

Data Mining

Steven Haenchen
CPA/ABV, CFE



Presentation to Mid-America Intergovernmental Audit Forum
Kansas City, Missouri May 1, 2008

Today's objectives

- Understand What Data Mining Is
- Options for looking for the needle in the haystack
- Considerations for looking for evidence when a crime is known or suspected
- Understand Limits to Data Mining



MAMIAF 5/1/2008

2

What Data Mining Is

- Get Data
- Extract Information
- Find Great, Exciting Results!



MAMIAF 5/1/2008

3

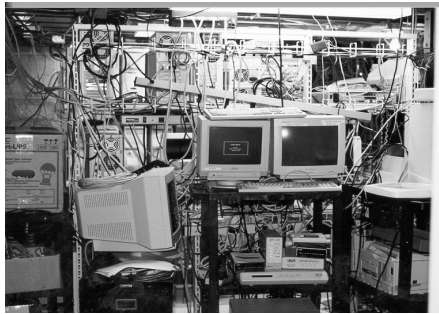
Get Data

- Data Warehouses are generally built by someone other than the data miner.
- Many platforms (mainframes, PCs, notebooks, paper files, Web pages, cell phones, fax machines, notepads)
- Many languages (English, Spanish, French, Legal, Medical, Numeric, Abbreviations, Code, Image)

MAMI AF 5/1/2008

4

Where do you start?



MAMI AF 5/1/2008

5

Start Simple

- Begin with data you can easily obtain
- Begin with data you understand
- Use simple techniques
- Use creativity and enthusiasm with an iterative process in mind!
- Expand the data
- Expand the mining

MAMI AF 5/1/2008

6

Data You Can Obtain

- Server data files (financial, human resources, marketing, email, Web site)
- PC/Notebook files (Spreadsheets, Letters, Memos, email, pictures)
- Choices:
 - Use original data or obtain a copy
 - Use original format or convert to “my” format

MAMIAF 5/1/2008

7

Data You Can Obtain - Example

- The Customer, Vender, and Employee master lists obtained from Finance in Excel
- The Employee master list obtained from Human Resources in Excel
- The Stock Investor list obtained from Treasury in Excel
- List of Business in Yellow Pages ignored

MAMIAF 5/1/2008

8

Data You Understand

- State the problem and formulate the hypothesis
 - Can't do that well if you don't understand the data
- Preprocess the Data
 - Benefits
 - Costs

MAMIAF 5/1/2008

9

Data You Understand - Example

- (State the problem and formulate the hypothesis)
 - Looking for fictitious payees
 - Fictitious payees have the same address of another payee
- (Preprocess the data)
 - All lists “normalized” to contain the same data fields for each record
 - Address fields “normalized” to standard abbreviations, spaces removed, upper case.

MAMIAF 5/1/2008

10

Use simple techniques

- Identify every address and count the times it occurs
- Begin looking at customers, vendors, etc. with addresses repeated the most often

MAMIAF 5/1/2008

11

Use creativity and enthusiasm with an iterative process in mind!

- 123 Apple (123APPLE) did not match with
- 123 Apple Street (123APPLEST)
- When matching addresses, only match the shorter of the two strings!

MAMIAF 5/1/2008

12

Use creativity and enthusiasm with an iterative process in mind!

- 123 Ap Place (123APPL) matched with
- 123 Apple Street (123APPLEST)

- Much more involved, but use a scoring method of each string to other with longer matches scored higher than shorter strings, each to the other, averaged. Only use score > xx%.
- Understand you will have some "noise".

MAMIAF 5/1/2008

13

Expand the data

- Obtain the amount paid each employee (expenses) and vendor
- Eliminate employees and vendors with payments less than \$xxxx

MAMIAF 5/1/2008

14

Use creativity and enthusiasm with an iterative process in mind!

- What makes sense for the business?
 - What states should employees reside in? Any outside those states or ZIP codes?
 - Same with customers and vendors
 - Investigate large payments to Post Office boxes (Is vendor known? In phone book?)
 - Do employee/vendor's expenses get charged to the same account every time?

