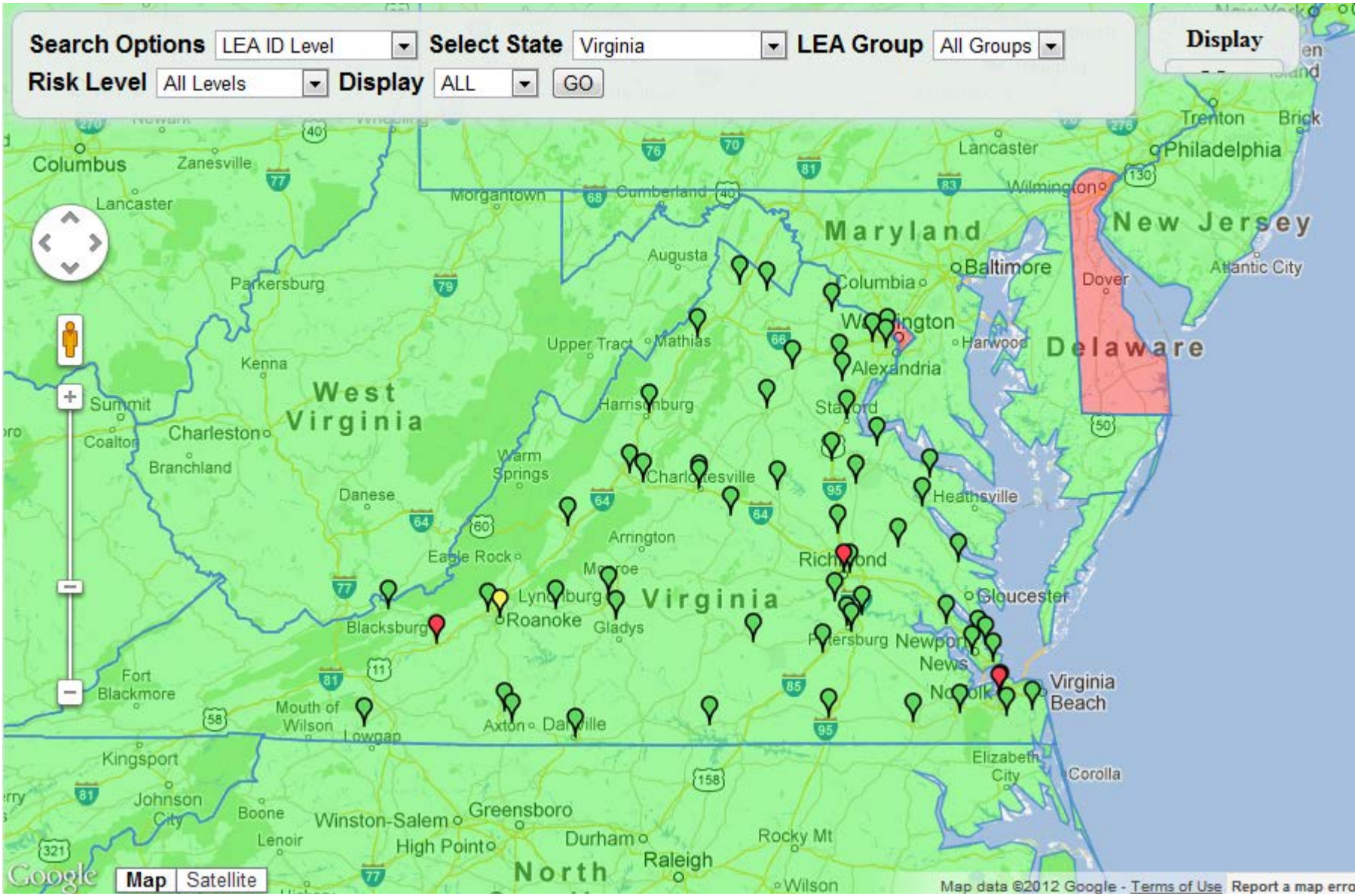
