Data Mining for Security Applications

Prof. Bhavani Thuraisingham

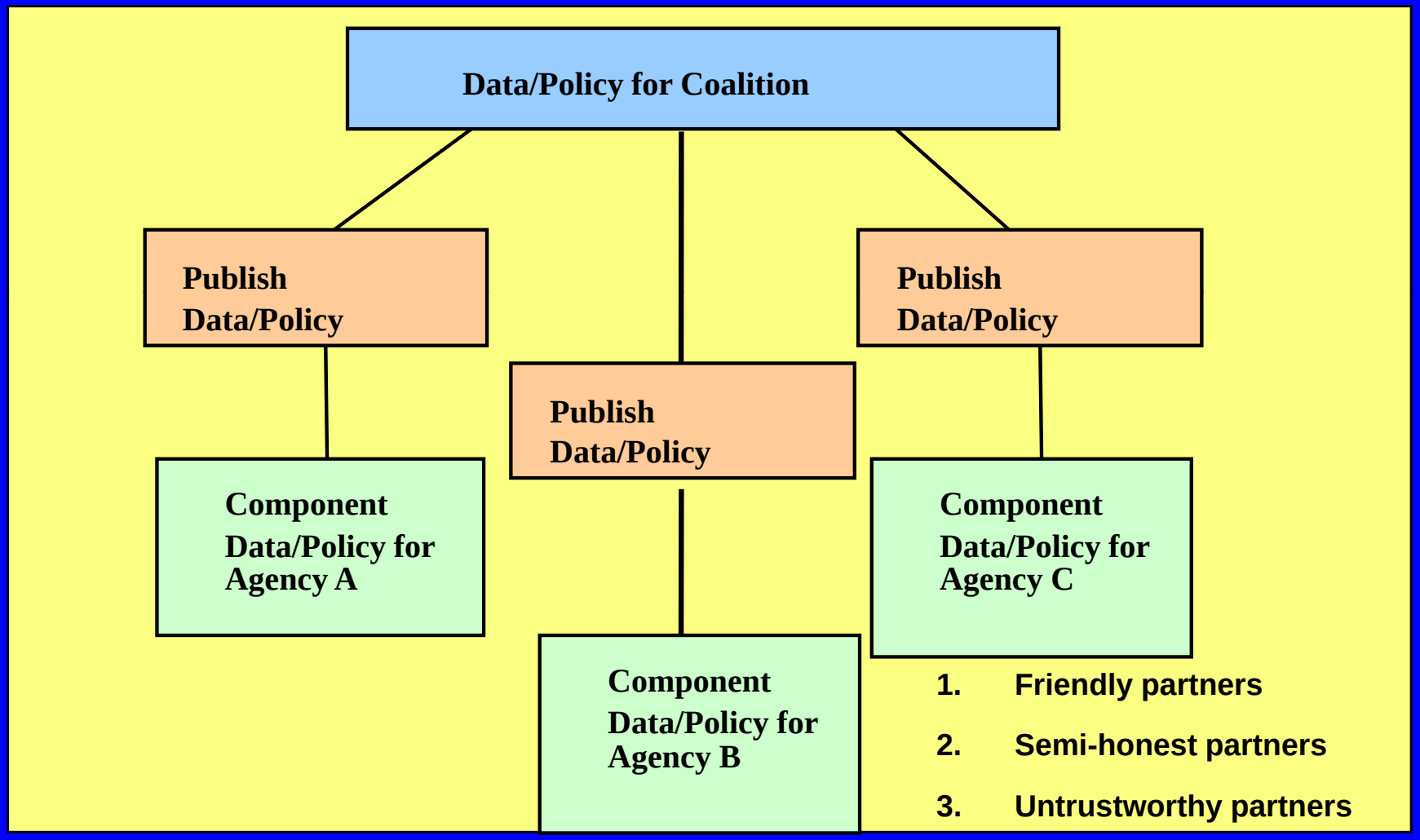
The University of Texas at Dallas

June 2006

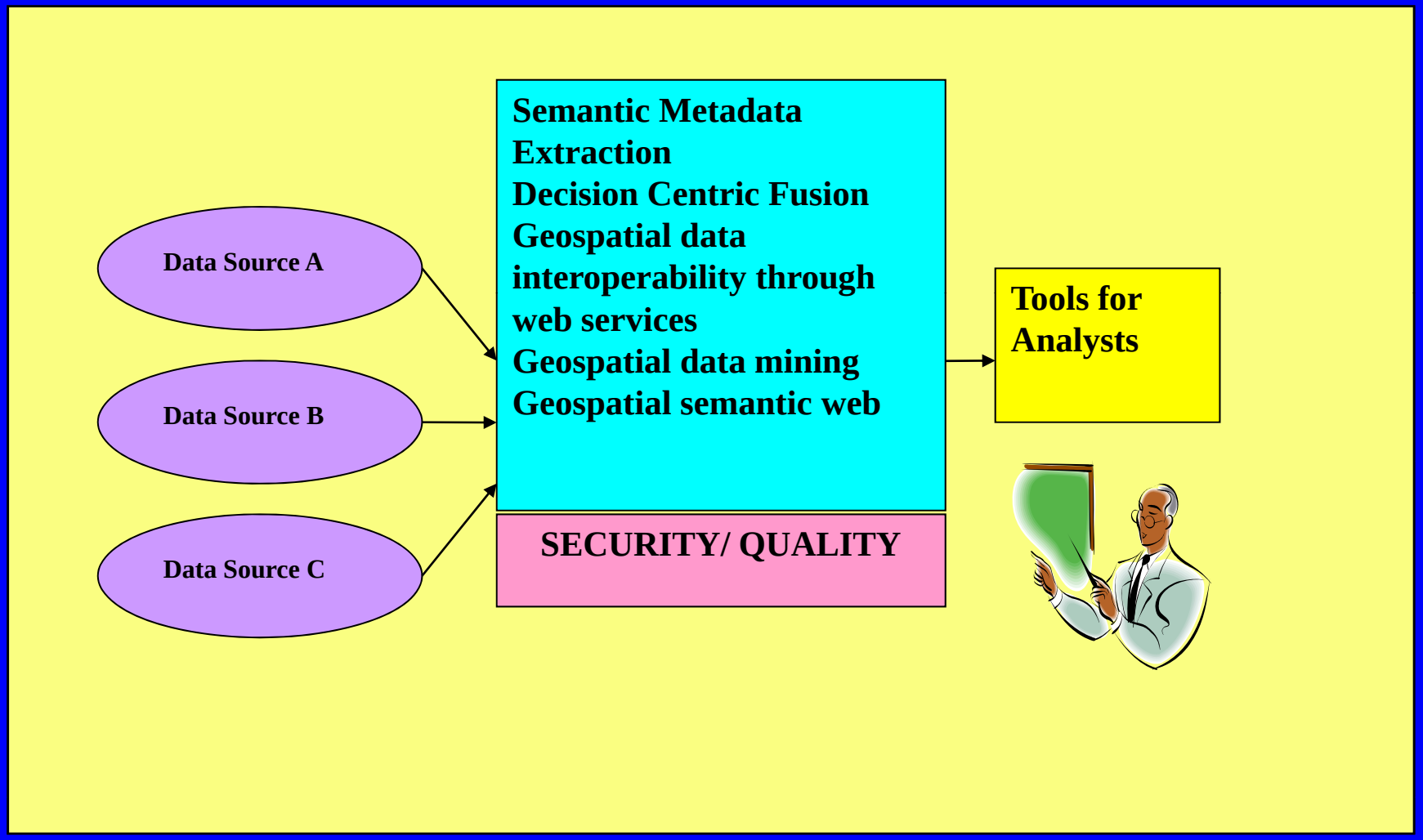
Outline

- Vision at U. Texas at Dallas on Assured Information Sharing
- Overview of Data Mining
- Data Mining for Cyber security applications (Joint work with Prof. Latifur Khan at the University of Texas at Dallas)
  - Intrusion Detection
  - Data Mining for Firewall Policy Management
  - Data Mining for Worm Detection
    - Malicious Executables
- Data Mining for National Security Applications
  - Non real-time and real-time threats
  - Surveillance
- Privacy and Data Mining
- Example Projects and Technology Transfer at UT Dallas.

