Data Mining

Necessity Is the Mother of Invention

Why Data Mining?—Potential Applications

- Data analysis and decision support
  - Market analysis and management
    - Target marketing, customer relationship management (CRM), market basket analysis, cross selling, market segmentation
  - Risk analysis and management
    - Forecasting, customer retention, improved underwriting, quality control, competitive analysis
  - Fraud detection and detection of unusual patterns (outliers)
- Other Applications
  - Text mining (news group, email, documents) and Web mining
  - Stream data mining
  - DNA and bio-data analysis

Market Analysis and Management

- Where does the data come from?
  - Credit card transactions, loyalty cards, discount coupons, customer complaint calls, plus (public) lifestyle studies
- Target marketing
  - Find clusters of "model" customers who share the same characteristics: interest, income level, spending habits, etc.
  - Determine customer purchasing patterns over time
- Cross-market analysis
  - Associations/co-relations between product sales, & prediction based on such association
- Customer profiling
  - What types of customers buy what products (clustering or classification)
- Customer requirement analysis
  - identifying the best products for different customers
  - predict what factors will attract new customers
- Provision of summary information
  - multidimensional summary reports
  - statistical summary information (data central tendency and variation)

Corporate Analysis & Risk Management

- Finance planning and asset evaluation
  - cash flow analysis and prediction
  - contingent claim analysis to evaluate assets
  - cross-sectional and time series analysis (financial-ratio, trend analysis, etc.)
- Resource planning
  - summarize and compare the resources and spending
- Competition
  - monitor competitors and market directions
  - group customers into classes and a class-based pricing procedure
  - set pricing strategy in a highly competitive market

Fraud Detection & Mining Unusual Patterns

- Approaches: Clustering & model construction for frauds, outlier analysis
- Applications: Health care, retail, credit card service, telecomm.
  - Auto insurance: ring of collisions
  - Money laundering: suspicious monetary transactions
  - Medical insurance
    - Professional patients, ring of doctors, and ring of references
    - Unnecessary or correlated screening tests
  - Telecommunications: phone-call fraud
    - Phone call model: destination of the call, duration, time of day or week. Analyze patterns that deviate from an expected norm
  - Retail industry
    - Analysts estimate that 38% of retail shrink is due to dishonest employees
  - Anti-terrorism

Other Applications

- Sports
  - IBM Advanced Scout analyzed NBA game statistics (shots blocked, assists, and fouls) to gain competitive advantage for New York Knicks and Miami Heat
- Astronomy
  - JPL and the Palomar Observatory discovered 22 quasars with the help of data mining
- Internet Web Surf-Aid
  - IBM Surf-Aid applies data mining algorithms to Web access logs for market-related pages to discover customer preference and behavior pages, analyzing effectiveness of Web marketing, improving Web site organization, etc.

Data Mining Functionalities

- Concept description: Characterization and discrimination
  - Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet regions
- Association (correlation and causality)
  - Diaper → Beer [0.5%, 75%]
- Classification and Prediction
  - Construct models (functions) that describe and distinguish classes or concepts for future prediction
    - E.g., classify countries based on climate, or classify cars based on gas mileage
  - Presentation: decision-tree, classification rule, neural network
  - Predict some unknown or missing numerical values

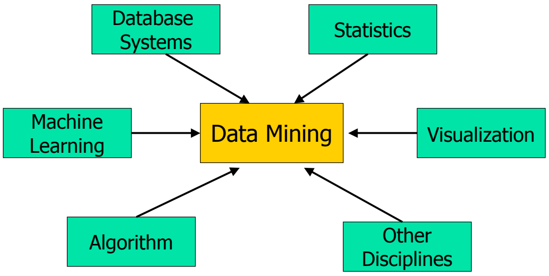
Data Mining Functionalities (2)

- Cluster analysis
  - Class label is unknown: Group data to form new classes, e.g., cluster houses to find distribution patterns
  - Maximizing intra-class similarity & minimizing interclass similarity
- Outlier analysis
  - Outlier: a data object that does not comply with the general behavior of the data
  - Noise or exception? No! useful in fraud detection, rare events analysis
- Trend and evolution analysis
  - Trend and deviation: regression analysis
  - Sequential pattern mining, periodicity analysis
  - Similarity-based analysis
- Other pattern-directed or statistical analyses

Are All the "Discovered" Patterns Interesting?

