The High ROI of Data Mining for Innovative Organizations

UVA CMIT
Chantilly, VA
July 9, 2010

John F. Elder IV, Ph.D.
elder@datamininglab.com

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

QuickTime™ and a decompressor are needed to see this picture.

Case Study Lessons Outline

- The Hype Cycle vs. Real Results
- 3 Major ways Data Mining helps:
  - Eliminate the bad:
    - Examples from IRS, Hewlett-Packard, Capital One
  - Discover the good:
    - Ex: Pfizer, Westwind Foundation
  - Streamline / Automate:
    - Ex: Lumidigm, Peregrine, SSA, Anheuser-Busch
- Lessons Learned:
  features of successful projects

QuickTime™ and a decompressor are needed to see this picture.

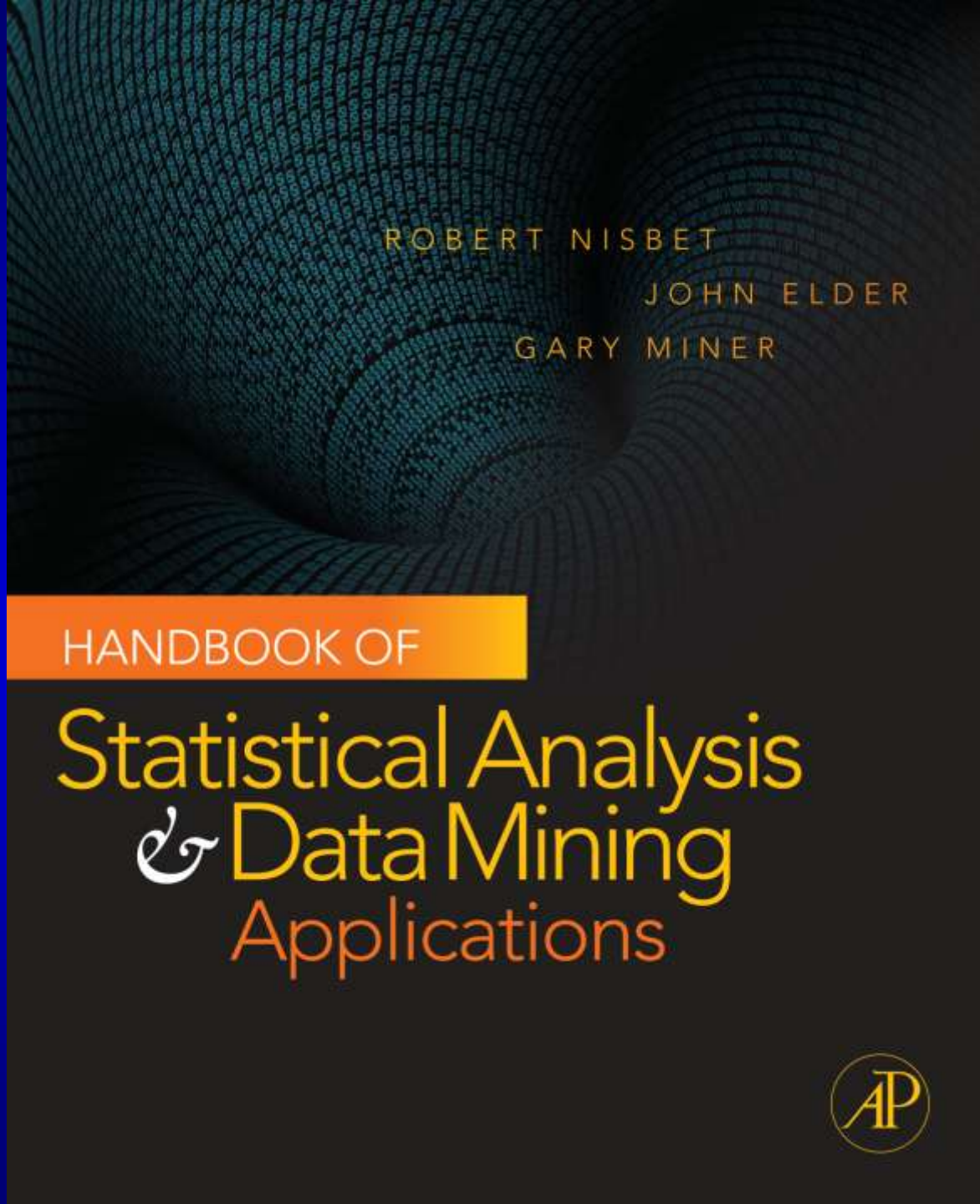
Book written *for* practitioners, *by* practitioners

Published by Elsevier's Academic Press in May 2009

Details and reviews at Amazon: www.tinyurl.com/bookERI

Won 2009 PROSE Award for Mathematics

QuickTime™ and a decompressor are needed to see this picture.



"Of course machines can think. After all, humans are just machines made of meat."
- MIT CS professor