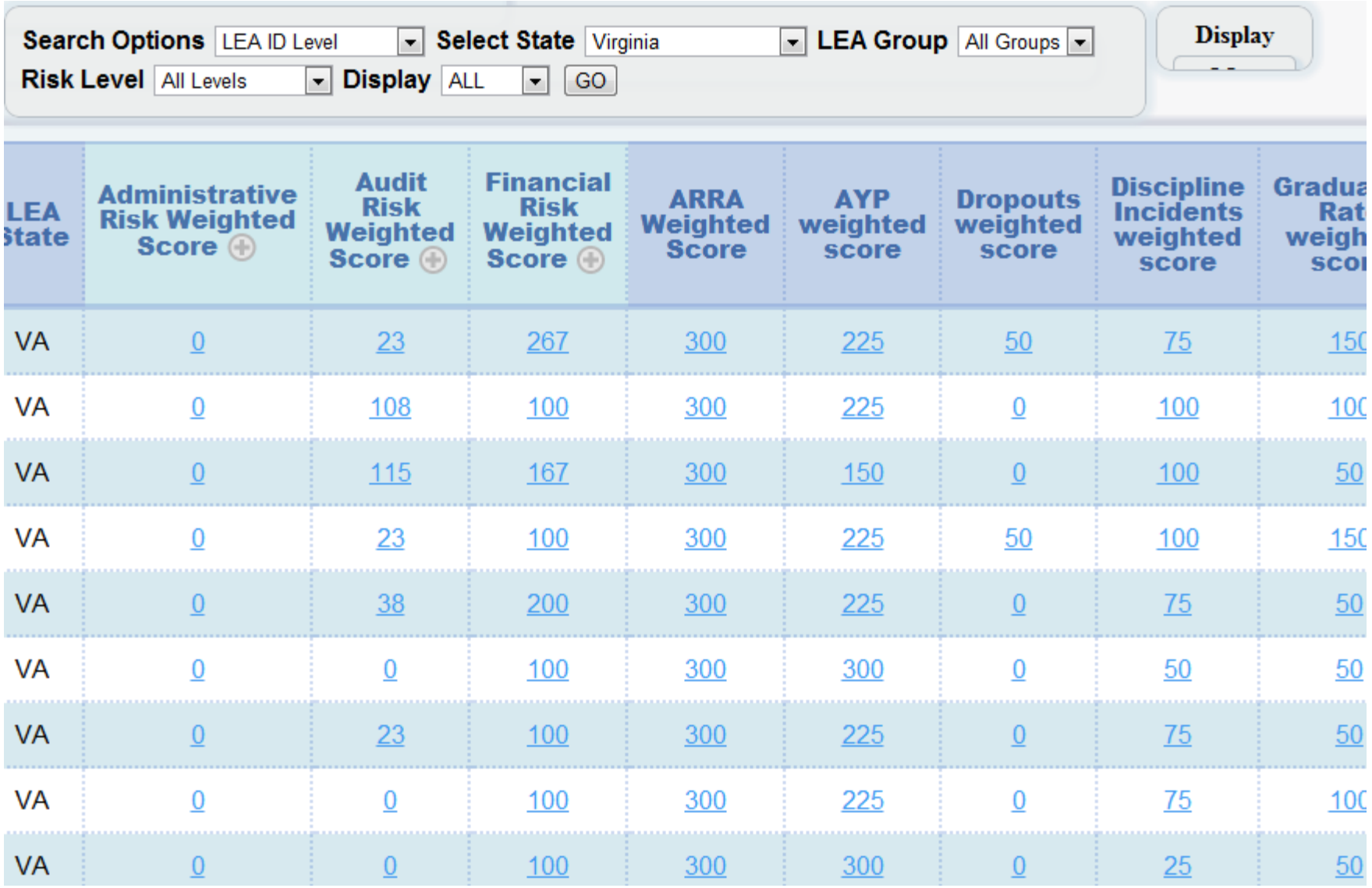