- Data mining may generate thousands of patterns: Not all of them are interesting
  - Suggested approach: Human-centered, query-based, focused mining
- **Interestingness measures**
  - A pattern is interesting if it is easily understood by humans, valid on new or test data with some degree of certainty, potentially useful, novel, or validates some hypothesis that a user seeks to confirm
- **Objective vs. subjective interestingness measures**
  - Objective: based on statistics and structures of patterns, e.g., support, confidence, etc.
  - Subjective: based on user's belief in the data, e.g., unexpectedness, novelty, actionability, etc.

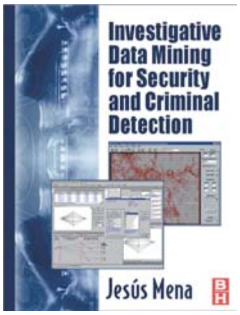
Data Mining: Confluence of Multiple Disciplines

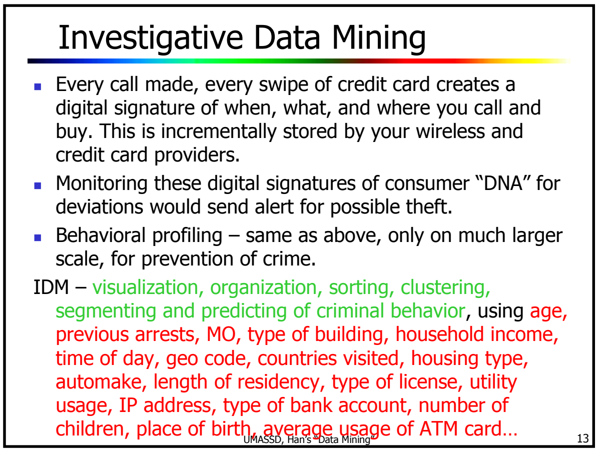


Data Mining: Classification Schemes

- General functionality
  - Descriptive data mining
  - Predictive data mining
- Different views, different classifications
  - Kinds of data to be mined
  - Kinds of knowledge to be discovered
  - Kinds of techniques utilized
  - Kinds of applications adapted

Former IRS agent and data miner

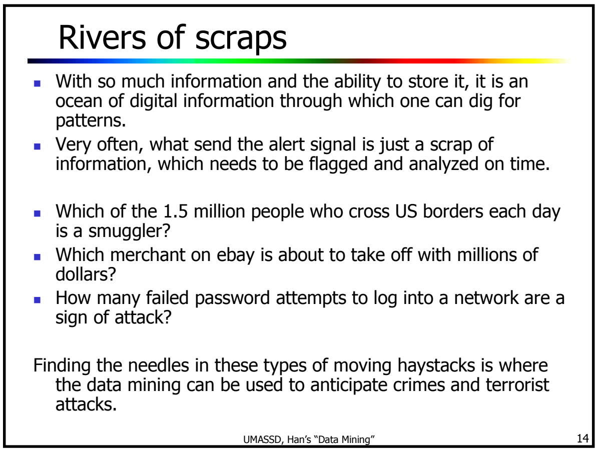

Investigative Data Mining for Security and Criminal Detection
Jesús Mena
BH



Investigative Data Mining

- Every call made, every swipe of credit card creates a digital signature of when, what, and where you call and buy. This is incrementally stored by your wireless and credit card providers.
- Monitoring these digital signatures of consumer "DNA" for deviations would send alert for possible theft.
- Behavioral profiling – same as above, only on much larger scale, for prevention of crime.

IDM – visualization, organization, sorting, clustering, segmenting and predicting of criminal behavior, using age, previous arrests, MO, type of building, household income, time of day, geo code, countries visited, housing type, automake, length of residency, type of license, utility usage, IP address, type of bank account, number of children, place of birth, average usage of ATM card…