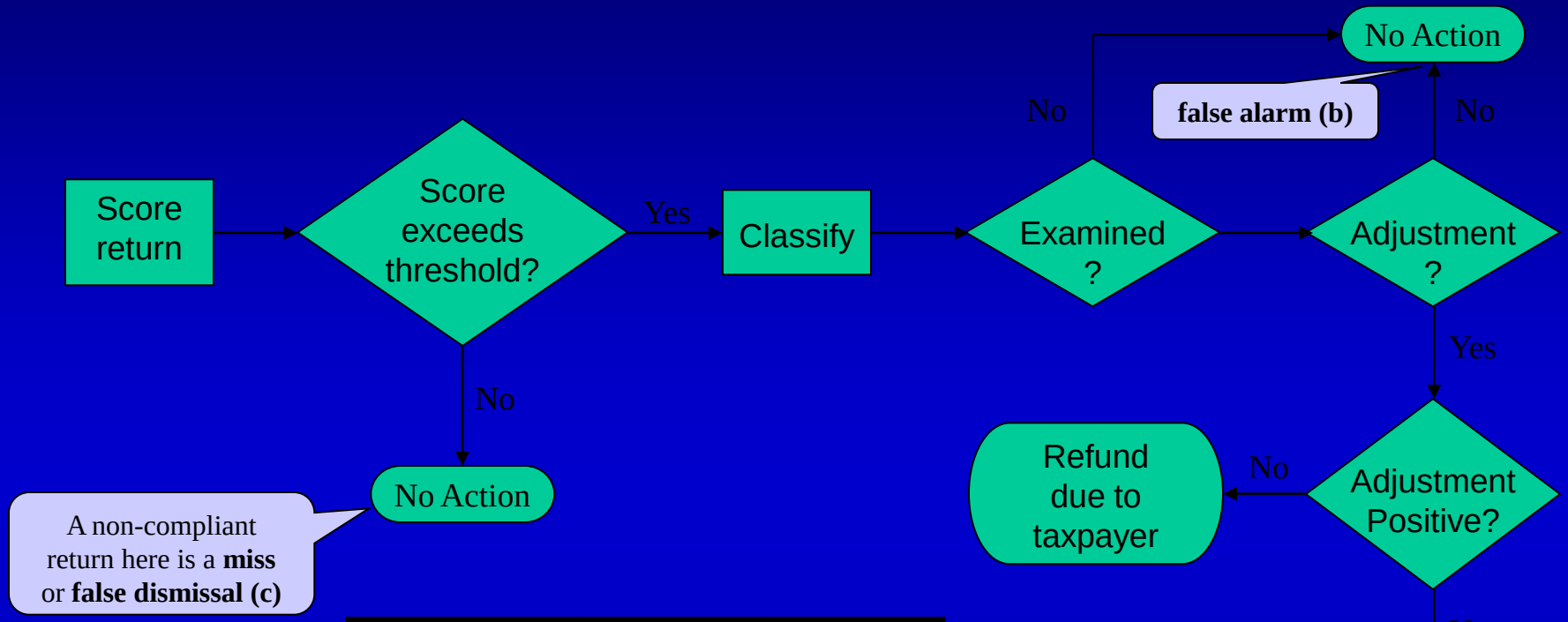
Human and computer strengths are more complementary than alike.

3 Major Ways Data Mining Helps

- Eliminate the Bad
  - Score credit risk, Detect fraud, Find outliers
- Highlight the Good
  - Time market trades, Discover new drugs, Uncover hidden value
- Streamline or Semi-Automate Decisions
  - Intuit products of interest, Assess status, Verify identity, Be aware of “what if” scenarios, Speed service

QuickTime™ and a decompressor are needed to see this picture.

Case 1) IRS Fraud Detection



Truth \ Prediction	Not Flagged	Flagged
Compliant	a	b
Non-Compliant	c	d

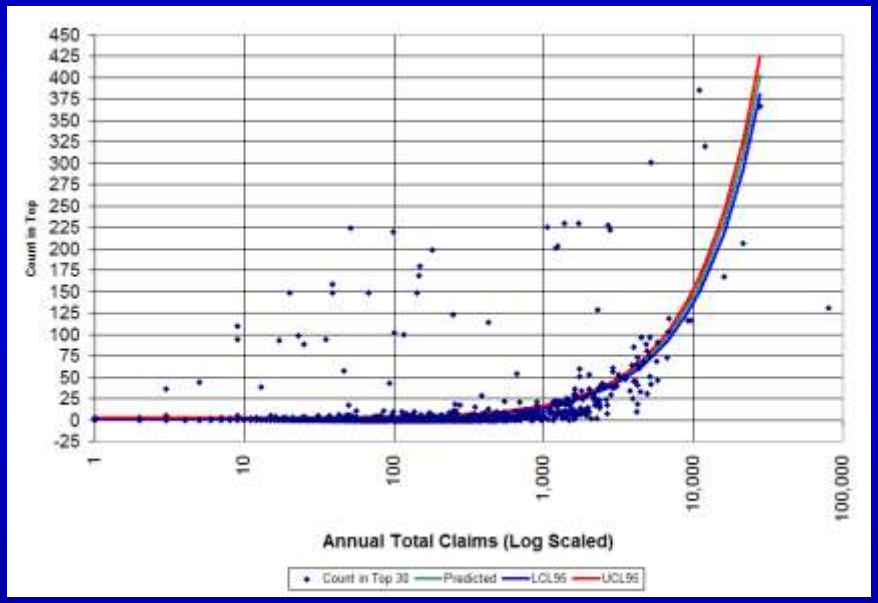
detection rate = d/(c+d)

workload = b+d

hit:scan = 1 : (b+d)/d

QuickTime™ and a decompressor are needed to see this picture.

Case 2) Hewlett-Packard: Service Fraud Detection

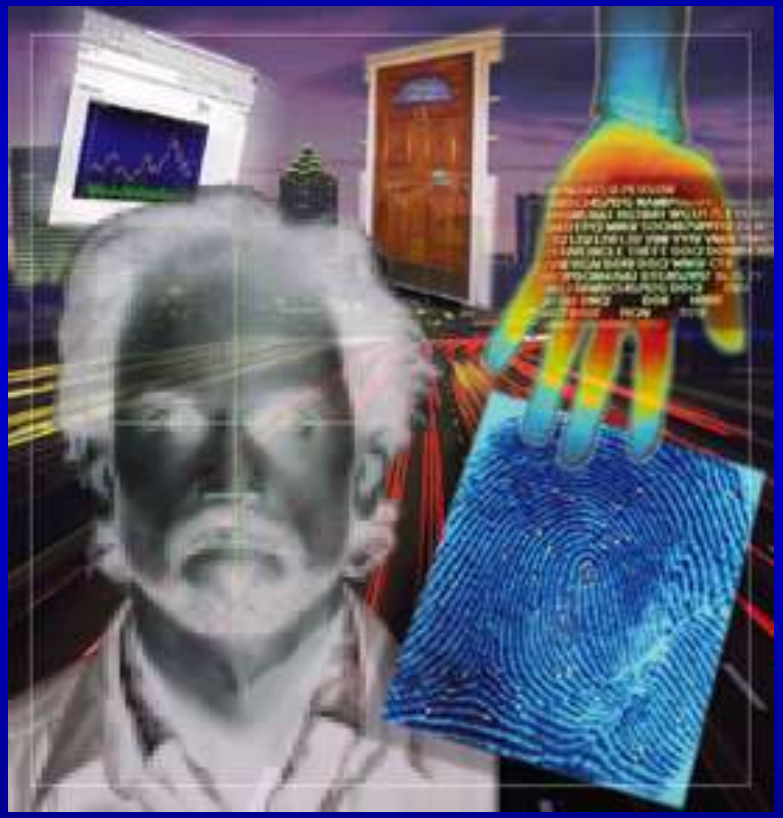


- Tips indicated fraud exists
- Goal: Learn from known cases to find unknown
- Automate current process, build model on known, score all data, investigate top
- Recovered $20M in 9 months
- Awards + promotions + growth Became profit center

QuickTime™ and a decompressor are needed to see this picture.