MAMIAF 5/1/2008

15

Expand the mining - Example

- Categorize employee/vendors by expense type
- Identify “relevant” types and exclude remainder
- For relevant types, determine mean and standard deviations of each invoice
- Analyze invoices that are “outliers”

MAMIAF 5/1/2008

16

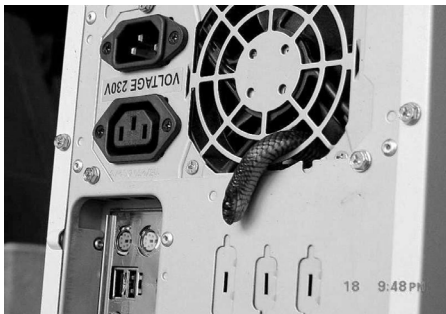
Extract Information

Category	Vendor	Report	Date	Description	No. of Invoices	Last Amount	Last Month	File	Entry
FILE	FILE HIDDEN SYSTEM READ			FILE HIDDEN SYSTEM READ					
FILE	FILE HIDDEN SYSTEM READ			FILE HIDDEN SYSTEM READ					
FILE	FILE HIDDEN SYSTEM READ			FILE HIDDEN SYSTEM READ					
FILE	FILE HIDDEN SYSTEM READ			FILE HIDDEN SYSTEM READ					

MAMIAF 5/1/2008

17

Find Great, Exciting Results!



MAMIAF 5/1/2008

18

Presto – You're a Data Miner!

AWESOME

MAMIAF 5/1/2008

19

Today's objectives

- Understand What Data Mining Is
- Options for looking for the needle in the haystack
- Considerations for looking for evidence when a crime is known or suspected
- Understand Limits to Data Mining



MAMIAF 5/1/2008

20

Refresher

- Stated the problem and formulated the hypothesis
- Collected the data
- Preprocessed the data
- Kept it Simple in an area with Knowledge

READY for Creativity and Enthusiasm!

MAMIAF 5/1/2008

21

Preprocessing

- Simple transformations
- Cleansing and scrubbing, outliers
- Integration
- Aggregation
- Normalizing
- Data smoothing
- Differences
- Ratios

MAMIAP 5/1/2008

22

Preprocessing (continued)

- Benefits – Simpler data, faster processing
- Costs – may not be as accurate as the original data; may not have the ability to expand testing with details no longer available

MAMIAP 5/1/2008

23

Analysis Techniques

- Missing data
- Outlier analysis
- Statistical inference
- Cluster analysis
- Decision trees
- Association rules
- Artificial neural networks
- Genetic algorithms
- Fuzzy logic
- Visualization

MAMIAP 5/1/2008

24

Missing Data

- Ignore the records
- Investigate the records
- Estimate the data
- Compute the probable data

- WHAT MAKES SENSE?
- Noise happens!



MAMIAP 5/1/2008

25

Outlier Analysis

- Ignore the records
- Investigate the records
- BOTH

- WHAT MAKES SENSE?
- Noise happens!



MAMIAP 5/1/2008

26

Statistical Inference

- Find your outliers
- Find repetitive transactions
- Find outliers after adjustment for cycles
- Find outliers after adjustment for relationships
- Benford's Law



MAMIAP 5/1/2008

27

Cluster Analysis

- Group similar together
 - Based on your knowledge
 - Based on algorithms
- Compare groups
- Compare group ratios
 - e.g. office supplies per employee by office; toner usage per office

MAMIAF 5/1/2008

28

Decision Trees

- Advanced cluster analysis



MAMIAF 5/1/2008

29

Association Rules

- Market-Basket Analysis
 - Airfare with motel?
 - Per diem with motel?
 - Hardware with installation charge?
- Algorithm Apriori
 - Airfare, Motel, Cell Phone relationships
- Web Mining
 - Hits
 - Paths (to, from)

MAMIAF 5/1/2008

30

Artificial Neural Networks

- Advanced topic
- Letting the computer do your work



MAMIAF 5/1/2008

31

Genetic Algorithms

- Advanced topic
- Mathematically letting the computer “evolve” your prediction
- You investigate variances from predication

MAMIAF 5/1/2008

32

Fuzzy Logic

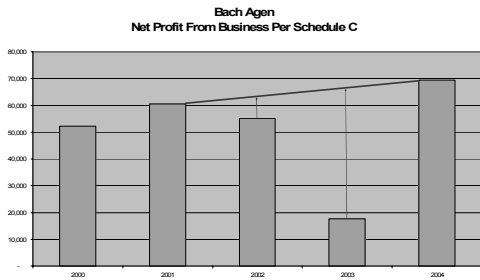
- Text Mining
- Canonical forms
- Keyword searching
- Keyword relationship
- Context
 - The boy hit it over the fence.
- Occurance



MAMIAF 5/1/2008

33

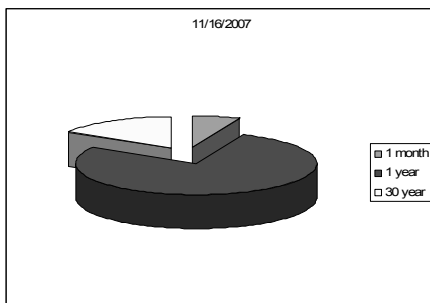
Visualization



MAMIAF 5/1/2008

37

Visualization



MAMIAF 5/1/2008

38

Today's objectives

- Understand What Data Mining Is
- Options for looking for the needle in the haystack
- Considerations for looking for evidence when a crime is known or suspected
- Understand Limits to Data Mining



MAMIAF 5/1/2008

39

Refresher

- Stated the problem and formulated the hypothesis

READY for Creativity and Enthusiasm!

