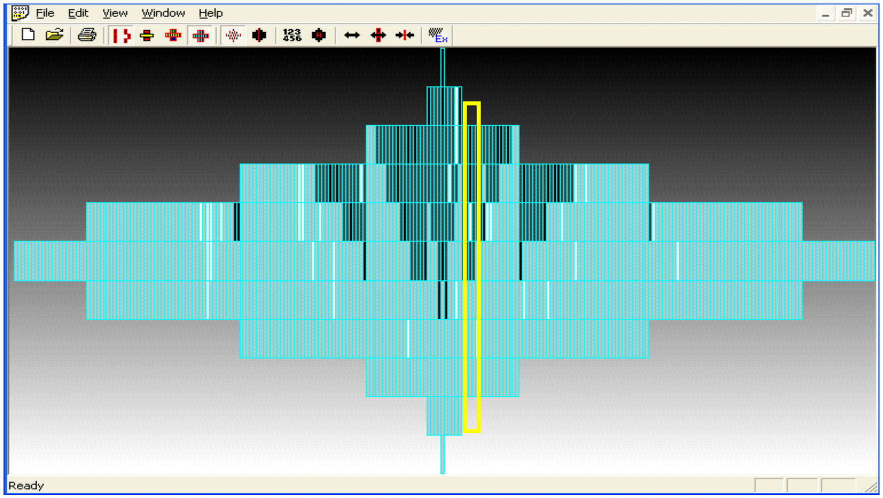
Visual Data Mining and Reasoning



Prof. Boris Kovalerchuk,

Dept. of Computer Science, Central Washington University, USA

Borisk@cwu.edu; *http://www.cwu.edu/~borisk*

Tutorial Notes

July 4, 2005

Abstract

Purpose. Why is the topic important?

Currently, with proliferation of visual techniques and hardware capabilities, visual data mining and reasoning has become feasible for a wide range of problems and large data sets. Visual data mining has several important advantages such as direct appeal to user understanding, and ability to show patterns that are extremely difficult to express analytically.

Learning objectives for the tutorial. What will participants learn?

This tutorial provides an overview of current techniques of visual data mining and reasoning, including newly developed iconic reasoning and visual monotone Boolean data mining. The tutorial will discuss current trends and future directions in the development of hybrid techniques of analytical and visual data mining.

This work has been partially supported by NGA and ARDA grants

Outline

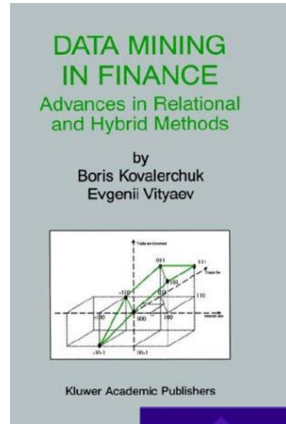
July 4, 2005

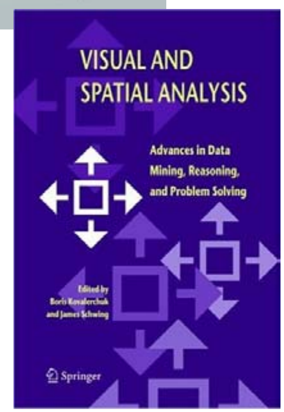
9:00 am – 12.00 pm

Break 10:00-10:30 am

Introduction

- **This tutorial discusses**
  - **capabilities and limitations of visual data mining and reasoning and**
  - **new opportunities that visual data mining and reasoning provide.**





- **The tutorial presents**
  - **Visual data mining methods conceptually and**
  - **demonstrate them with case studies**
    - **for breast cancer diagnostics based on X-ray mammographic images**
    - **visual correlation/link discovery for terrorism analysis and**
    - **social analysis.**