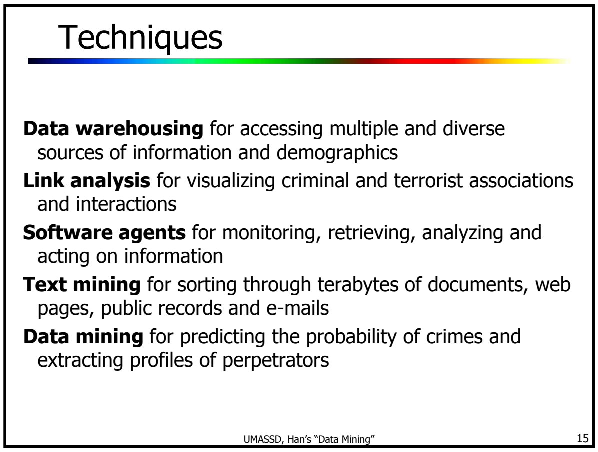


Rivers of scraps

- With so much information and the ability to store it, it is an ocean of digital information through which one can dig for patterns.
- Very often, what send the alert signal is just a scrap of information, which needs to be flagged and analyzed on time.

- Which of the 1.5 million people who cross US borders each day is a smuggler?
- Which merchant on ebay is about to take off with millions of dollars?
- How many failed password attempts to log into a network are a sign of attack?

Finding the needles in these types of moving haystacks is where the data mining can be used to anticipate crimes and terrorist attacks.



Techniques

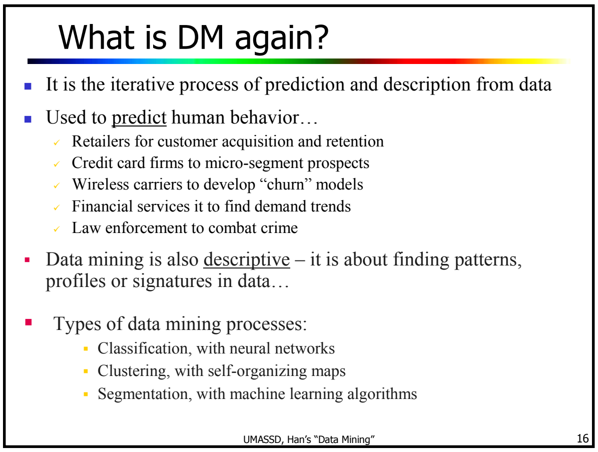
Data warehousing for accessing multiple and diverse sources of information and demographics

Link analysis for visualizing criminal and terrorist associations and interactions

Software agents for monitoring, retrieving, analyzing and acting on information

Text mining for sorting through terabytes of documents, web pages, public records and e-mails

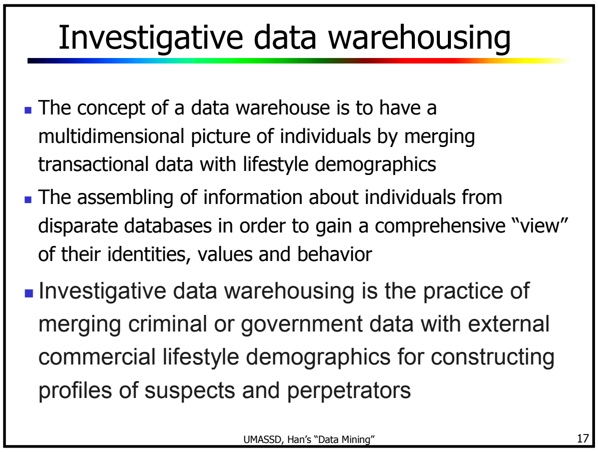
Data mining for predicting the probability of crimes and extracting profiles of perpetrators



What is DM again?

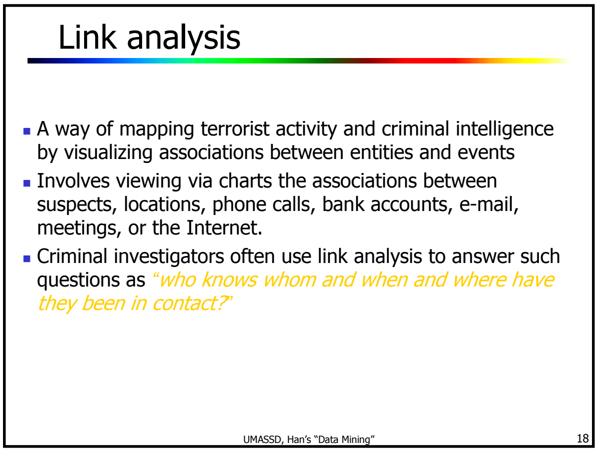
- It is the iterative process of prediction and description from data
- Used to predict human behavior…
  - ✓ Retailers for customer acquisition and retention
  - ✓ Credit card firms to micro-segment prospects
  - ✓ Wireless carriers to develop "churn" models
  - ✓ Financial services it to find demand trends
  - ✓ Law enforcement to combat crime
- Data mining is also descriptive – it is about finding patterns, profiles or signatures in data…