MAMIAF 5/1/2008

40

Gather the Data

- Determine if indeed a crime has been committed
- Determine status of crime (ongoing?)
- Review organization security and audit policy
- Determine need for law enforcement assistance

MAMIAF 5/1/2008

41

Gather the Data (continued)

- Handling of Evidence
 - Relevant
 - Supported by a foundation for its introduction into court
 - Legally obtained
 - Properly identified
 - Properly preserved
- Integrity of the Evidence

MAMIAF 5/1/2008

42

Gather the Data (continued)

- Depends on the allegations
 - System data
 - Access logs (servers, routers, physical security)
 - Hard drive images
 - Cell phone storage
 - USB drives
 - Email server
 - Backup tapes
 - CDs, DVDs, diskettes
 - Fax logs, fax memory
 - Surveillance video

MAMIAF 5/1/2008

43

Assumptions

- Financial data is examined
- Losses calculated
 - Usually an iterative process as you learn more
 - Make sure calculations are reasonable
- Looking for additional evidence – primarily on subject's imaged drive(s)

MAMIAF 5/1/2008

44

Preprocess Data

- Heterogeneous data
- Recovered files that had been deleted
- Partially recovered files that had been deleted
- Deleted files found elsewhere: backup tapes or backups to server drives, diskettes, USB drives, etc.

MAMIAF 5/1/2008

45

Data Mining

- Identify concepts specific to the case
 - Names
 - Accounts
 - Amounts
 - Etc.
- Search data
 - Physically
 - Logically

MAMIAF 5/1/2008

46

Data Mining

- Iterative process

**READY for Creativity and
Enthusiasm!**

MAMIAF 5/1/2008

47

Fuzzy Logic

- Text Mining
- Canonical forms
- Keyword searching
- Keyword relationship
- Context
 - The boy hit it over the fence.
- Occurrence



MAMIAF 5/1/2008

48

Special Consideration

- Hidden files
- Files with false extensions
 - Renamed
 - Doc in Xls
 - CRC validation tables
- Encrypted files
- Messages embedded in pictures

MAMIAF 5/1/2008

49

Special Consideration

- Infected files
- Passwords
- Pulling the plug
- The interview

MAMIAF 5/1/2008

50

Limits

- Must understand computer hardware and systems
 - Company email in email server if deleted by subject, but Internet access to personal email is not
 - What information is hidden in the registry?
 - What information is hidden in temp files and cookies?
 - What are retention processes in place (not policies)

MAMIAF 5/1/2008

51

Today's objectives

- Understand What Data Mining Is
- Options for looking for the needle in the haystack
- Considerations for looking for evidence when a crime is known or suspected
- Understand Limits to Data Mining



MAMIAF 5/1/2008

52

Understand Limits to Data Mining

- It is tempting to develop a theory to fit an oddity found in the data.
- One can find evidence to support any preconception if you let the computer churn long enough.
- A finding makes more sense if there is a plausible theory for it. But a beguiling story can disguise weaknesses in the data.
- The more factors or features in a data set the computer considers, the more likely the program will find a relationship, valid or not.

Peter Coy, Business Week

MAMIAF 5/1/2008

53

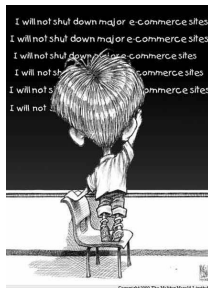
Questions???

Steven Haenchen
CPAABV, CFE
MCSO, MCDBA, MCSA, MCSE
Haenchen Valuation Services, Inc.
Communicating Complex Messages Clearly
10436 Oakmont
Overland Park, KS 66215
Web: www.haenchenvaluations.com
Email: Haenchen@HaenchenValuations.com or
HaenchenSL@hotmail.com
Work: (913) 825-5235
Cell: (913) 488-7187



Communicating Complex Messages Clearly
Business Valuations
Haenchen Valuation Services

MAMIAF 5/1/2008



54