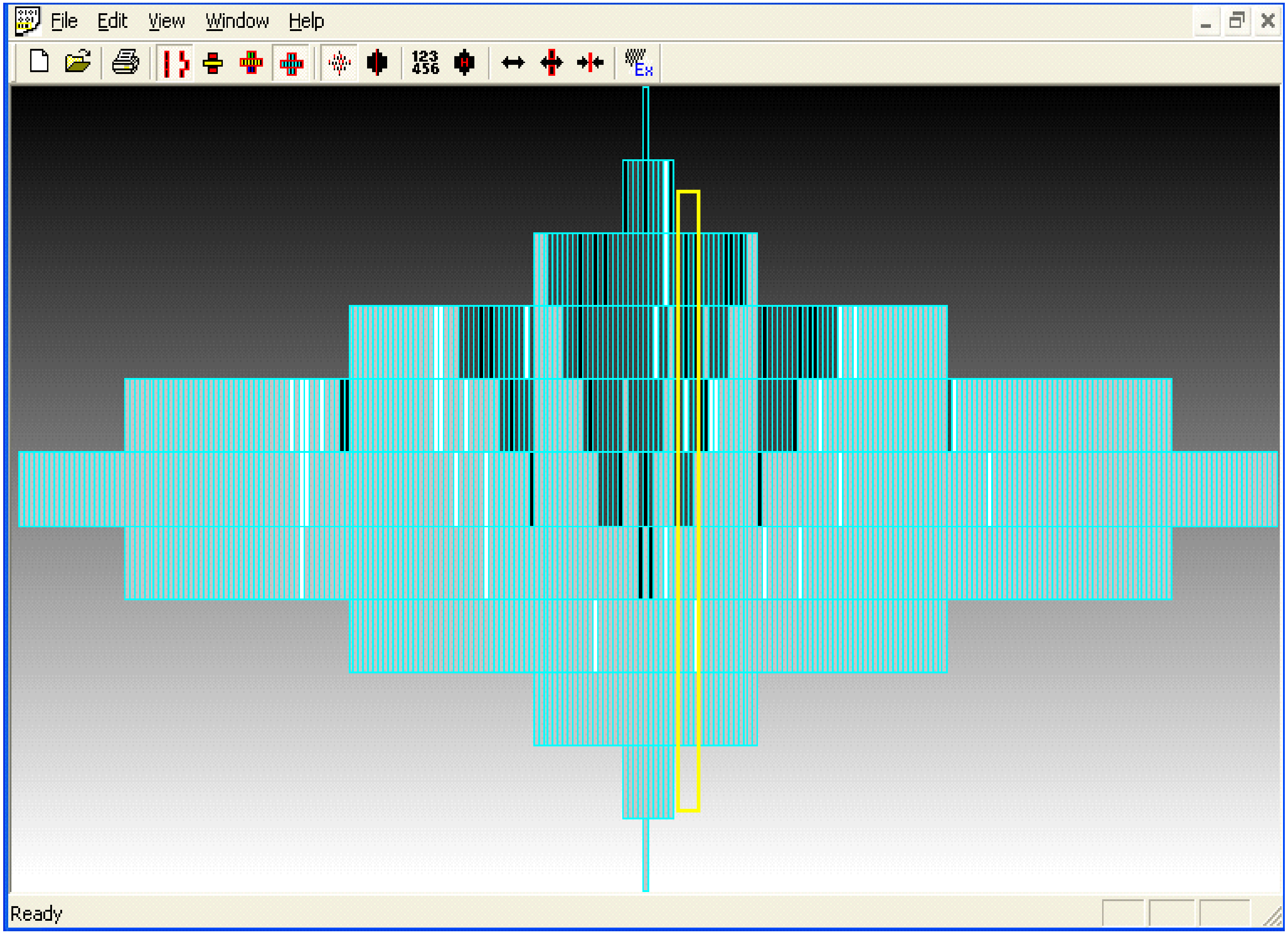
International Symposium of Visual Data Mining , July 6-8, 2005, London

- **Visual data mining research seeks to enhance the knowledge discovery process through**
  - **the use of graphical representations of data mining *results and processes* and**
  - **the *combination* of visual and computational approaches to data exploration.**
  - **By leveraging human perceptions of the visual space, patterns that might not otherwise be discovered will be identified through the use of visual data mining techniques.**