Case 3) Lumidigm: Bio-Metrics

(id not by what you have or know, but what you are)



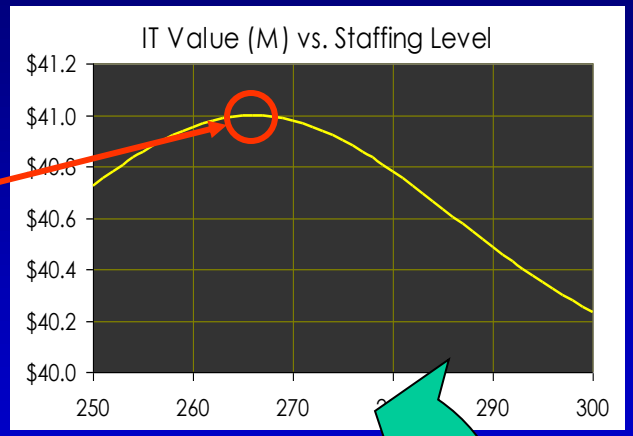
- Original Goal:
  Diagnose disease
- Data: infra-red skin reflections at multiple frequencies
- Problem: hard to remove person specifics
- Idea: Turn problem into gain: new bio-metric
- Q: *Who* are you?
- Use to validate ID for safety or convenience

Solution in search of the right Application

50000
45000
40000
35000
30000
25000
20000
15000
10000
5000
0

QuickTime™ and a
decompressor
are needed to see this picture.

Case 4) Peregrine Systems: Business Service Modeling
Or "Sim-City" for IT



Optimal Operating Point: 266 Staff


Environment Choices

Environment Facts/Constraints


Staffing Choices

Incident Distribution

Probability Distributions

IT Cost

IT Value

Service Simulation

IT Risk


Service Objectives

Resolution Distribution

Business Service Relationships

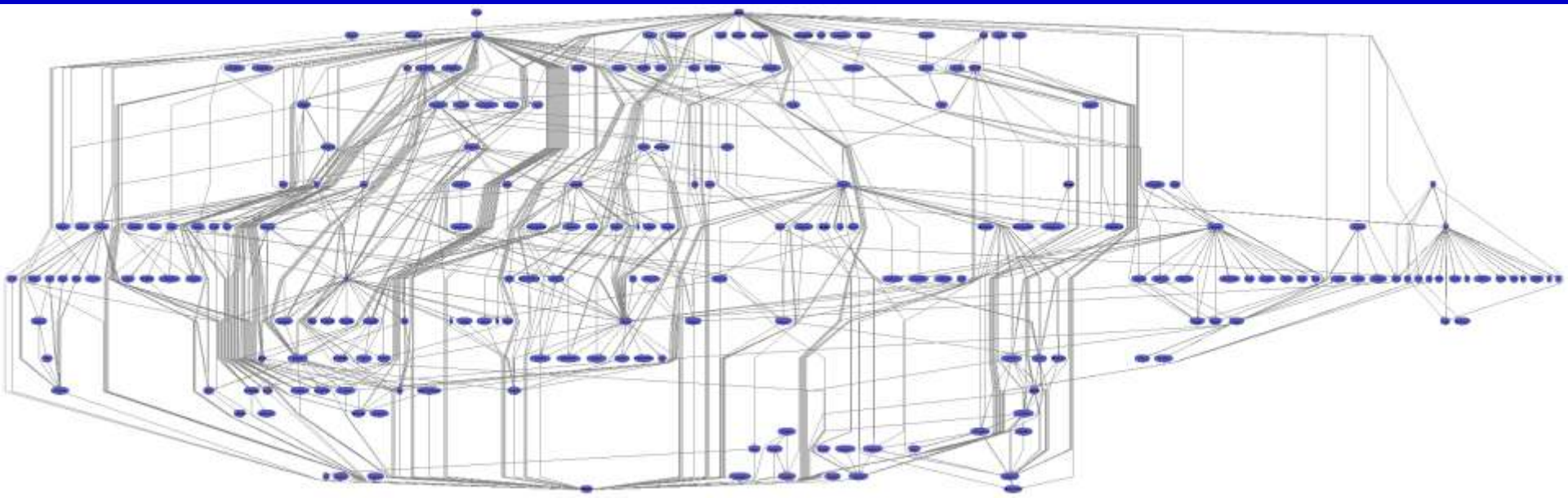
Business Value

Choice
Goal to Optimize
Facts/constraints
Prediction
Simulation
Calculation

QuickTime™ and a decompressor are needed to see this picture.

Goal: Simulation & Optimization

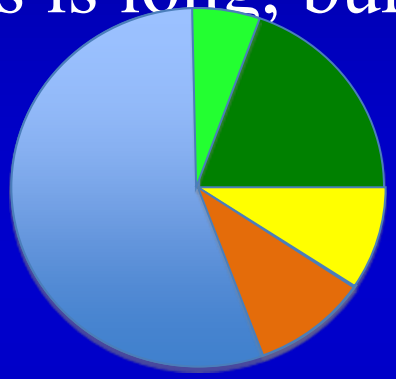
- Build tool to manage IT Service (help tickets)
- New product critical to company (& worked!)
- Analytic breakthrough: keep uncertainty to end
- Success: purchased by HP; increased HP sales



Case 5) Social Security Administration Disability Approval

- **Pain:** Approval process is long, bureaucratic

Up to 2 Years !



1/3 of cases eventually approved

1/5 of cases approved immediately!

With Text Mining,

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

QuickTime™ and a decompressor are needed to see this picture.