- Types of data mining processes:
  - Classification, with neural networks
  - Clustering, with self-organizing maps
  - Segmentation, with machine learning algorithms



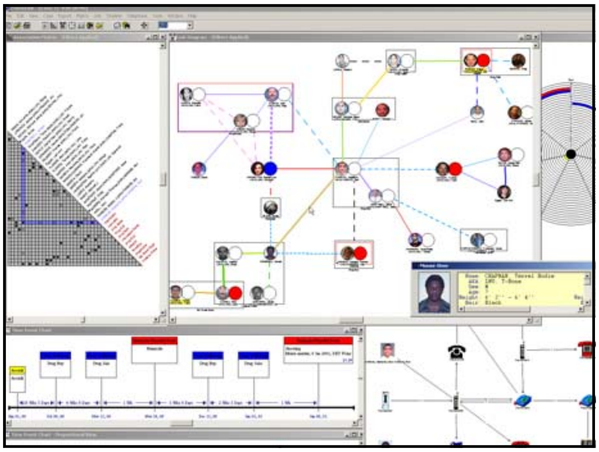
Investigative data warehousing

- The concept of a data warehouse is to have a multidimensional picture of individuals by merging transactional data with lifestyle demographics
- The assembling of information about individuals from disparate databases in order to gain a comprehensive "view" of their identities, values and behavior
- Investigative data warehousing is the practice of merging criminal or government data with external commercial lifestyle demographics for constructing profiles of suspects and perpetrators



Link analysis

- A way of mapping terrorist activity and criminal intelligence by visualizing associations between entities and events
- Involves viewing via charts the associations between suspects, locations, phone calls, bank accounts, e-mail, meetings, or the Internet.
- Criminal investigators often use link analysis to answer such questions as *"who knows whom and when and where have they been in contact?"*

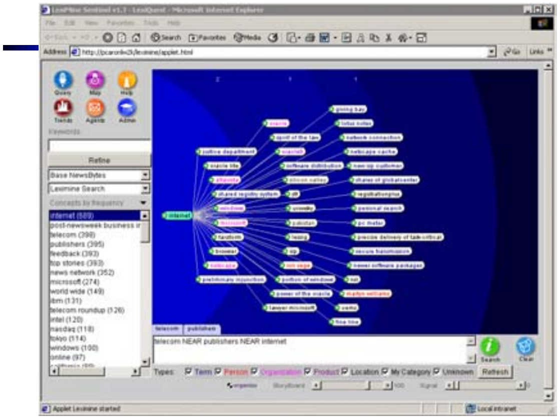


Text mining

- Text mining software is being used to categorize and route content to specific users based on several technologies:
  1. Rule-based, manually
  2. Statistical, naïve Bayes or nearest neighbor
  3. Support Vector (yes/no)
  4. Probabilistic Latent Semantic Analysis
- Text mining can extract key concepts and match keywords from unstructured data

Text mining investigations

- Investigators and analysts can sort, organize and analyze gigabytes of text during their inquiries
- Applied to the problem of searching and locating names or terms used in e-mails, wireless phone calls, faxes, instant messages, chat rooms, etc.
- Police in the UK are using text mining to organize criminal cases – institutionalizing their knowledge of criminal activities by perpetrators' MOs



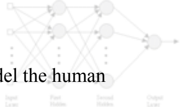
Intelligent agents

- Agents are software programs that perform user delegated tasks autonomously, such as retrieve specific data over networks
- Increasingly used in the area of intrusion detection, for monitoring systems and networks – to deter hacker attacks
- Agents can be assigned tasks, such as mining a database and communicating its results

Software detectives

- Agents are automated programs running independently over networks
- Agents represent concepts of reasoning and autonomous learning
- Agents can be used with machine learning for remote data mining
- Distributed data mining firms: InfoGlide and InferX