Assured Information Sharing (AIS)



Geospatial Data Management for AIS



What is Data Mining?

Information Harvesting

Data Mining

Knowledge Mining

Knowledge Discovery in Databases

Data Dredging

Data Archaeology

Data Pattern Processing

Database Mining

Knowledge Extraction

Siftware

The process of discovering meaningful new correlations, patterns, and trends by sifting through large amounts of data, often previously unknown, using pattern recognition technologies and statistical and mathematical techniques (Thuraisingham, Data Mining, CRC Press 1998)

What's going on in data mining?

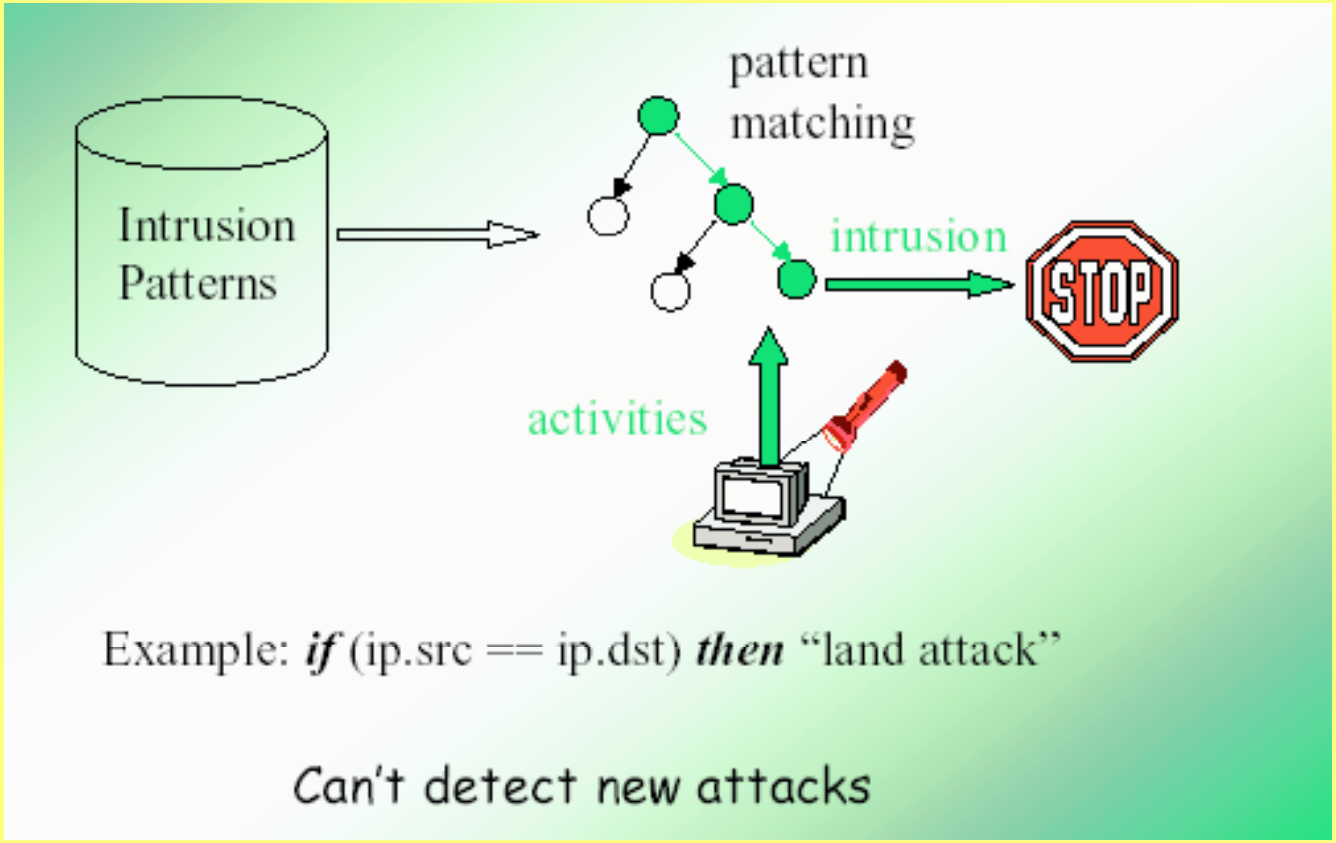
- What are the technologies for data mining?
  - Database management, data warehousing, machine learning, statistics, pattern recognition, visualization, parallel processing
- What can data mining do for you?
  - Data mining outcomes: Classification, Clustering, Association, Anomaly detection, Prediction, Estimation, . . .
- How do you carry out data mining?
  - Data mining techniques: Decision trees, Neural networks, Market-basket analysis, Link analysis, Genetic algorithms, . . .
- What is the current status?
  - Many commercial products mine relational databases
- What are some of the challenges?
  - Mining unstructured data, extracting useful patterns, web mining, Data mining, security and privacy

Data Mining for Intrusion Detection: Problem

- An intrusion can be defined as “any set of actions that attempt to compromise the integrity, confidentiality, or availability of a resource”.
- Attacks are:
  - Host-based attacks
  - Network-based attacks
- Intrusion detection systems are split into two groups:
  - Anomaly detection systems
  - Misuse detection systems
- Use audit logs
  - Capture *all* activities in network and hosts.
  - But the amount of data is huge!

