

Needles in a Haystack

Data Mining and Predictive Analytics
to Prioritize Leads and Highlight Risk

Association of Inspectors General
May 19, 2011

Elder Research, Inc.
300 West Main Street, Suite 301
Charlottesville, Virginia 22903
434-973-7673
www.datamininglab.com

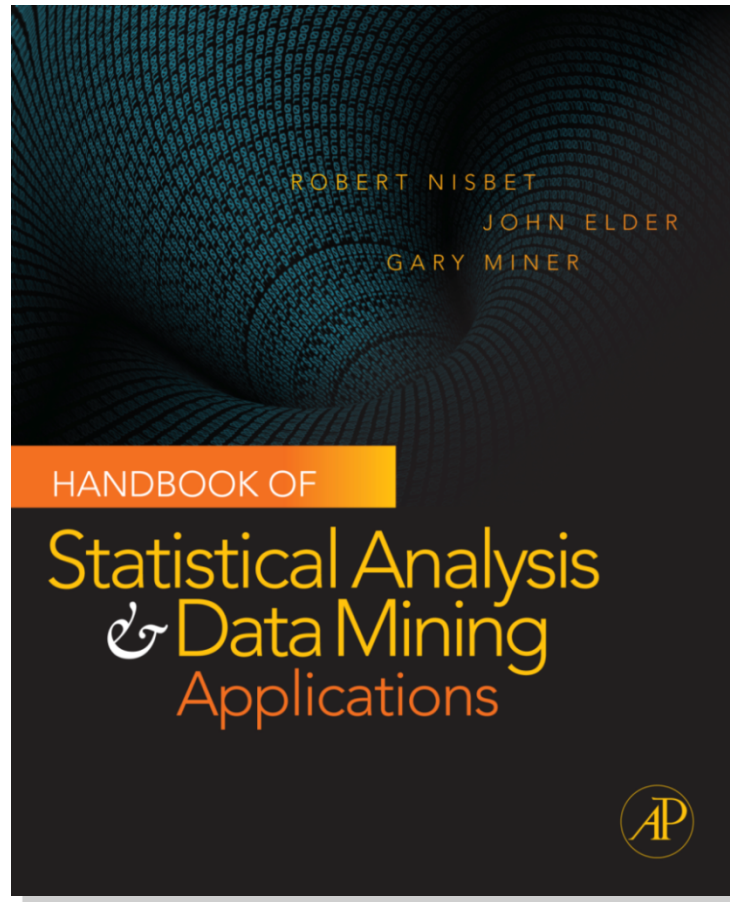
John Elder, PhD
elder@datamininglab.com

A brief introduction to Elder Research, Inc.

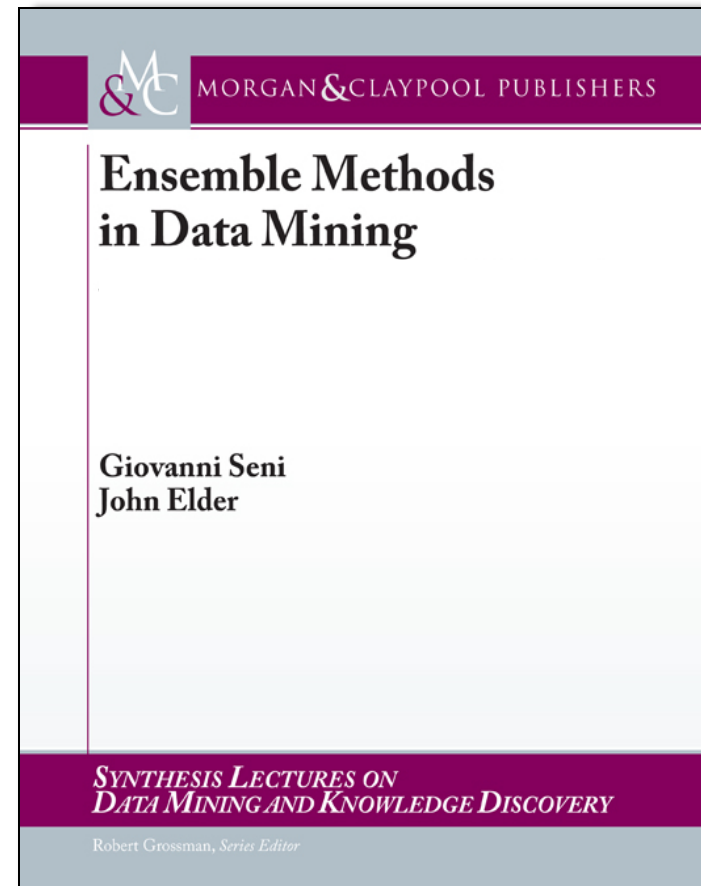
- Largest, most experienced consultancy in Data Mining and Predictive Analytics (Founded in 1995, ~20 analysts)
- Experts in predictive modeling, fraud detection, link analysis, and text mining
- World's top lab of COTS and custom analytic software
- Researchers hold patents, PhDs, and/or TS clearances
- Federal clients include DoD, DHS, SSA, Treasury, NGIC, DFAS, DoE, Army, Navy, Federal OIG
- Commercial clients include HP, Anheuser-Busch, Capital One, HSBC, Oppenheimer Funds, AstraZeneca, Georgetown University, Pfizer