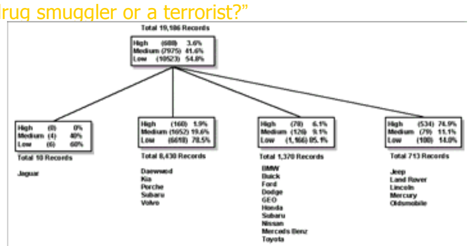
Neural networks



- Neural networks are software system that model the human process of learning and remembering
- Can be used for classifying patterns of digital and physical criminal evidence and predicting crimes
- One of the first and most successful applications is in the area of credit card fraud detection by HNC
- Neural networks must be ***trained,*** this type of software is really about "remembering"
- Networks are trained by exposing them to samples, enabling them to recognize patterns
- Networks have been used to match the forensic *"signature"* of kerosene in arson cases by criminalists at the California Department of Justice

Machine learning

- A pivotal technology for profiling perpetrators
- Algorithms automate the manual process of discovering key features (attributes) and intervals (ranges) in databases
- They can answer such questions as "when is fraud most likely to take place?" or, "what are the characteristics of a drug smuggler or a terrorist?"



Profiling via Pattern Recognition

- Turvey emphasizes that "A full forensic analysis must be performed on all available physical evidence before profiling can begin."
- The same is true with investigative data mining, the tools are different but the methodology is the same
- A data mining profile is based on the evidence of digital observations found in the data

Data features

- Age
- Sex
- Race
- Residence
- Intelligence level
- Occupation
- Marital status
- Living arrangements
- Type and condition of vehicle
- Motivating factors
- Arrest record
- Provocation factors
- Possible interrogation techniques

Profile

Using demographics, insurance files, DMV records, etc. a data mining analysis may yield this type of profile:

```
If INSURER STATUS = None
AND # OF CROSSING THIS WEEK = 8
AND TITLE OWNERSHIP = Owned
AND VEHICLE MAKE = Jeep
AND DRIVER CITY = Out of State
AND DEMOGRAPHIC NEIGHBORHOD = High Rise Renters
THEN Potential Smuggler = 72% Probability
```

Case study – clustering burglars

- West Midland Police, UK
- Bogus official burglars (persons gaining access to premises by deception with the intention to steal property)
- Tool: SOM neural network (used for clustering the input information and outputting a topological map with the found patterns). Clementine software package.
- Problem: volume of such burglaries over a wide geographical area makes it difficult to link crimes committed by the same offender(s).

SOM

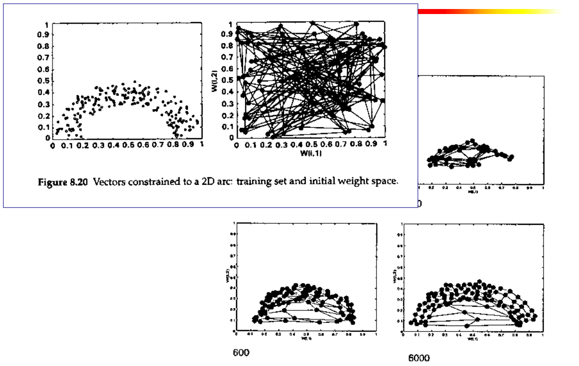


Figure 8.20 Vectors constrained to a 2D arc: training set and initial weight space.

Figure 8.21 Vectors constrained to a 2D arc: "snapshots" in weight space during training.

Data selection, cleaning, and coding

- Elderly victims (memory, correctness of description…)
- When crime is reported, an officer attends the scene and takes a crime report
- Correctness of crime report, errors in entries, vagueness of description
- Unification of reports (providing specific fields for age, gender, height, hair color and length, build, accent, race, number of accomplices).

Coding of information

- relativity of information (height, weight, hair length, age)
- information with continuous, binary, nominal, and ordinal scales

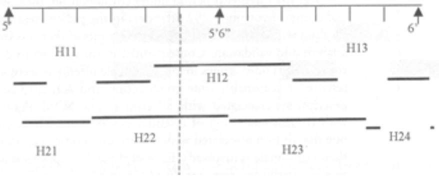
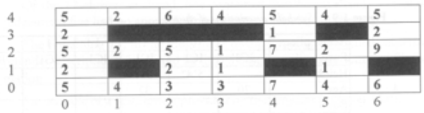