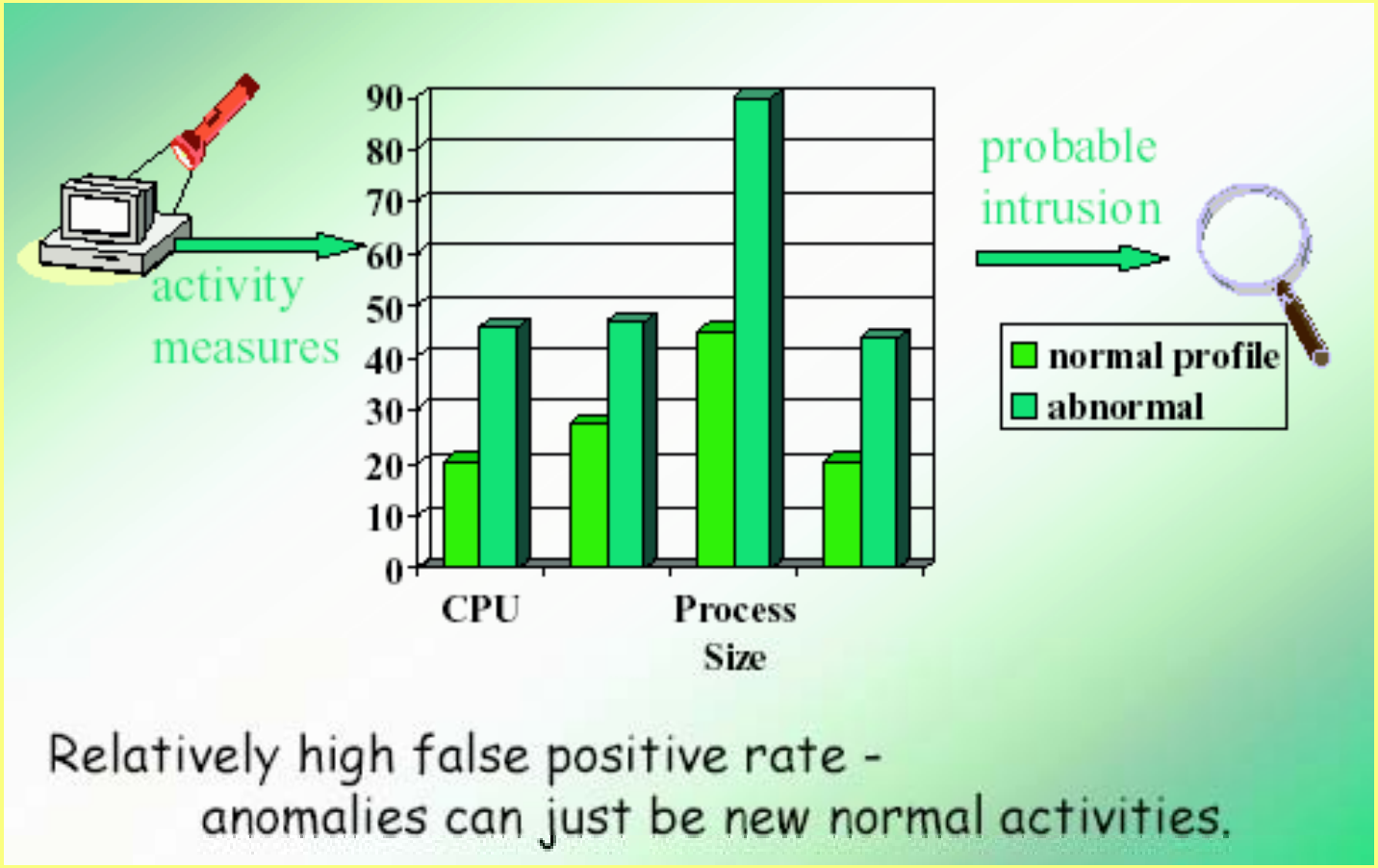
Misuse Detection

- **Misuse Detection**

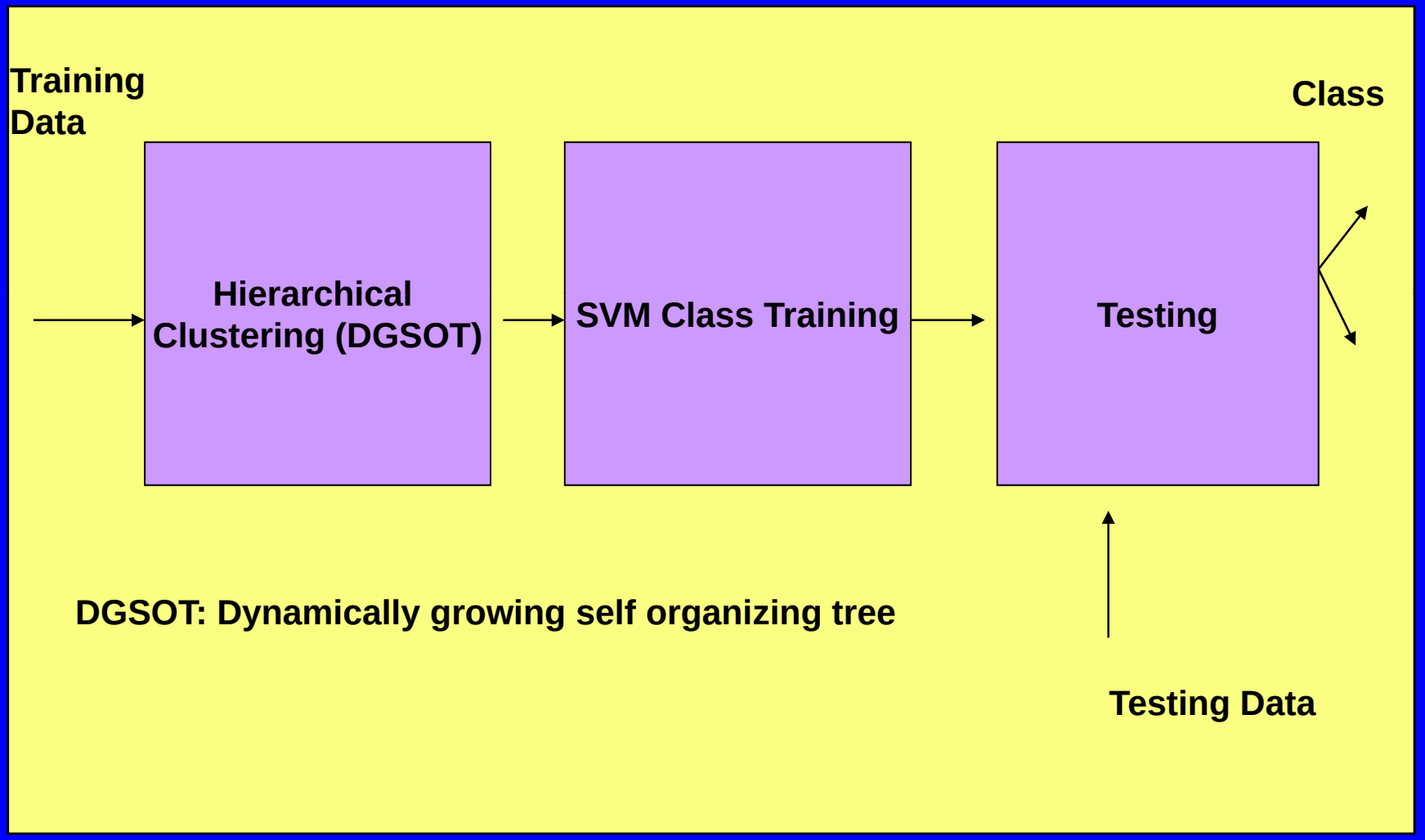


Problem: Anomaly Detection

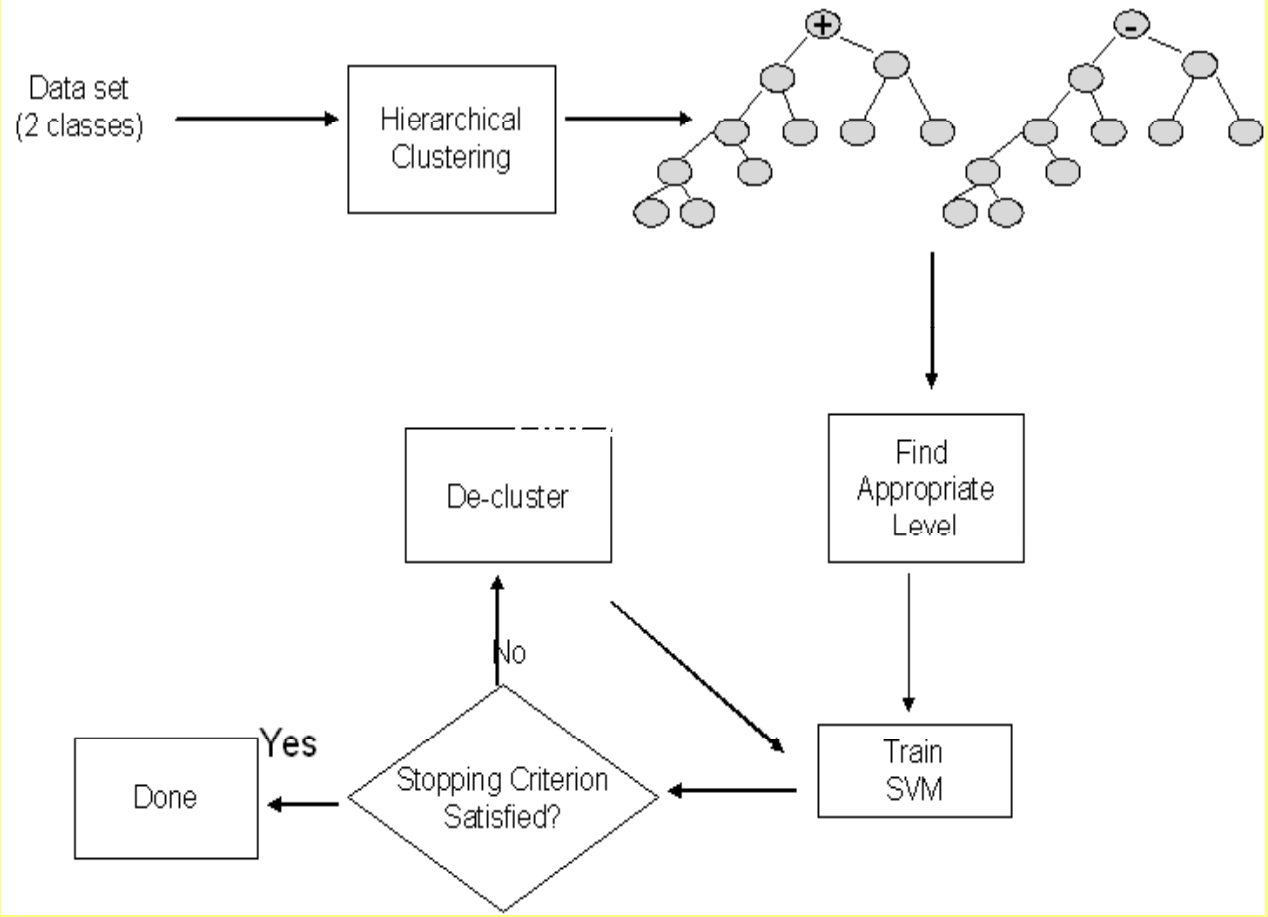
- **Anomaly Detection**



Our Approach: Overview



Our Approach: Hierarchical Clustering



Hierarchical clustering with SVM flow chart

Results

Training Time, FP and FN Rates of Various Methods

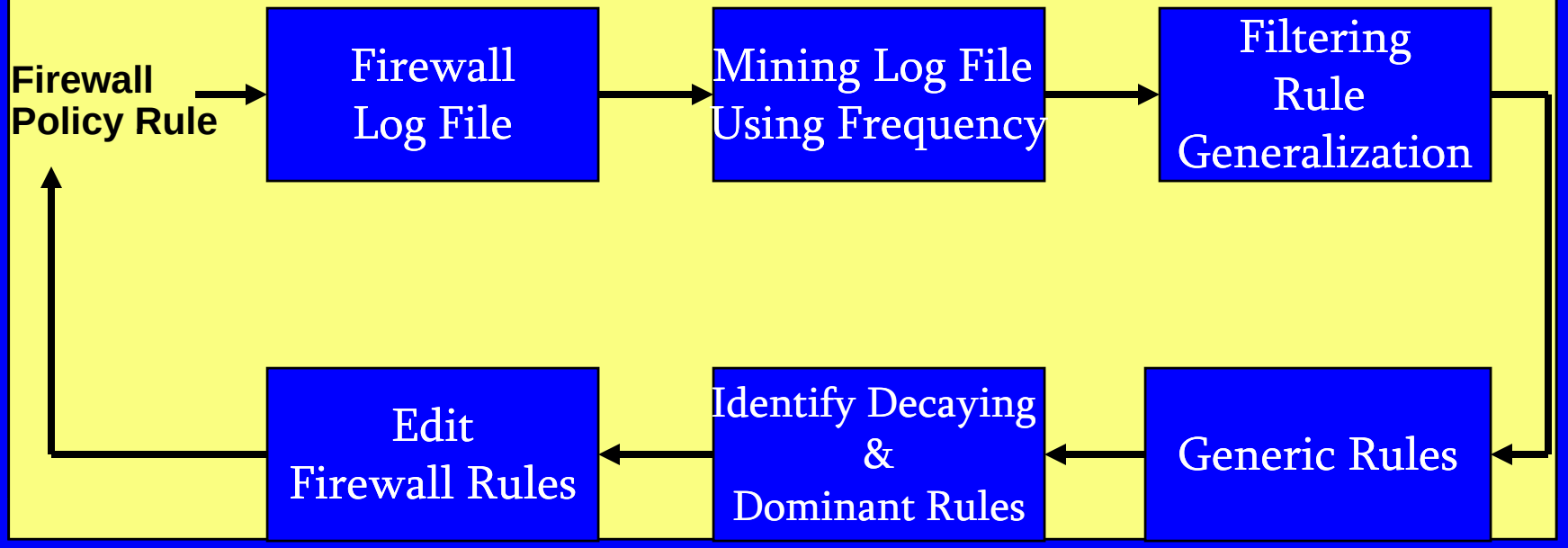
Methods	Average Accuracy	Total Training Time	Average FP Rate (%)	Average FN Rate (%)
Random Selection	52%	0.44 hours	40	47
Pure SVM	57.6%	17.34 hours	35.5	42
SVM+Rocchio Bundling	51.6%	26.7 hours	44.2	48
SVM + DGSOT	69.8%	13.18 hours	37.8	29.8

Analysis of Firewall Policy Rules Using Data Mining Techniques

- Firewall is the *de facto* core technology of today's network security
  - First line of defense against external network attacks and threats
  - Firewall controls or governs network access by allowing or denying the incoming or outgoing network traffic according to firewall policy rules.
  - Manual definition of rules often result in in anomalies in the policy
  - Detecting and resolving these anomalies manually is a tedious and an error prone task
- Solutions:
  - Anomaly detection:
    - Theoretical Framework for the resolution of anomaly; A new algorithm will simultaneously detect and resolve any anomaly that is present in the policy rules