Books

Book written *for* practitioners
by practitioners (May 2009)



How to combine models for
improved predictions (Feb 2010)



(Won 2009 PROSE Award for Mathematics)

The 9 Levels of Analytics (part 1 of 2)

Descriptive Techniques (1-6):

1 – Standard Reporting

“How much did we sell last quarter?”

2 – Custom Reporting or “Slicing and Dicing” the Data (Excel)

“How many investigations did we perform in each state last year?”

3 – Queries/drilldowns (SQL, OLAP)

“Which contractors received >\$10 million in sole-source contracts last year?”

4 – Dashboards/alerts (Business Intelligence)

“In what sectors have customer complaints grown since last quarter?”

5 – Statistical Analysis

“Is frequency of customer communication correlated with satisfaction?”

6 – Clustering (Unsupervised Learning)

“How many fundamentally different types of behaviors are in the data and what do they generally look like?”

The 9 Levels of Analytics (part 2 of 2)

Predictive Techniques (7-9):

7 – Predictive Modeling

“Which contracts are most likely to be fraudulent?”

8 – Optimization & Simulation

“What number of investigators would we put on each case to maximize expected return?”

9 – Next Generation Analytics – Text Mining & Link Analysis

“Do the transactions reveal a coordinated set of people likely to be a fraud ring?”

Case Study (level 7, predictive model): *Internal Revenue Service*

The Problem:

- Limited staff time to review tax returns most likely to be fraudulent
- Fraud can take many different forms, and changes often

The Solution:

- ERI led data mining technical team to build Electronic Fraud Detection System to find tax return fraud
- Built fraud detection models that increased hit rate by a **factor of 25**
- Pilot implementation was so successful that roll-out to other sites was accelerated by a year

Case Study (level 6, metrics + scoring): *A Large Federal OIG*

The Problem:

- Federal agency managed \$33 billion in postal contracts in FY2009
- Fraud can occur at any stage of a contract's lifecycle, from pre-solicitation to award and performance
- Fraud can take many different forms, including kickbacks, collusion among bidders, and overcharging

The Solution:

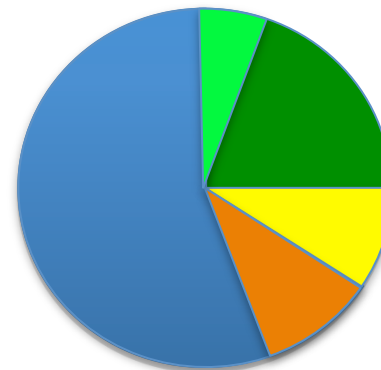
- Integrated data from disparate data sources
- Developed 30+ fraud indicators to produce high-quality leads
- Investigators **confirmed that the leads are promising**; they overlap well with identified fraudsters
- Now, ERI is building a **customized contract fraud discovery tool** to enhance investigative productivity

Case Study (levels 7 & 9, prediction + text): *Social Security Disability Approval*



- Pain: Approval process is long, bureaucratic

**Up to
2 Years !**



With Text Mining,
1/5 of cases approved
immediately!

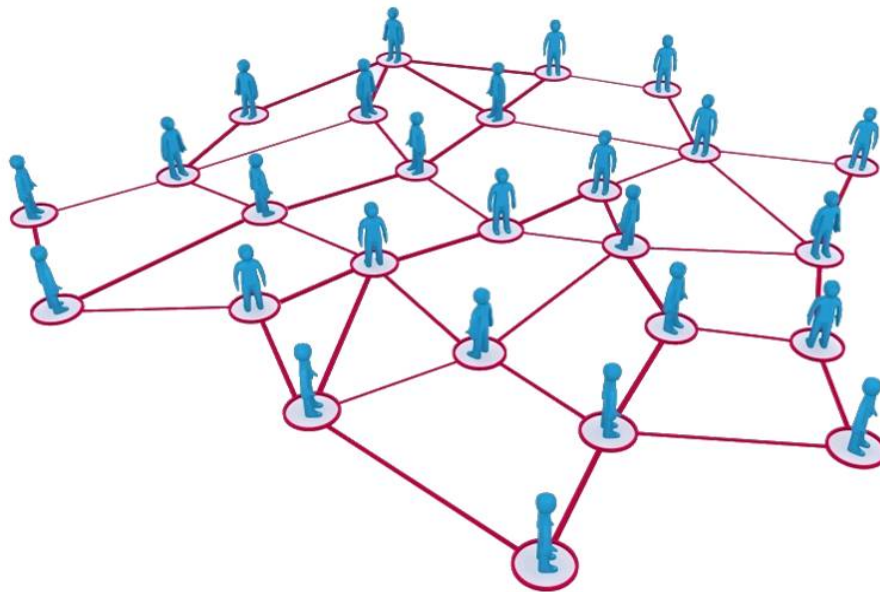
1/3 of cases
eventually approved

1/2 of appeals overturn
original decision

- Goal: Fast-track “easy” cases
- Challenge: Free-text on disability application
- Result: 20% of Approvals possible immediately and with greater consistency