Text is Messy and Complex

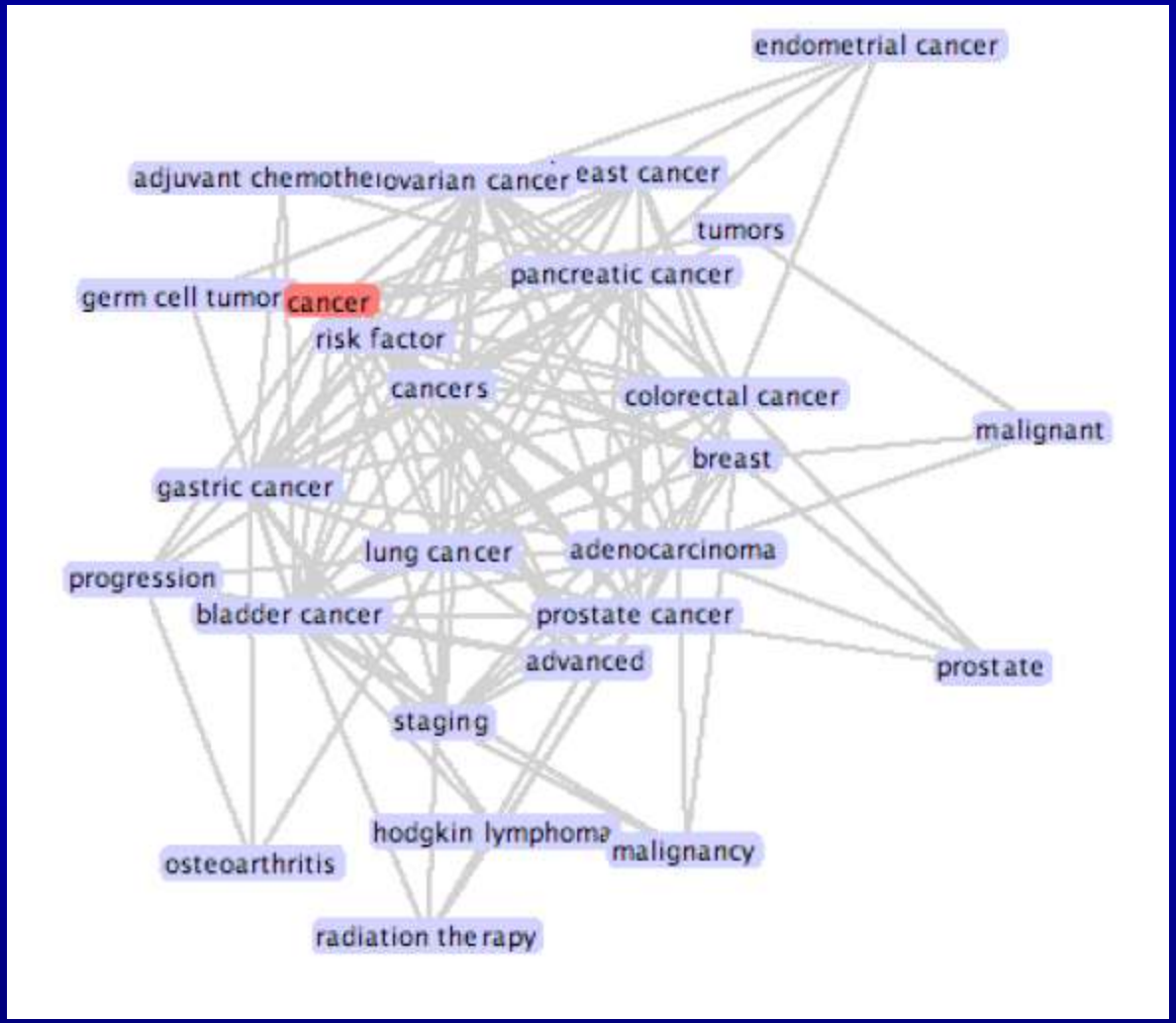
- Multi-Word Phrases (Concepts or Lemmas)
- Stemming (ex: Learn = Learning, Learned, Learns…)
- Synonyms (ex: ALS = Lou Gherig's Disease)
- Misspellings

Ex: 51 phrases found for "Learning Disability"

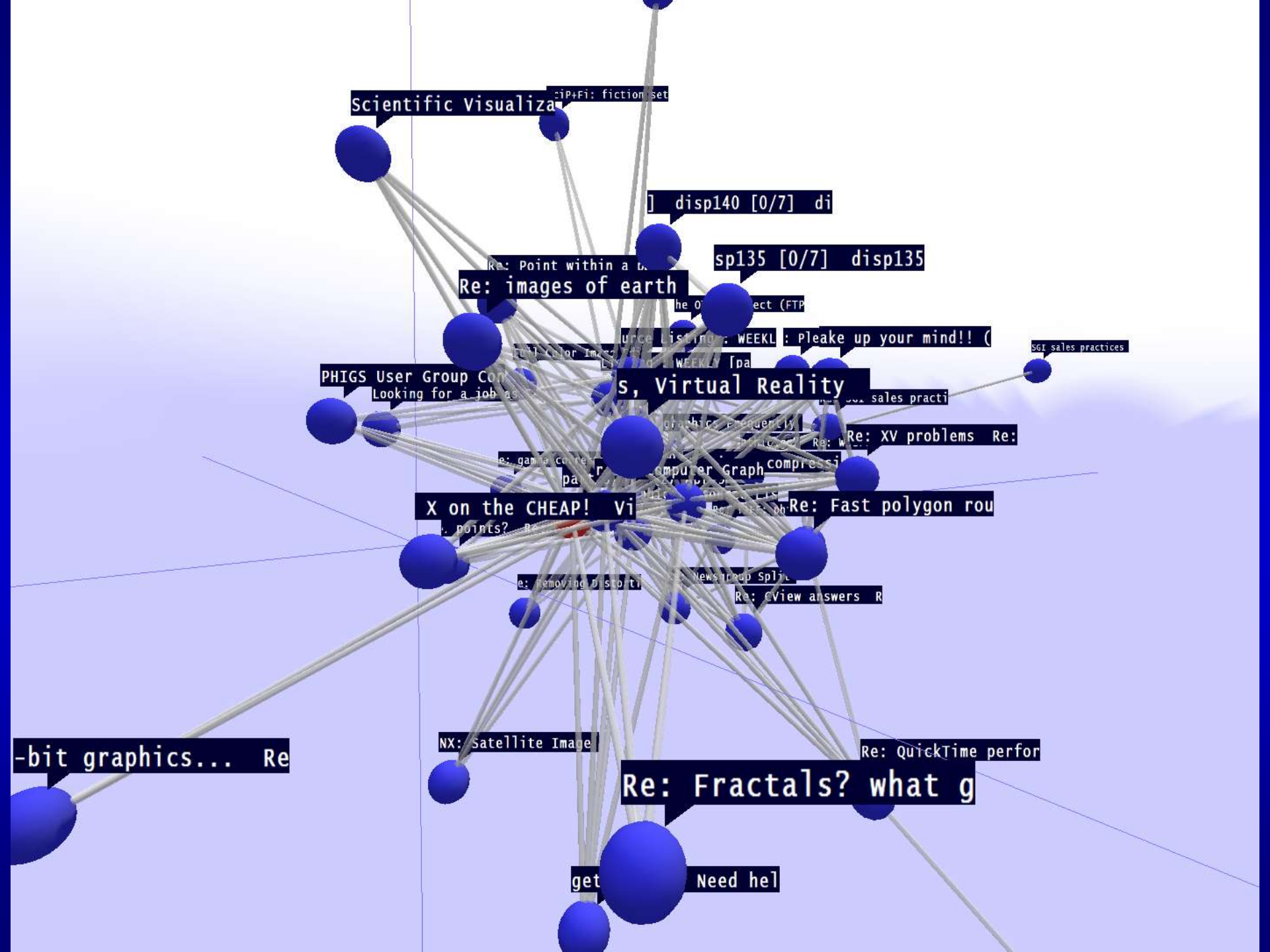
learning disablitiy, learning deisability, learning disablity, learining disabilities, learning disabiality, learningdisabilty, learning disabiilty, learning disabilitty, learning disabilty, learning disablety, learning disabilitiy, learnoing disability, learning disabilties, learning disabiltiy, learning dsblty, learning disibility, learnings disabilty, learningdisability, larning disabilities, learning disabilitiies, learning disabilitties, learning disibilities, learning diasability, learning dasability, learnning disability, learning disabilities, lerning disability, learning disabilites, learneing disability, learninig disability, learning disaiblities, leraning disability, learning disaiblity, learnings disability, learning disabilitys, learning disabillity, learnings disabilities, learning diasbility, learning disabiliites, learning dsiability, learning disabliity, learning disibilty, learning disbilities, learning disbality, learning disbility, learning disabilit, learningdisabilities, learningi disability, lerniung disabilities, learning disabliities, learning disaability, learning disablities