Figure 1.7 *Illustrative example of the encoding of height as zero or one.*

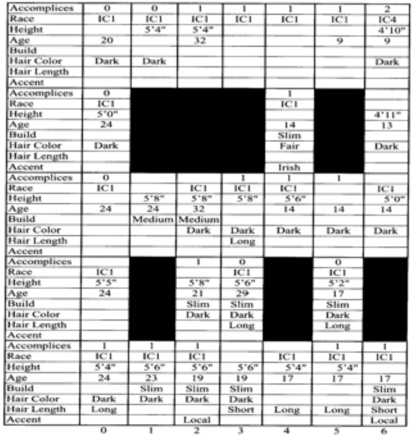
Application and findings

Missing values, how to manage?

Combining neighboring clusters?

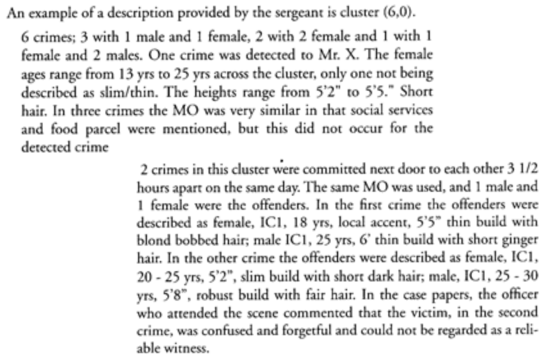


4	5	2	6	4	5	4	5
3	2	■	■	■	1	■	2
2	5	2	5	1	7	2	9
1	2	■	2	1	■	1	■
0	5	4	3	3	7	4	6
	0	1	2	3	4	5	6



4	Accomplices	0	0	1	1	1	1	2
	Race	IC1	IC1	IC1	IC1	IC1	IC1	IC4
	Height		5'4"	5'4"				4'10"
	Age	20		32			9	9
	Build							
	Hair Color	Dark	Dark					Dark
	Hair Length							
	Accent							
3	Accomplices	0	■	■	■	1	■	
	Race	IC1	■	■	■	IC1	■	
	Height	5'0"	■	■	■		■	4'11"
	Age	24	■	■	■	14	■	13
	Build		■	■	■	Slim	■	
	Hair Color	Dark	■	■	■	Fair	■	Dark
	Hair Length		■	■	■		■	
	Accent		■	■	■	Irish	■	
2	Accomplices	0			1	1	1	
	Race	IC1		IC1	IC1	IC1		IC1
	Height		5'8"	5'8"	5'8"	5'6"		5'0"
	Age	24	24	32		14	14	14
	Build		Medium	Medium				
	Hair Color			Dark	Dark	Dark	Dark	Dark
	Hair Length				Long			
	Accent							
1	Accomplices		■	1	0	■	0	■
	Race	IC1	■		IC1	■	IC1	■
	Height	5'5"	■	5'8"	5'6"	■	5'2"	■
	Age	24	■	21	29	■	17	■
	Build		■	Slim	Slim	■	Slim	■
	Hair Color		■	Dark	Dark	■	Dark	■
	Hair Length		■		Long	■	Long	■
	Accent		■			■		■
0	Accomplices	1	1	1			1	1
	Race	IC1	IC1	IC1		IC1	IC1	IC1
	Height	5'4"	5'6"	5'6"	5'6"	5'4"	5'4"	
	Age	24	23	19	19	17	17	17
	Build		Slim	Slim	Slim			Slim
	Hair Color	Dark	Dark	Dark	Dark			Dark
	Hair Length	Long			Short	Long	Long	Short
	Accent			Local				Local
		0	1	2	3	4	5	6

Validation

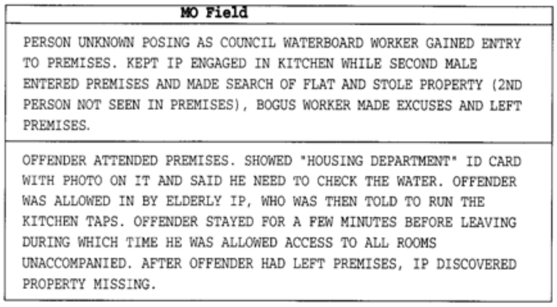


An example of a description provided by the sergeant is cluster (6,0).