Complex Network Analysis

- “Social Network Analysis” (SNA) or “Link Analysis”
- Can find critical relationships (links), or key agents (nodes)
- Uses multiple kinds of relationships between individuals to visualize networks
- *Predictive Link Analysis*
 - build models on networks to, for example, find fraud “rings”



Birds of a feather flock together: fraud status is relatively contagious among related individuals

Case Study (level 9, links)

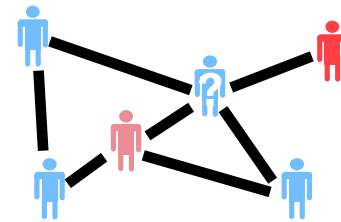
Securities Fraud Ring Detection



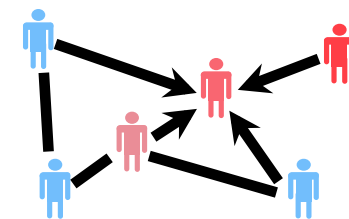
- Cluster individuals based on firm, branch, and geography
- Using network information boosted predictive performance 10-20%
- Model and experts performed roughly equivalently – but on different groups.
- Model **augments, but cannot replace**, expert analysis



Use intrinsic information about an individual in question



Add information about related individuals



Make a prediction about individual in question

How to Manage Data/Text Mining Projects

- Assess data assets (what treasure could be hidden in our sludge?)
 - Data caretaker must be on-board
- Identify pain points in current production process
 - What improvements would make the biggest impact?
- Brainstorm ideal process
 - External expertise is extremely useful here
- Conduct a pilot project. Simultaneously:
 - “Hit a single”: e.g., automate key task, create dashboard
 - “Swing for fences”: attack core weakness
- Have key staff work closely with analytic experts
 - Transfer technology to inside
 - Internalize essential steps
- Prove ROI. Make allies and decision-makers look good!



John F. Elder IV Chief Scientist, Elder Research, Inc.

DR. JOHN ELDER HEADS A DATA MINING CONSULTING TEAM WITH OFFICES IN CHARLOTTESVILLE VIRGINIA AND WASHINGTON DC (WWW.DATAMININGLAB.COM). FOUNDED IN 1995, ELDER RESEARCH, INC. FOCUSES ON INVESTMENT, COMMERCIAL AND SECURITY APPLICATIONS OF ADVANCED ANALYTICS, INCLUDING TEXT MINING, STOCK SELECTION, IMAGE RECOGNITION, PROCESS OPTIMIZATION, CROSS-SELLING, BIOMETRICS, DRUG EFFICACY, CREDIT SCORING, MARKET TIMING, AND FRAUD DETECTION.

JOHN OBTAINED A BS AND MEE IN ELECTRICAL ENGINEERING FROM RICE UNIVERSITY, AND A PHD IN SYSTEMS ENGINEERING FROM THE UNIVERSITY OF VIRGINIA, WHERE HE'S AN ADJUNCT PROFESSOR TEACHING OPTIMIZATION OR DATA MINING. PRIOR TO 16 YEARS AT ERI, HE SPENT 5 YEARS IN AEROSPACE DEFENSE CONSULTING, 4 HEADING RESEARCH AT AN INVESTMENT MANAGEMENT FIRM, AND 2 IN RICE'S *COMPUTATIONAL & APPLIED MATHEMATICS* DEPARTMENT.

DR. ELDER HAS AUTHORED INNOVATIVE DATA MINING TOOLS, IS A FREQUENT KEYNOTE SPEAKER, AND WAS CO-CHAIR OF THE 2009 *KNOWLEDGE DISCOVERY AND DATA MINING* CONFERENCE, IN PARIS. DR. ELDER WAS HONORED TO SERVE FOR 5 YEARS ON A PANEL APPOINTED BY THE PRESIDENT TO GUIDE TECHNOLOGY FOR NATIONAL SECURITY. HIS BOOK WITH BOB NISBET AND GARY MINER, *HANDBOOK OF STATISTICAL ANALYSIS & DATA MINING APPLICATIONS*, WON THE PROSE AWARD FOR MATHEMATICS IN 2009. HIS BOOK WITH GIOVANNI SENI, *ENSEMBLE METHODS IN DATA MINING: IMPROVING ACCURACY THROUGH COMBINING PREDICTIONS*, WAS PUBLISHED IN FEBRUARY 2010.