QuickTime™ and a decompressor are needed to see this picture.

ERI new technology: Discover web of word relations



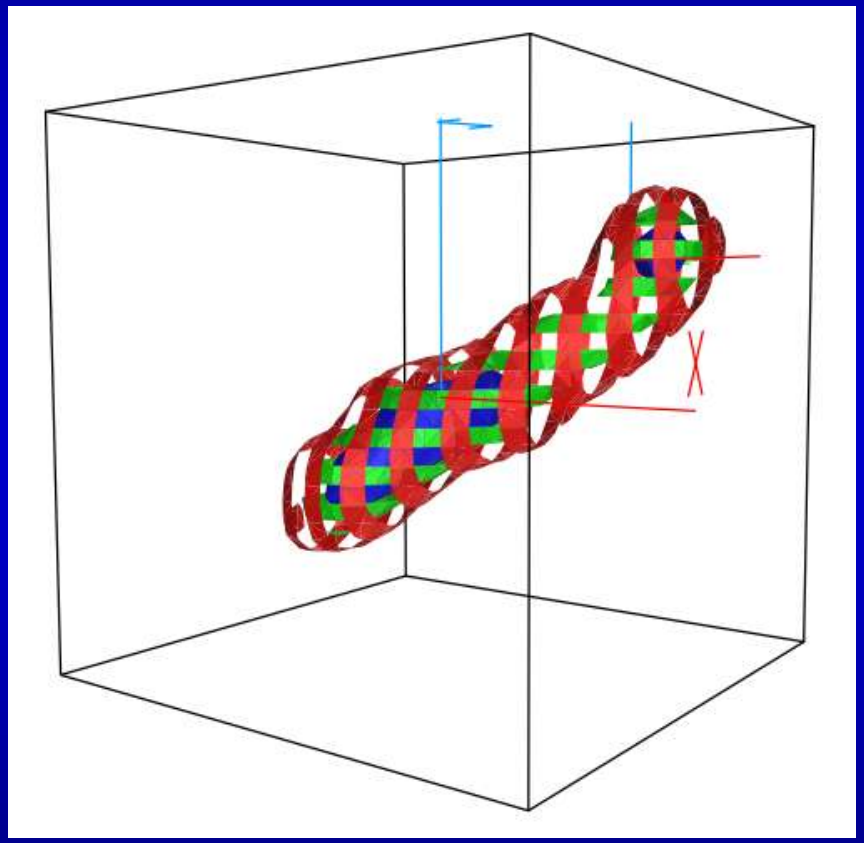
QuickTime™ and a decompressor are needed to see this picture.


Scientific Visualiza
] disp140 [0/7] di
sp135 [0/7] disp135
Re: images of earth
Pleake up your mind!! (
SGI sales practices
PHIGS User Group Co
Looking for a job
s, Virtual Reality
Re: XV problems Re:
X on the CHEAP! Vi
Re: Fast polygon rou
-bit graphics... Re
NX: Satellite Image
Re: QuickTime perfor
Re: Fractals? what g
get
Need hel