Table
Bar Chart
Pie Chart
Calgary
Ontario
Quebec
Washington
Montana
North Dakota
Minnesota
Oregon
Idaho
Wyoming
South Dakota
Wisconsin
Michigan
Nebraska
Iowa
Chicago
Illinois
Nevada
Utah
Colorado
Kansas
Missouri
Indiana
Ohio
Pennsylvania
New York
California
Los Angeles
San Diego
Arizona
Phoenix
New Mexico
Oklahoma
Dallas
Texas
Arkansas
Tennessee
Kentucky
West Virginia
Virginia
North Carolina
South Carolina
Mississippi
Alabama
Georgia
Louisiana
Houston
San Antonio
Florida
Jacksonville
Toronto
Ottawa
Montreal
New Brunswick
Maine
Nova Scotia
Prince Edward Island
Vermont
New Hampshire
Massachusetts
Rhode Island
Connecticut
New Jersey
Delaware
Maryland
District of Columbia
Gulf of California
Gulf of Mexico
Mexico
Mexico City
Havana
Cuba
Dominican Republic
Port-au-Prince
Puerto Rico
Google
Map
Satellite
Map data ©2012 Google, INEGI, MapLink, Tele Atlas - Terms of Use


Search Options
LEA ID Level
Select State
Virginia
LEA Group
All Groups
Risk Level
All Levels
Display
ALL
GO
Display
Maryland
New Jersey
Delaware
West Virginia
Virginia
North
Columbus
Zanesville
Lancaster
Parkersburg
Morgantown
Cumberland
Augusta
Baltimore
Columbia
Dover
Atlantic City
Trenton
Philadelphia
Wilmington
Harwood
Alexandria
Upper Tract
Mathias
Harrisonburg
Kenna
Summit
Coalton
Charleston
Branchland
Danese
Warm Springs
Charlottesville
Heathsville
Arrington
Eagle Rock
Monroe
Richmond
Lynchburg
Roanoke
Gladys
Blacksburg
Gloucester
Newport News
Norfolk
Virginia Beach
Fort Blackmore
Mouth of Wilson
Lowgap
Axton
Danville
Kingsport
Johnson City
Boone
Winston-Salem
Greensboro
Durham
High Point
Raleigh
Rocky Mt
Wilson
Elizabeth City
Corolla
Google
Map
Satellite
Map data ©2012 Google - Terms of Use
Report a map erro



LEA State	Administrative Risk Weighted Score	Audit Risk Weighted Score	Financial Risk Weighted Score	ARRA Weighted Score	AYP weighted score	Dropouts weighted score	Discipline Incidents weighted score	Gradua Rat weigh scor
VA	0	23	267	300	225	50	75	150
VA	0	108	100	300	225	0	100	100
VA	0	115	167	300	150	0	100	50
VA	0	23	100	300	225	50	100	150
VA	0	38	200	300	225	0	75	50
VA	0	0	100	300	300	0	50	50
VA	0	23	100	300	225	0	75	50
VA	0	0	100	300	225	0	75	100
VA	0	0	100	300	300	0	25	50

E-Fraud Data Analytical Model

- Student fraud rings have become a rapidly growing crime activity that now have targeted the U.S. Department of Education (ED) FSA programs.

- ED processed over 19 million applications for student financial aid and disbursed over $90 Billion in FSA funds in SY2010.

Record Filtering process

- Initial data analysis showed many false positives and an overwhelming number of records.
- Needed to develop a process to limit the records yet keep the riskiest ones.
- Conducted an assessment of the data again focusing on what we had learned from the IS case data and identified three key indicators:
  - [redacted]
  - [redacted]
  - [redacted]
- Determined these indicators as the primary filtering mechanism.

Fine Tuning of Data

- Developed a set of Risk Weighting factors ranging from 3 to 0.
- Purpose to be able rank from highest to lowest ranking of identified fraud groupings.
- Enhanced the risk scoring mechanism to identify and omit scoring on known frequency anomalies relating to [redacted] and certain [redacted].

Student Fraud Ring Filter Results

15 Post Secondary Schools selected as part of the Proof of Concept Project.

School Code	Total Student Population	Filtered Student Population
999999	21,035	**89**
999999	4,171	**26**
999999	12,020	**183**
999999	65,457	**129**
999999	19,193	**8**
999999	5,566	**4**
999999	5,332	**5**
999999	3,572	**28**
999999	44,130	**366**
999999	11,727	**51**
999999	6,106	**11**
999999	7,787	**86**
999999	40,441	**100**
999999	1,701	**19**
999999	48,598	**62**

Is Your Model Statistically Supportable?

- Verify that your sample data being used in the proof of concept project is representative of the total population.
- Address concerns of bias by modifying established fraud indicator parameters looking for abnormal/unexpected variances.
- Use data outside of the sample but from the total population to reaffirm expected outcomes.
- Bringing a statistician onboard from the outset of the project, or at a minimum assess the planned project methodology is recommended.

Feedback from the Field

- Conclusion model had identified all known fraud rings from SY2010 test set. Statistically this is very rare, which further gives us a sense of the value generated.
- Identified new previously unknown fraud rings.
- Added additional students to fraud rings under investigation.

Audit Independence Concerns?

- Tell story here of unintended consequences ….

Build risk models as part of a performance audit!

The End

Questions?