    - Traffic Mining: Mine the traffic and detect anomalies

Traffic Mining

- **To bridge the gap between what is written in the firewall policy rules and what is being observed in the network is to analyze traffic and log of the packets– traffic mining**
  - **Network traffic trend may show that some rules are out-dated or not used recently**



Traffic Mining Results

1: TCP,INPUT,129.110.96.117,ANY,*.*.*.*,80,DENY
2: TCP,INPUT,*.*.*.*,ANY,*.*.*.*,80,ACCEPT
3: TCP,INPUT,*.*.*.*,ANY,*.*.*.*,443,DENY
4: TCP,INPUT,129.110.96.117,ANY,*.*.*.*,22,DENY
5: TCP,INPUT,*.*.*.*,ANY,*.*.*.*,22,ACCEPT
6: TCP,OUTPUT,129.110.96.80,ANY,*.*.*.*,22,DENY
7: UDP,OUTPUT,*.*.*.*,ANY,*.*.*.*,53,ACCEPT
8: UDP,INPUT,*.*.*.*,53,*.*.*.*,ANY,ACCEPT
9: UDP,OUTPUT,*.*.*.*,ANY,*.*.*.*,ANY,DENY
10: UDP,INPUT,*.*.*.*,ANY,*.*.*.*,ANY,DENY
11: TCP,INPUT,129.110.96.117,ANY,129.110.96.80,22,DENY
12: TCP,INPUT,129.110.96.117,ANY,129.110.96.80,80,DENY
13: UDP,INPUT,*.*.*.*,ANY,129.110.96.80,ANY,DENY
14: UDP,OUTPUT,129.110.96.80,ANY,129.110.10.*,ANY,DENY
15: TCP,INPUT,*.*.*.*,ANY,129.110.96.80,22,ACCEPT
16: TCP,INPUT,*.*.*.*,ANY,129.110.96.80,80,ACCEPT
17: UDP,INPUT,129.110.*.*,53,129.110.96.80,ANY,ACCEPT
18: UDP,OUTPUT,129.110.96.80,ANY,129.110.*.*,53,ACCEPT