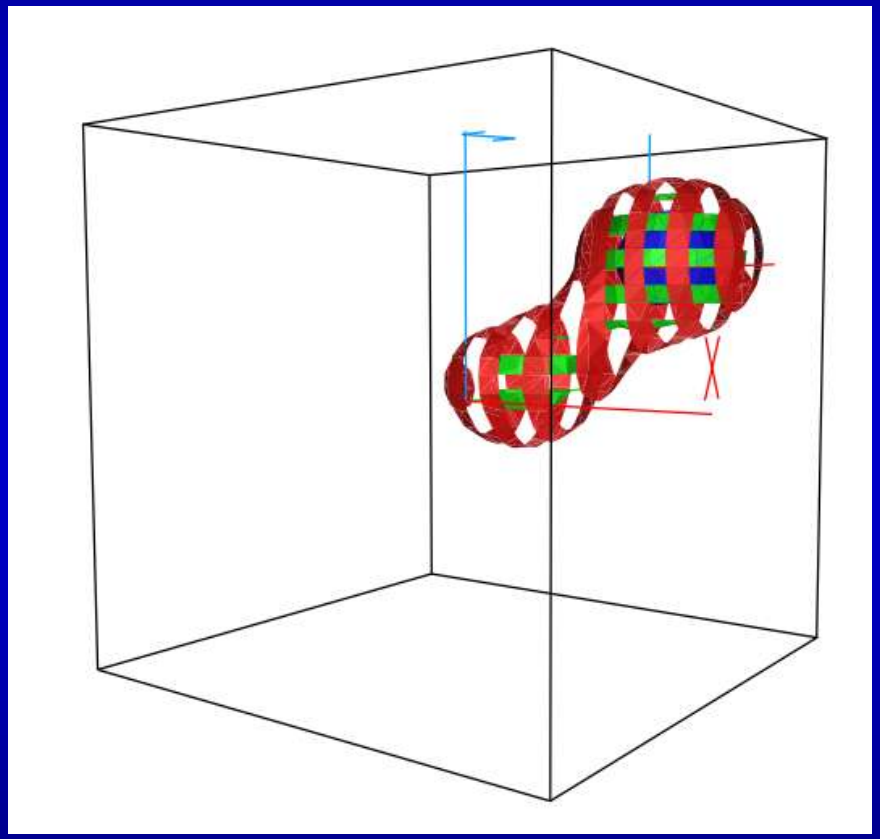
Case 6) Pharmacia & Upjohn (Pfizer)

New Drug Effective?

Placebo



Drug



Density surfaces enclose ascending quartiles of data

QuickTime™ and a decompressor are needed to see this picture.

Bottom Line

- IRS fraud detection
- HP consumer fraud detection
- Lumidigm biometric identification
- Peregrine optimization
- SSA text mining
- Pharmacia & Upjohn (Pfizer) drug efficacy

QuickTime™ and a
decompressor
are needed to see this picture.

Lessons Learned:
Necessary Ingredients for Analytic Project Success

- Gain Expected: either:
  - Leverageable - an incremental improvement will matter, OR
  - "Low-hanging fruit" - nobody's yet (dared) attack the problem
- Interdisciplinary Team: experts needed in business area, statistics, algorithms, and databases
- Data Vigilance: capture and maintain the accumulating information stream
- Time: learning occurs over multiple cycles
- Business Champion essential!

… then Data Mining can add extraordinary value

QuickTime™ and a decompressor are needed to see this picture.

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

Dr. John Elder heads a data mining consulting team with offices in Charlottesville, Virginia, Washington DC, Mountain View, California and Manhasset New York (*www.datamininglab.com*). Founded in 1995, Elder Research, Inc. focuses on Federal, commercial, investment, and security applications of advanced analytics, including text mining, stock selection, image recognition, biometrics, process optimization, cross-selling, drug efficacy, credit scoring, risk management, and fraud detection.

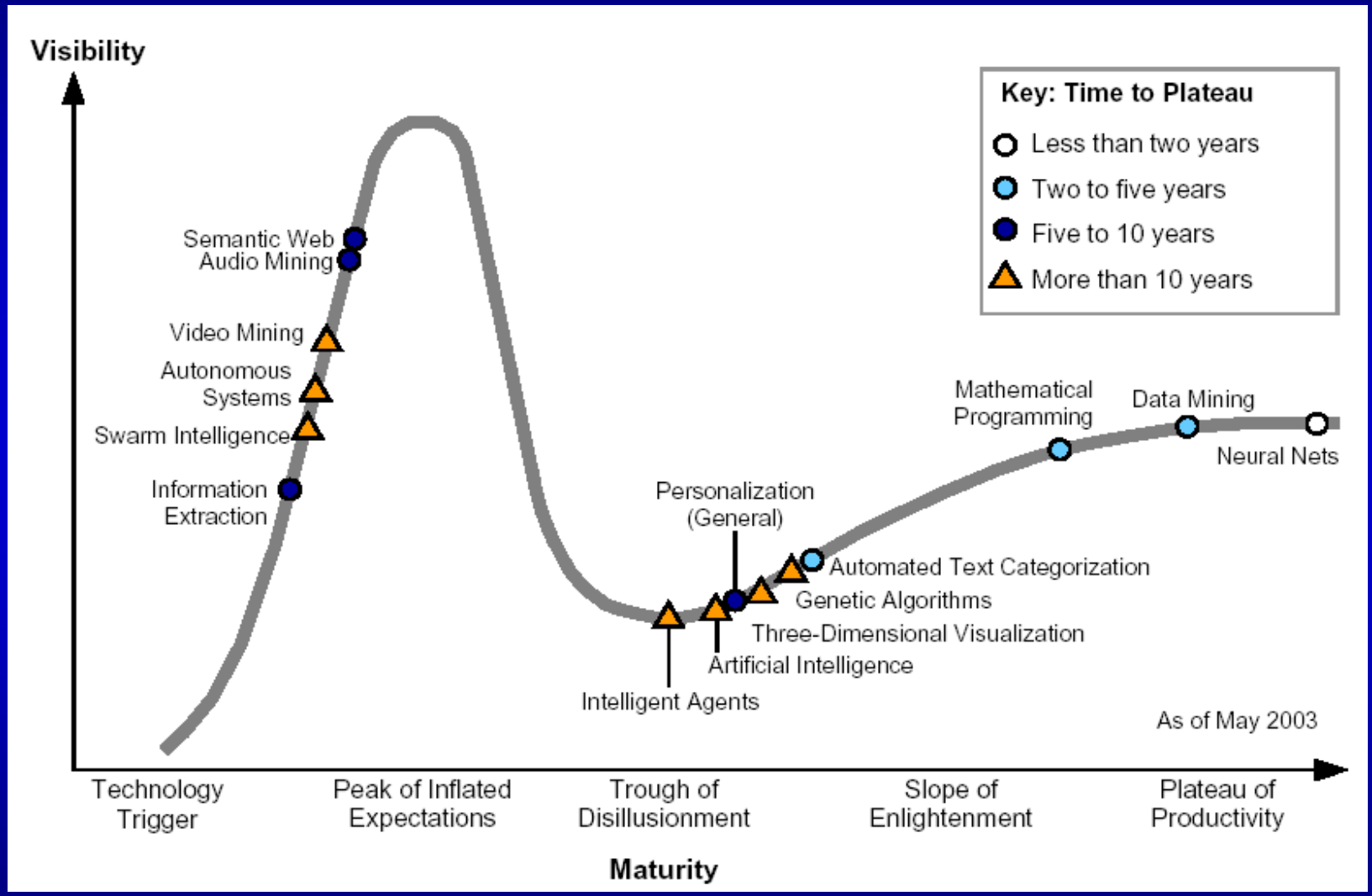
John obtained a BS and MEE in Electrical Engineering from Rice University, and a PhD in Systems Engineering from the University of Virginia, where he's an adjunct professor teaching Optimization or Data Mining. Prior to 15 years at ERI, he spent 5 years in aerospace defense consulting, 4 heading research 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. John's courses on analysis techniques – taught at dozens of universities, companies, and government labs – are noted for their clarity and effectiveness. Dr. Elder was honored to serve for 5 years on a panel appointed by the President to guide technology for National Security. His book on Data Mining, with Bob Nisbet and Gary Miner, won the 2009 PROSE Award for Mathematics. A book on Ensembles with Giovanni Seni was published in 2010.