6 crimes; 3 with 1 male and 1 female, 2 with 2 female and 1 with 1 female and 2 males. One crime was detected to Mr. X. The female ages range from 13 yrs to 25 yrs across the cluster, only one not being described as slim/thin. The heights range from 5'2" to 5'5." Short hair. In three crimes the MO was very similar in that social services and food parcel were mentioned, but this did not occur for the detected crime

2 crimes in this cluster were committed next door to each other 3 1/2 hours apart on the same day. The same MO was used, and 1 male and 1 female were the offenders. In the first crime the offenders were described as female, IC1, 18 yrs, local accent, 5'5" thin build with blond bobbed hair; male IC1, 25 yrs, 6' thin build with short ginger hair. In the other crime the offenders were described as female, IC1, 20 - 25 yrs, 5'2", slim build with short dark hair; male, IC1, 25 - 30 yrs, 5'8", robust build with fair hair. In the case papers, the officer who attended the scene commented that the victim, in the second crime, was confused and forgetful and could not be regarded as a reliable witness.

Free field information



MO Field

PERSON UNKNOWN POSING AS COUNCIL WATERBOARD WORKER GAINED ENTRY TO PREMISES. KEPT IP ENGAGED IN KITCHEN WHILE SECOND MALE ENTERED PREMISES AND MADE SEARCH OF FLAT AND STOLE PROPERTY (2ND PERSON NOT SEEN IN PREMISES), BOGUS WORKER MADE EXCUSES AND LEFT PREMISES.

OFFENDER ATTENDED PREMISES. SHOWED "HOUSING DEPARTMENT" ID CARD WITH PHOTO ON IT AND SAID HE NEED TO CHECK THE WATER. OFFENDER WAS ALLOWED IN BY ELDERLY IP, WHO WAS THEN TOLD TO RUN THE KITCHEN TAPS. OFFENDER STAYED FOR A FEW MINUTES BEFORE LEAVING DURING WHICH TIME HE WAS ALLOWED ACCESS TO ALL ROOMS UNACCOMPANIED. AFTER OFFENDER HAD LEFT PREMISES, IP DISCOVERED PROPERTY MISSING.

Multi-Dimensional View of Data Mining

- **Data to be mined**
  - Relational, data warehouse, transactional, stream, object-oriented/relational, active, spatial, time-series, text, multi-media, heterogeneous, legacy, WWW
- **Knowledge to be mined**
  - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
  - Multiple/integrated functions and mining at multiple levels
- **Techniques utilized**
  - Database-oriented, data warehouse (OLAP), machine learning, statistics, visualization, etc.
- **Applications adapted**
  - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, Web mining, etc.

Where to Find References?

- Data mining and KDD (SIGKDD: CDROM)
  - Conferences: ACM-SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, PAKDD, etc.
  - Journal: Data Mining and Knowledge Discovery, KDD Explorations
- Database systems (SIGMOD: CD ROM)
  - Conferences: ACM-SIGMOD, ACM-PODS, VLDB, IEEE-ICDE, EDBT, ICDT, DASFAA
  - Journals: ACM-TODS, IEEE-TKDE, JIIS, J. ACM, etc.
- AI & Machine Learning
  - Conferences: Machine learning (ML), AAAI, IJCAI, COLT (Learning Theory), etc.
  - Journals: Machine Learning, Artificial Intelligence, etc.
- Statistics
  - Conferences: Joint Stat. Meeting, etc.
  - Journals: Annals of statistics, etc.
- Visualization
  - Conference proceedings: CHI, ACM-SIGGraph, etc.
  - Journals: IEEE Trans. visualization and computer graphics, etc.

Recommended Reference Books

- R. Agrawal, J. Han, and H. Mannila, Readings in Data Mining: A Database Perspective, Morgan Kaufmann (in preparation)
- U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996
- U. Fayyad, G. Grinstein, and A. Wierse, Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2001
- D. J. Hand, H. Mannila, and P. Smyth, Principles of Data Mining, MIT Press, 2001
- T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer-Verlag, 2001
- T. M. Mitchell, Machine Learning, McGraw Hill, 1997
- G. Piatetsky-Shapiro and W. J. Frawley. Knowledge Discovery in Databases. AAAI/MIT Press, 1991
- S. M. Weiss and N. Indurkhya, Predictive Data Mining, Morgan Kaufmann, 1998
- I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2001

Thank you !!!