Rule 1, Rule 2: ==> GENRERALIZATION
Rule 1, Rule 16: ==> CORRELATED
Rule 2, Rule 12: ==> SHADOWED
Rule 4, Rule 5: ==> GENRERALIZATION
Rule 4, Rule 15: ==> CORRELATED
Rule 5, Rule 11: ==> SHADOWED

Anomaly Discovery Result

Worm Detection: Introduction

- What are worms?
  - Self-replicating program; Exploits software vulnerability on a victim; Remotely infects other victims
- Evil worms
  - Severe effect; Code Red epidemic cost $2.6 Billion
- Goals of worm detection
  - Real-time detection
- Issues
  - Substantial Volume of Identical Traffic, Random Probing
- Methods for worm detection
  - Count number of sources/destinations; Count number of failed connection attempts
- Worm Types
  - Email worms, Instant Messaging worms, Internet worms, IRC worms, File-sharing Networks worms
- Automatic signature generation possible
  - EarlyBird System (S. Singh -UCSD); Autograph (H. Ah-Kim - CMU)

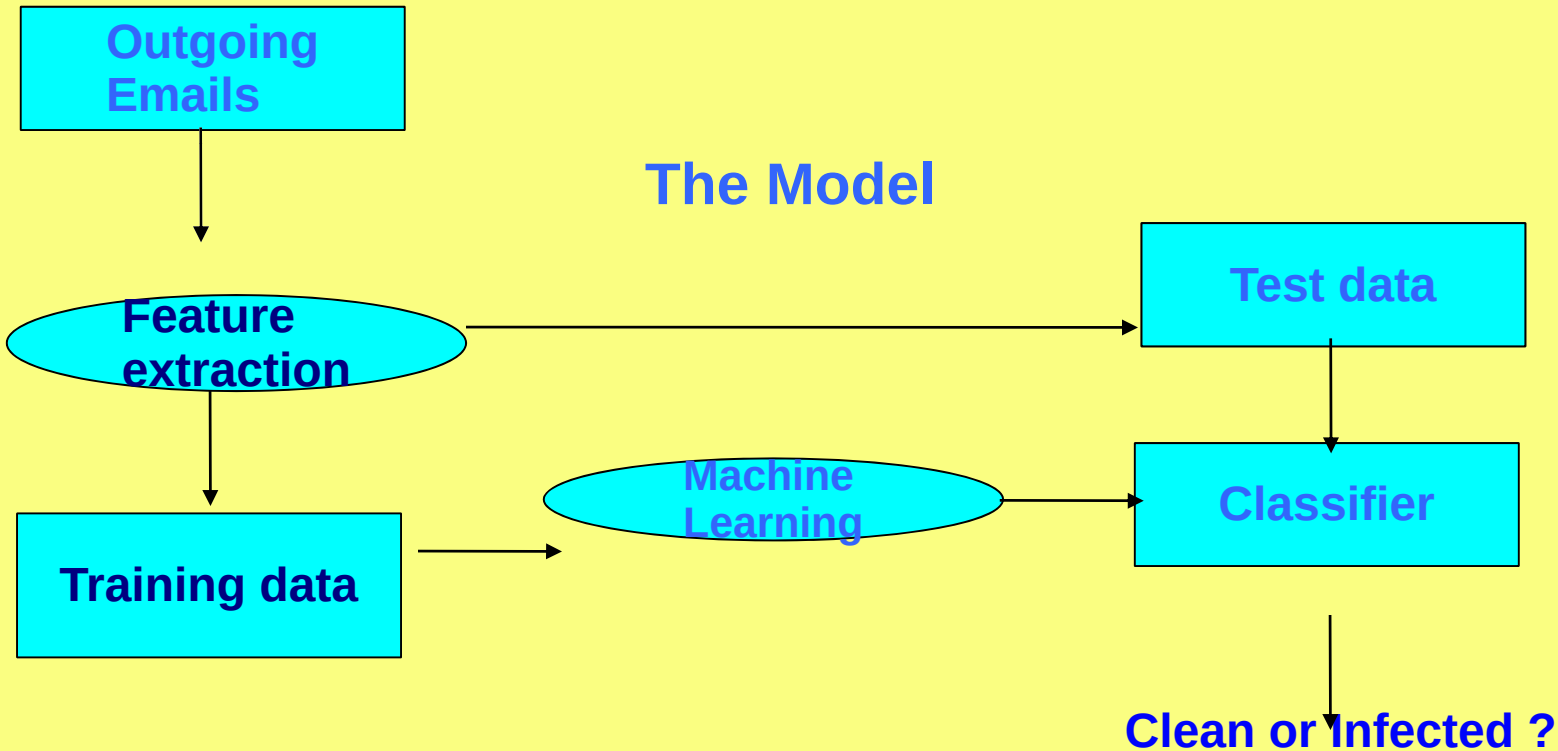
Email Worm Detection using Data Mining

Task:

given some training instances of both "normal" and "viral" emails, induce a hypothesis to detect "viral" emails.

We used:

Naïve Bayes
SVM



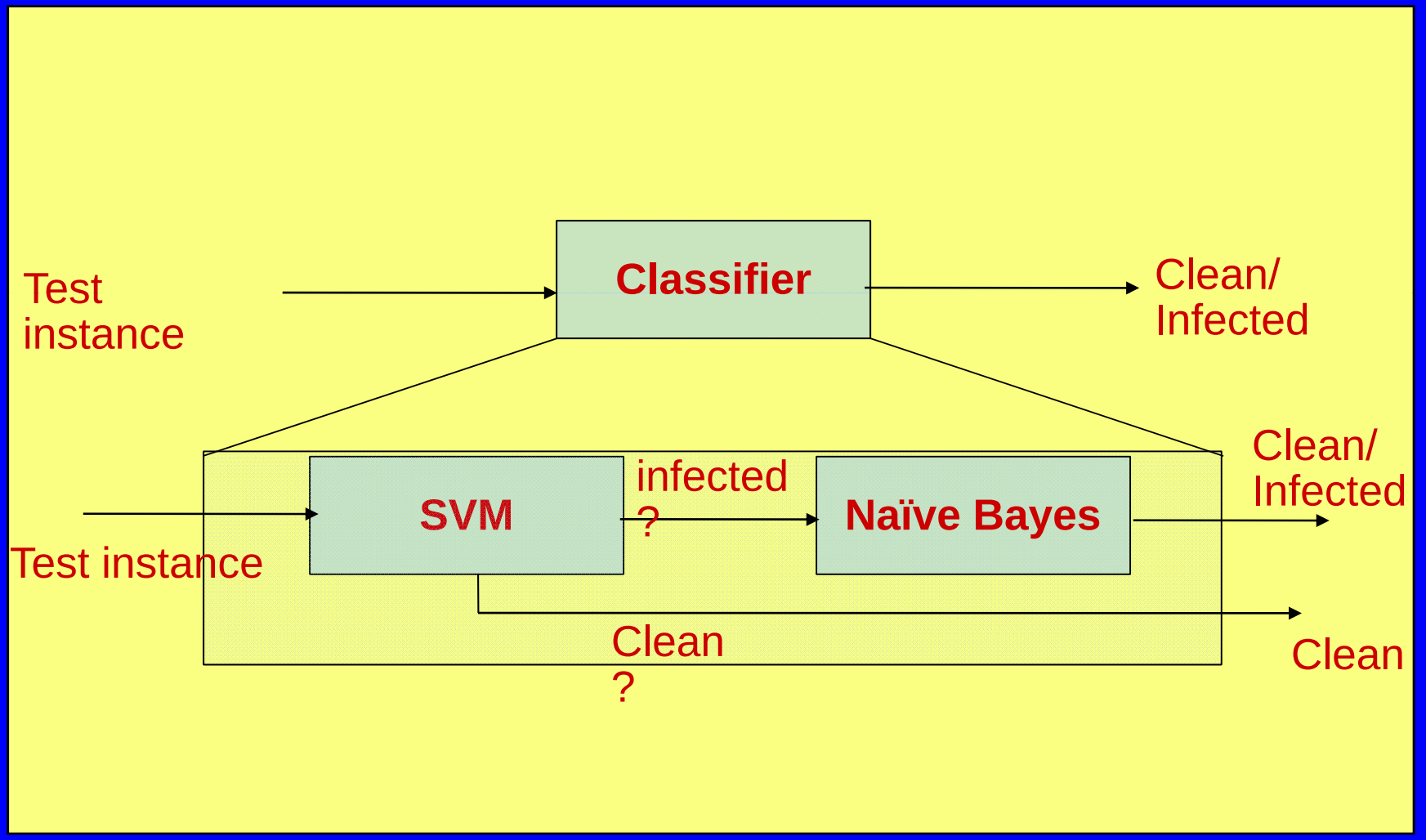
Assumptions

- Features are based on outgoing emails.
- Different users have different “normal” behaviour.
- Analysis should be per-user basis.
- Two groups of features
  - Per email (#of attachments, HTML in body, text/binary attachments)
  - Per window (mean words in body, variable words in subject)
- Total of 24 features identified
- Goal: Identify “normal” and “viral” emails based on these features

Feature sets

- **Per email features**
  - = **Binary valued Features**
    - **Presence of HTML; script tags/attributes; embedded images; hyperlinks;**
    - **Presence of binary, text attachments; MIME types of file attachments**
  - = **Continuous-valued Features**
    - **Number of attachments; Number of words/characters in the subject and body**
- **Per window features**
  - = **Number of emails sent; Number of unique email recipients; Number of unique sender addresses; Average number of words/characters per subject, *body; average word length:;* Variance in number of words/characters per subject, body; Variance in word length**
  - = **Ratio of emails with attachments**

Data Mining Approach



Data set

- Collected from UC Berkeley.
  - Contains instances for both normal and viral emails.
- Six worm types:
  - bagle.f, bubbleboy, mydoom.m,
  - mydoom.u, netsky.d, sobig.f
- Originally Six sets of data:
  - training instances: normal (400) + five worms (5x200)
  - testing instances: normal (1200) + the sixth worm (200)
- Problem: Not balanced, no cross validation reported
- Solution: re-arrange the data and apply cross-validation

Our Implementation and Analysis

- **Implementation**
  - **Naïve Bayes: Assume "Normal" distribution of numeric and real data; smoothing applied**
  - **SVM: with the parameter settings: one-class SVM with the radial basis function using "gamma" = 0.015 and "nu" = 0.1.**
- **Analysis**
  - **NB alone performs better than other techniques**
  - **SVM alone also performs better if parameters are set correctly**
  - **mydoom.m and VBS.Bubbleboy data set are not sufficient (very low detection accuracy in all classifiers)**
  - **The feature-based approach seems to be useful only when we have**

    identified the relevant features

    gathered enough training data

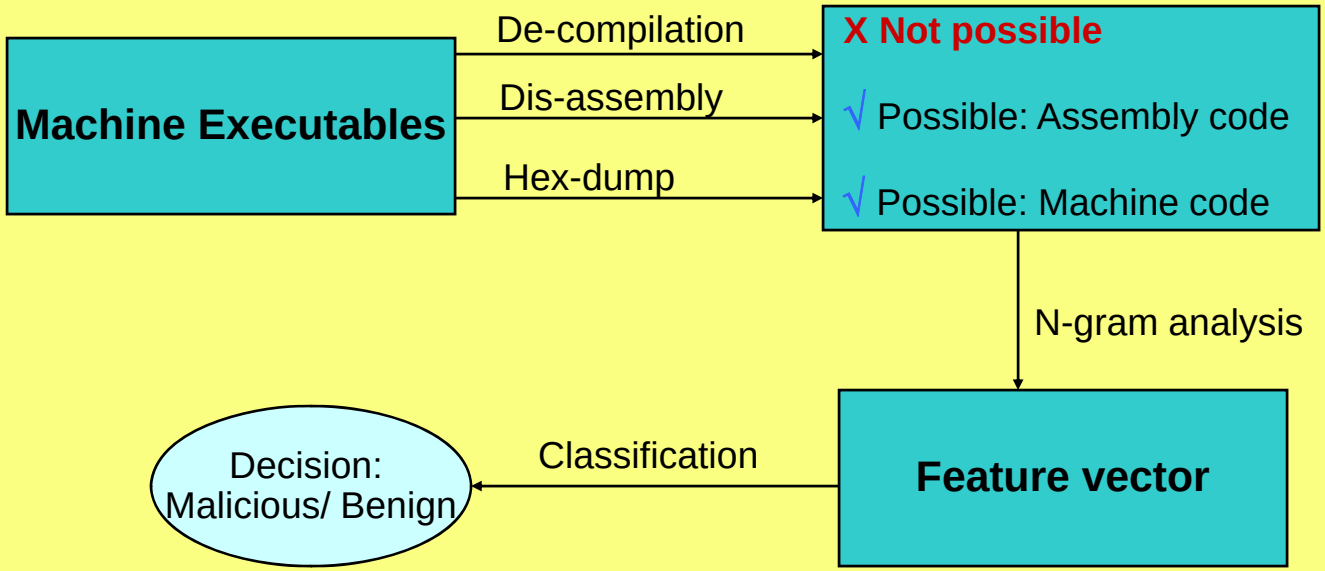
    Implement classifiers with best parameter settings

Detecting Malicious Executables: Recent Approaches

- Content-based techniques
  - Signature detection by human
  - Automated signature detection
  - Data mining
- But these approaches are based on analysis of the machine code
  - Hard to come up with meaningful patterns

Detecting Malicious Executables. New Ideas

- ✗ **If we could re-generate the source code**
  We could easily detect malicious codes, But this is not possible with current state-of-the art.
- ✓ **But we can re-generate the assembly code**
  From most of the executables, Analyze the code, Apply data mining to detect malicious code (New!)