John is a follower of Christ and the proud father of 5.

QuickTime™ and a
decompressor
are needed to see this picture.

Data Mining and the Hype Cycle



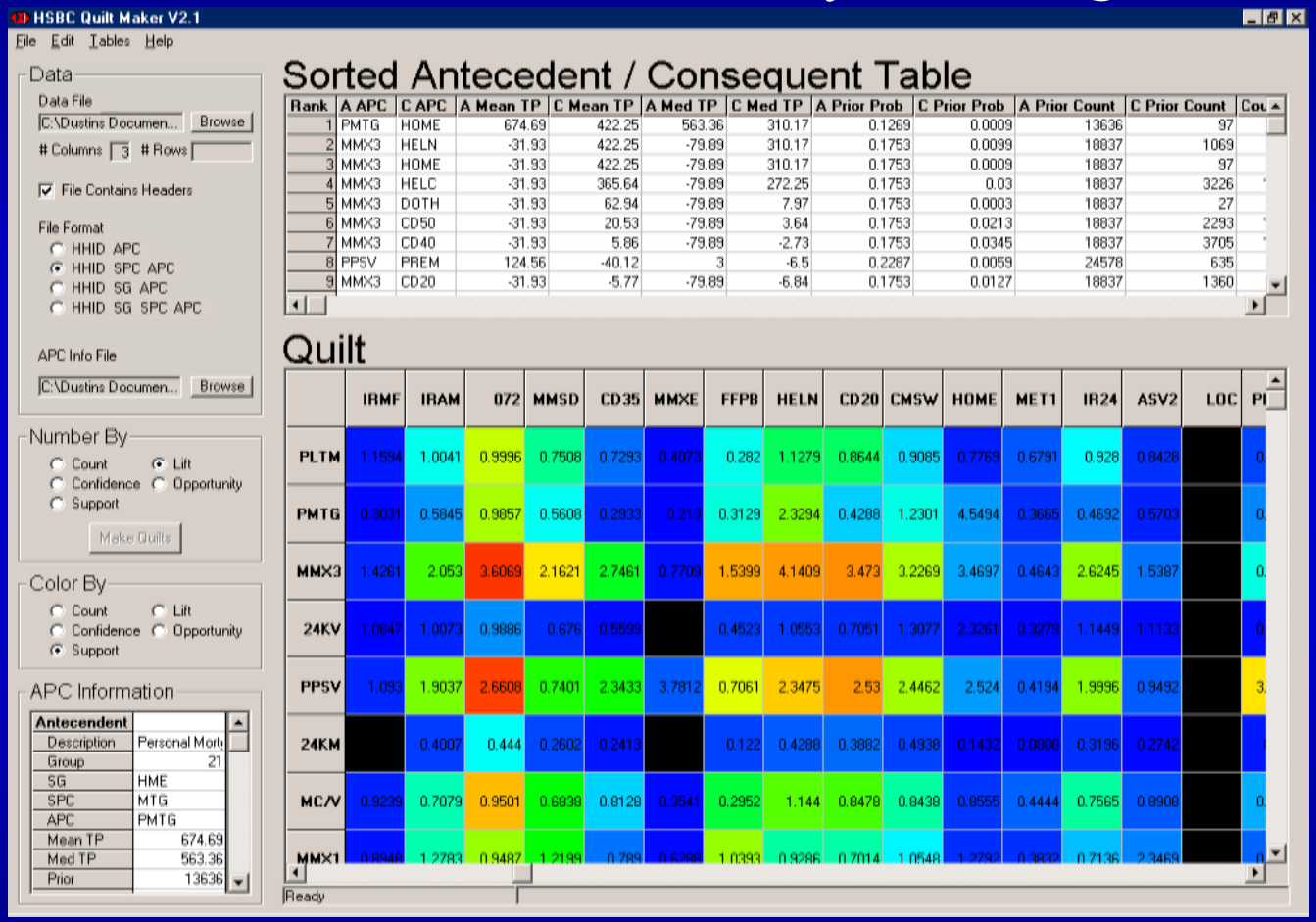
QuickTime™ and a decompressor are needed to see this picture.

Source: Gartner Group

Case 7) HSBC Cross-sell / Up-sell

- Q: What product will interest a customer next?
- Reason: Better target marketing campaigns
- New: Turn contact (a cost) into gain
- Data: transaction history

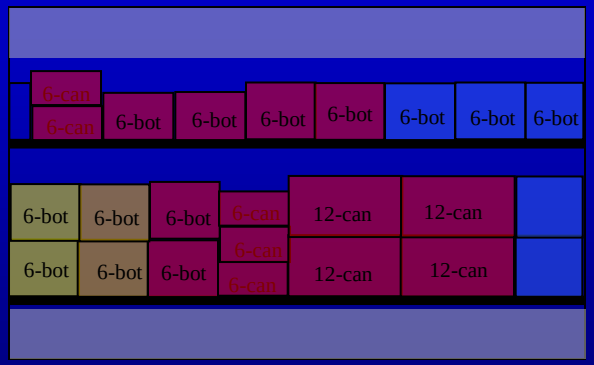
Case 7) Key Technology:
Custom Association Discovery Tool: *QuiltMaker*



Visualize and Quantify next-most-interesting profitable product

QuickTime™ and a decompressor are needed to see this picture.