Detecting Malicious Executables ..Testing New Ideas

- Collected 838 Malicious and 597 Benign executables
  - Malicious code from VX Heavens (http://vx.netlux.org)
- Dis-assembled into assembly code
  - Using s/w found in http://www.geocities.com/~sangcho/index.html
- Computed n-grams of assembly instructions
- Used n-grams as features
- Applied data mining techniques to determine classification accuracies

Detecting Malicious Executables ..Results

Comparison with "n-gram mining" of machine code

Classifiers used (from WEKA)	Assembly Instruction 1-gram feature accuracy (%)	Machine Instruction 1-gram feature accuracy (%)
Support Vector Machine	**89.34**	**85.88**
Naïve Bayes	**64.95**	**68.57**
ID3	**88.15**	**82.20**
BAGGING	**89.97**	**86.37**
Adaboosting	**82.65**	**79.28**
Average	**83.01**	**80.46**

Other Applications of Data Mining in Security

- Insider Threat Analysis – both network/host and physical
- Fraud Detection
- Protecting children from inappropriate content on the Internet
- Digital Identity Management
- Detecting identity theft
- Biometrics identification and verification
- Digital Forensics
- National Security / Counter-terrorism
- Surveillance

Data Mining Needs for Counterterrorism: Non-real-time Data Mining

- Gather data from multiple sources
  - Information on terrorist attacks: who, what, where, when, how
  - Personal and business data: place of birth, ethnic origin, religion, education, work history, finances, criminal record, relatives, friends and associates, travel history, . . .
  - Unstructured data: newspaper articles, video clips, speeches, emails, phone records, . . .
- Integrate the data, build warehouses and federations
- Develop profiles of terrorists, activities/threats
- Mine the data to extract patterns of potential terrorists and predict future activities and targets
- Find the “needle in the haystack” - suspicious needles?
- Data integrity is important
- Techniques have to SCALE

Data Mining Needs for Counterterrorism: Real-time Data Mining

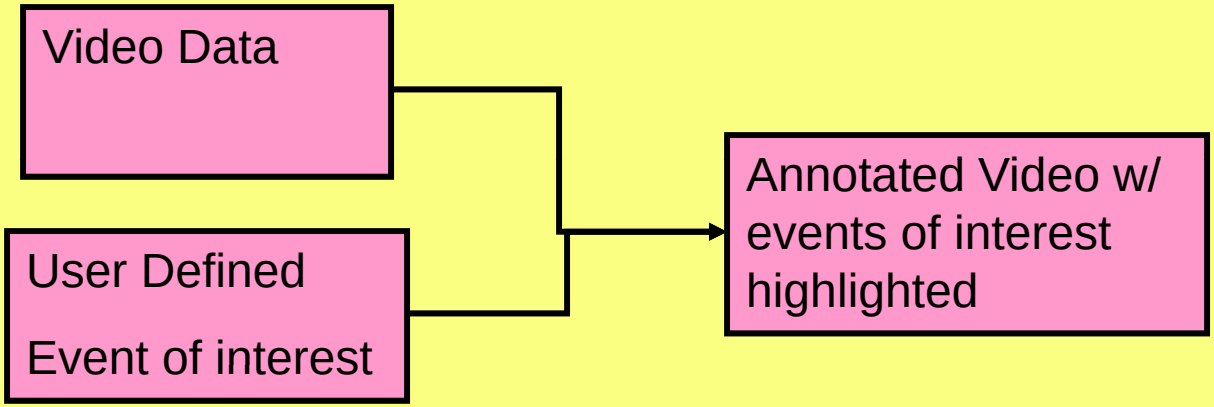
- Nature of data
  - Data arriving from sensors and other devices
    - =Continuous data streams
  - Breaking news, video releases, satellite images
  - Some critical data may also reside in caches
- Rapidly sift through the data and discard unwanted data for later use and analysis (non-real-time data mining)
- Data mining techniques need to meet timing constraints
- Quality of service (QoS) tradeoffs among timeliness, precision and accuracy
- Presentation of results, visualization, real-time alerts and triggers

Data Mining for Surveillance Problems Addressed

- Huge amounts of surveillance and video data available in the security domain
- Analysis is being done off-line usually using "Human Eyes"
- Need for tools to aid human analyst (pointing out areas in video where unusual activity occurs)

Semantic Gap

- **Using our proposed system:**
- **Greatly Increase video analysis efficiency**



The disconnect between the low-level features a machine sees when a video is input into it and the high-level semantic concepts (or events) a human being sees when looking at a video clip

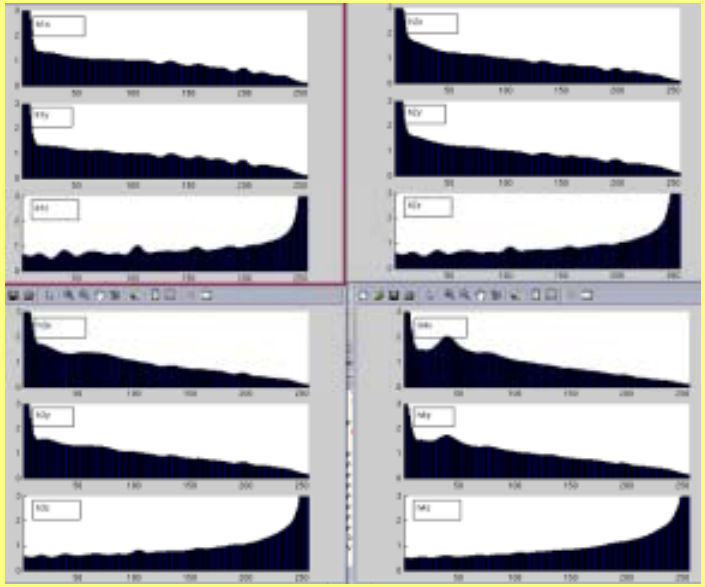
Low-Level features: color, texture, shape
High-level semantic concepts: presentation, newscast, boxing match

Our Approach

- **Event Representation**
  - **Estimate distribution of pixel intensity change**
- **Event Comparison**
  - **Contrast the event representation of different video sequences to determine if they contain similar semantic event content.**
- **Event Detection**
  - **Using manually labeled training video sequences to classify unlabeled video sequences**

Event Representation

- Measures the quantity and type of changes occurring within a scene
- A video event is represented as a set of x, y and t intensity gradient histograms over several temporal scales.
- Histograms are normalized and smoothed



Event Comparison and Detection

- Determine if the two video sequences contain similar high-level semantic concepts (events).
- Produces a number that indicates how close the two compared events are to one another.
- The lower this number is the closer the two events are.
  A robust event detection system should be able to
  - Recognize an event with reduced sensitivity to actor (e.g. clothing or skin tone) or background lighting variation.
  - Segment an unlabeled video containing multiple events into event specific segments

Labeled Video Events

- These events are manually labeled and used to classify unknown events

- Walking1 Running1 Waving2



Labeled Video Events

	walking 1	walking 2	walking 3	running 1	running 2	running 3	running 4	waving 2
walking 1	0	0.27625	0.24508	1.2262	1.383	0.97472	1.3791	10.961
walking 2	0.27625	0	0.17888	1.4757	1.5003	1.2908	1.541	10.581
walking 3	0.24508	0.17888	0	1.1298	1.0933	0.88604	1.1221	10.231
running 1	1.2262	1.4757	1.1298	0	0.43829	0.30451	0.39823	14.469
running 2	1.383	1.5003	1.0933	0.43829	0	0.23804	0.10761	15.05
running 3	0.97472	1.2908	0.88604	0.30451	0.23804	0	0.20489	14.2
running 4	1.3791	1.541	1.1221	0.39823	0.10761	0.20489	0	15.607
waving2	10.961	10.581	10.231	14.469	15.05	14.2	15.607	0

Example Experiment

Problem: Recognize and classify events irrespective of direction (right-to-left, left-to-right) and with reduced sensitivity to spatial variations (Clothing)

“Disguised Events”- Events similar to testing data except subject is dressed differently Compare Classification to “Truth” (Manual Labeling)

Disguised Walking 1

walking1	walking2	walking3	running1	running2	running3	running4	waving2
0.97653	0.45154	0.59608	1.5476	1.4633	1.5724	1.5406	12.225

Classification: Walking

Video Analysis Tool

- Using the event detection scheme we generate a video description document detailing the event composition of a specific video sequence
- This XML document annotation may be replaced by a more robust computer-understandable format (e.g. the VEML video event ontology language). Takes annotation document as input and organizes the corresponding video segment accordingly.
- Functions as an aid to a surveillance analyst searching for “Suspicious” events within a stream of video data.
- Activity of interest may be defined dynamically by the analyst during the running of the utility and flagged for analysis.

Directions

- Enhancements to the work
  - Working toward bridging the semantic gap and enabling more efficient video analysis
  - More rigorous experimental testing of concepts
  - Refine event classification through use of multiple machine learning algorithm (e.g. neural networks, decision trees, etc…). Experimentally determine optimal algorithm.
- Develop a model allowing definition of simultaneous events within the same video sequence
- Security and Privacy
  - Define an access control model that will allow access to surveillance video data to be restricted based on semantic content of video objects
  - Biometrics applications
  - Privacy preserving surveillance

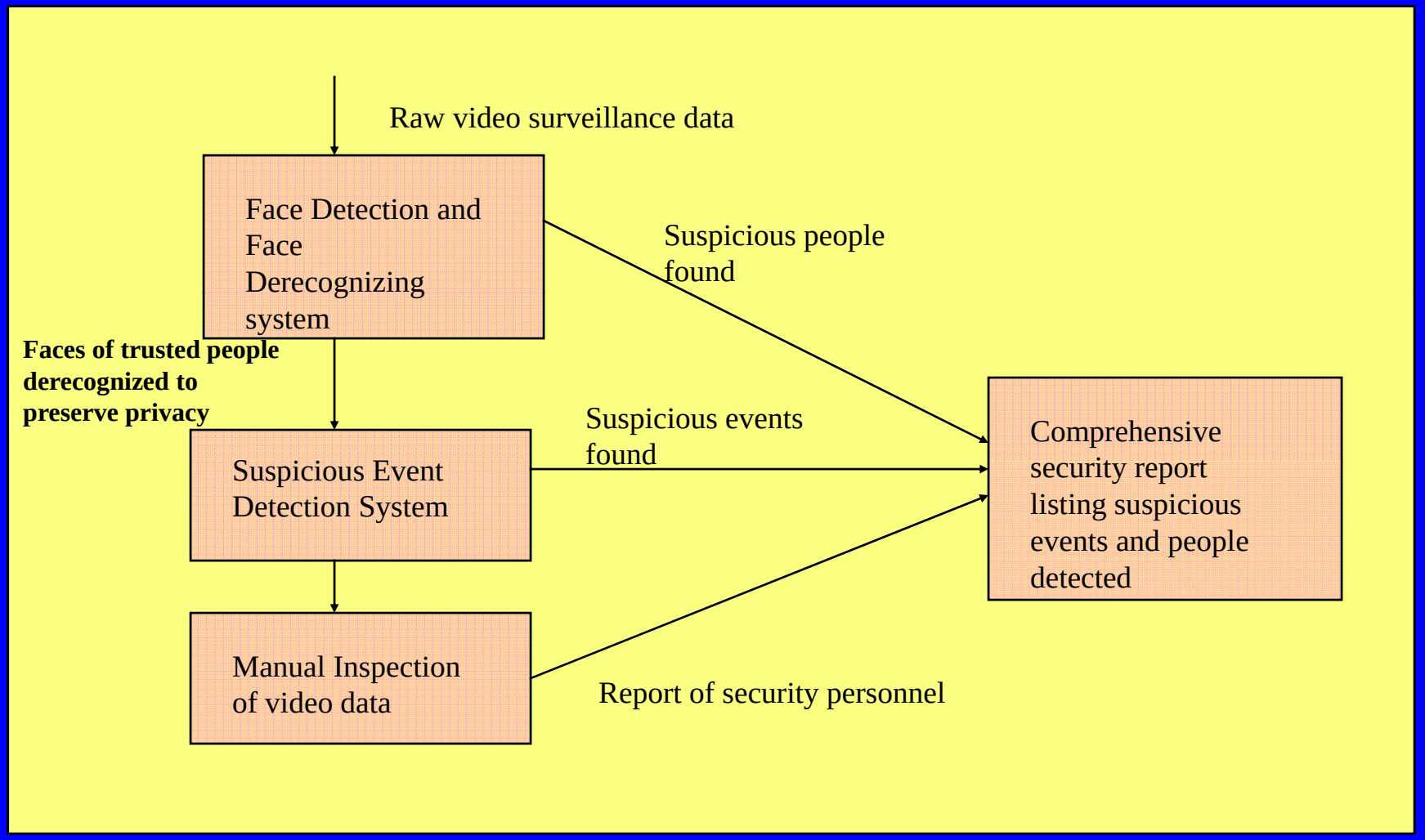
Data Mining as a Threat to Privacy

- Data mining gives us “facts” that are not obvious to human analysts of the data
- Can general trends across individuals be determined without revealing information about individuals?
- Possible threats:
  - Combine collections of data and infer information that is private
    - Disease information from prescription data
    - Military Action from Pizza delivery to pentagon
- Need to protect the associations and correlations between the data that are sensitive or private

Some Privacy Problems and Potential Solutions

- Problem: Privacy violations that result due to data mining
  - Potential solution: Privacy-preserving data mining
- Problem: Privacy violations that result due to the Inference problem
  - Inference is the process of deducing sensitive information from the legitimate responses received to user queries
  - Potential solution: Privacy Constraint Processing
- Problem: Privacy violations due to un-encrypted data
  - Potential solution: Encryption at different levels
- Problem: Privacy violation due to poor system design
  - Potential solution: Develop methodology for designing privacy-enhanced systems

Surveillance and Privacy



Data Mining and Privacy: Friends or Foes?

- They are neither friends nor foes
- Need advances in both data mining and privacy
- Data mining is a tool to be used by analysis and decision makers
  - Due to also positives and false negatives, need human in the loop
- Need to design flexible systems
  - Data mining has numerous applications including in security
  - For some applications one may have to focus entirely on “pure” data mining while for some others there may be a need for “privacy-preserving” data mining
  - Need flexible data mining techniques that can adapt to the changing environments
- Technologists, legal specialists, social scientists, policy makers and privacy advocates MUST work together

Example Projects at the University of Texas at Dallas

- Assured Information Sharing
  - Secure Semantic Web Technologies
  - Social Networks and Game theory applications
  - Privacy/Security Preserving Data Mining
- Geospatial Data Management
  - Geospatial data mining and data security
  - Geospatial semantic webs
- Data Mining for National Security
  - Suspicious Event Detention
  - Privacy preserving Surveillance
  - Automatic Face Detection
- Cross Cutting Themes
  - Data Mining for Security Applications (e.g., Intrusion detection, Mining Arabic Documents); Dependable Systems; Secure Grids

Some Technology Transfer Activities

- **AIS**
  - **Working with Collin County (near Dallas TX) to transfer AIS research to an operational Fusion Center for Emergency Management**
  - **Will Work with AFOSR to transfer the AIS technology to services and the GIG**
- **Geospatial Data**
  - **Contract with Raytheon IIS Division for Geospatial data management research; Partnership with Raytheon to transfer technology to operational programs**
  - **MOU between OGC, Raytheon and Oracle for Interoperability Experiments**
- **Data Mining for National Security**
  - **Working on DHS project with ADB Consulting**