Case 8) Anheuser-Busch Image Recognition



- Q: How do A-B products appear in the store?
- Reason: Discover and implement configurations that work
- Currently: Takes 4 hours to record configurations
- Goal: automate
- Data: images

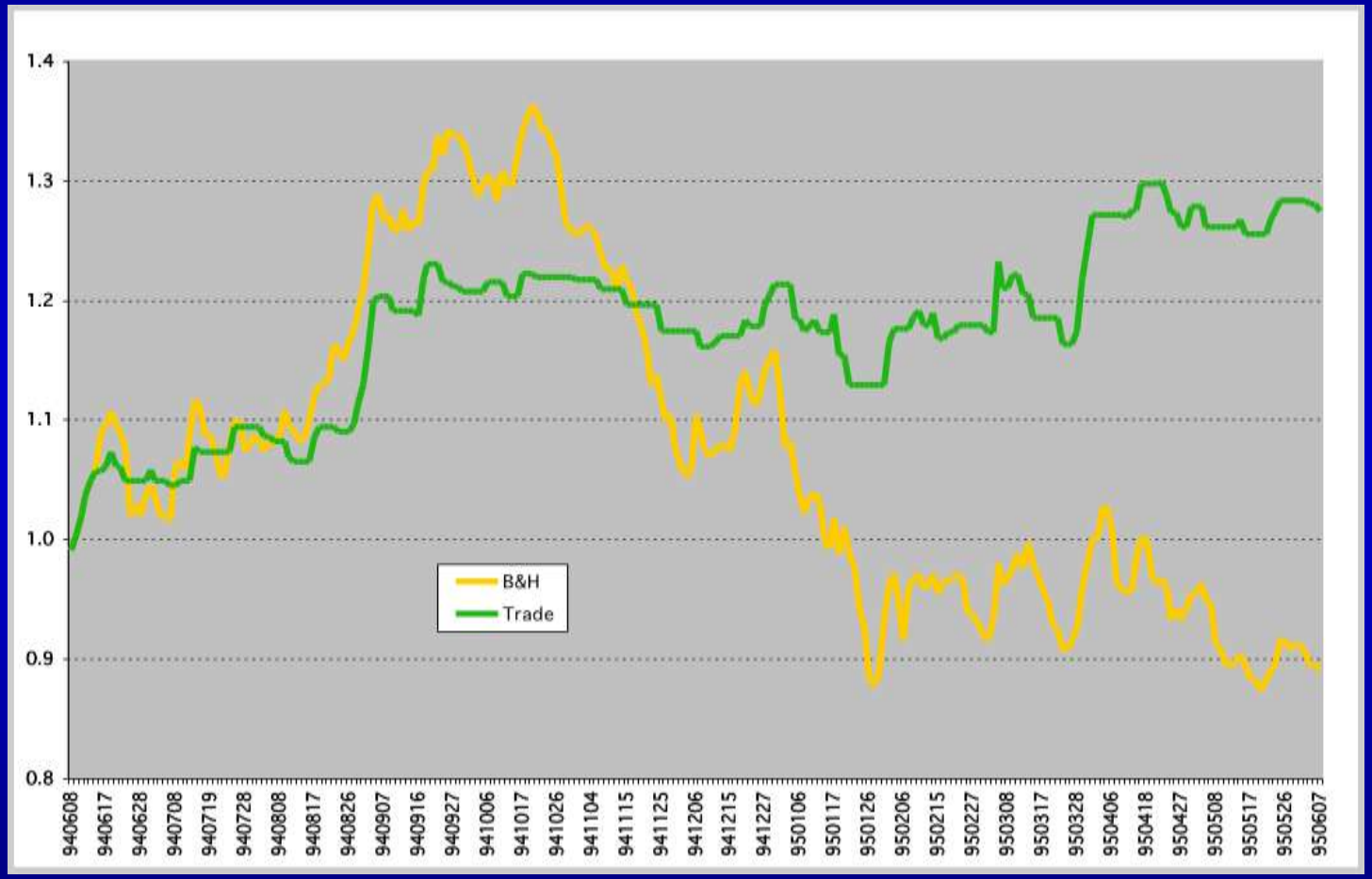
QuickTime™ and a decompressor are needed to see this picture.

- Automation 90% accurate
  -> 10-fold speed-up
- Id'd "stock outs" and "competitor creep"
- Key technological breakthrough: iterative identification

"Planogram": symbolic summary of product configuration. Used for planning and analysis

QuickTime™ and a decompressor are needed to see this picture.

Case 9) WestWind Foundation: Hedge Fund Strategy to Time Market



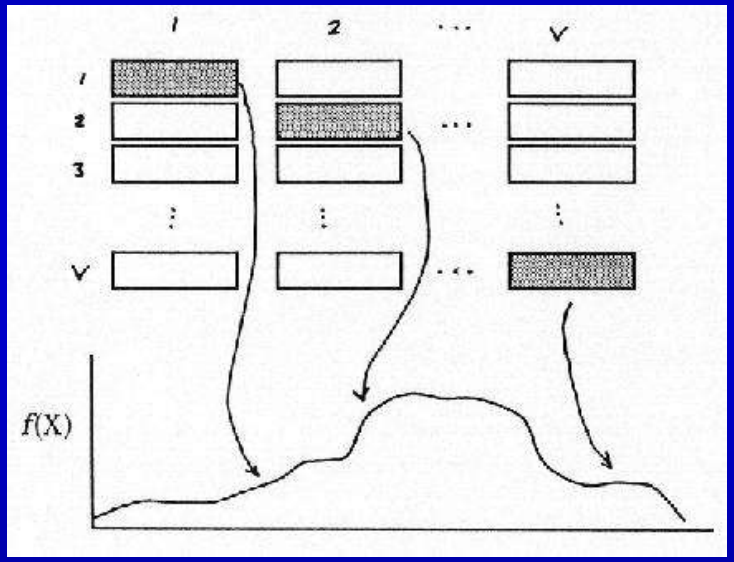
QuickTime™ and a decompressor are needed to see this picture.

Resampling Technology used to determine whether gains were "real":

READ file "fund_1yr" date position return
MULTIPLY position return trade
SUM trade original
PRINT original

REPEAT 1000
  SHUFFLE position pos
  MULTIPLY pos return trade
  SUM trade total
  SCORE total Z
END
HISTOGRAM Z

COUNT Z > original better
DIVIDE better 1000 prop_bet
PRINT prop_bet



QuickTime™ and a decompressor are needed to see this picture.