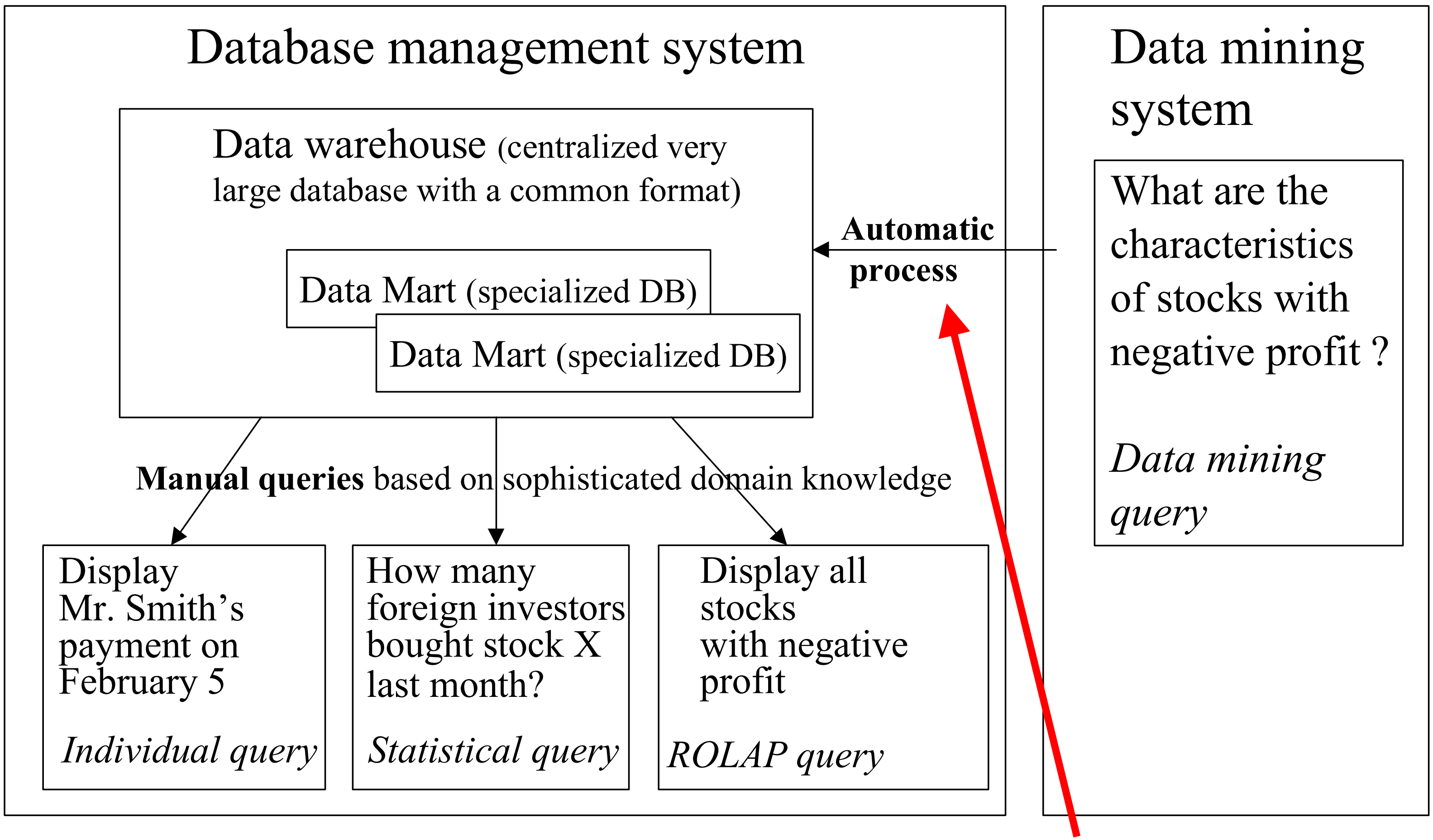
Topics of interest:

- **Data mining algorithm visualization**
- **Combining visual and computational methods of data exploration**
- **Temporal and spatial visual data mining**
- **Collaborative visualization and mining**
- **Evaluation of visual data mining methods**
- **Cognitive approaches and explanations for visual data mining**

1. Visual Data Mining Process

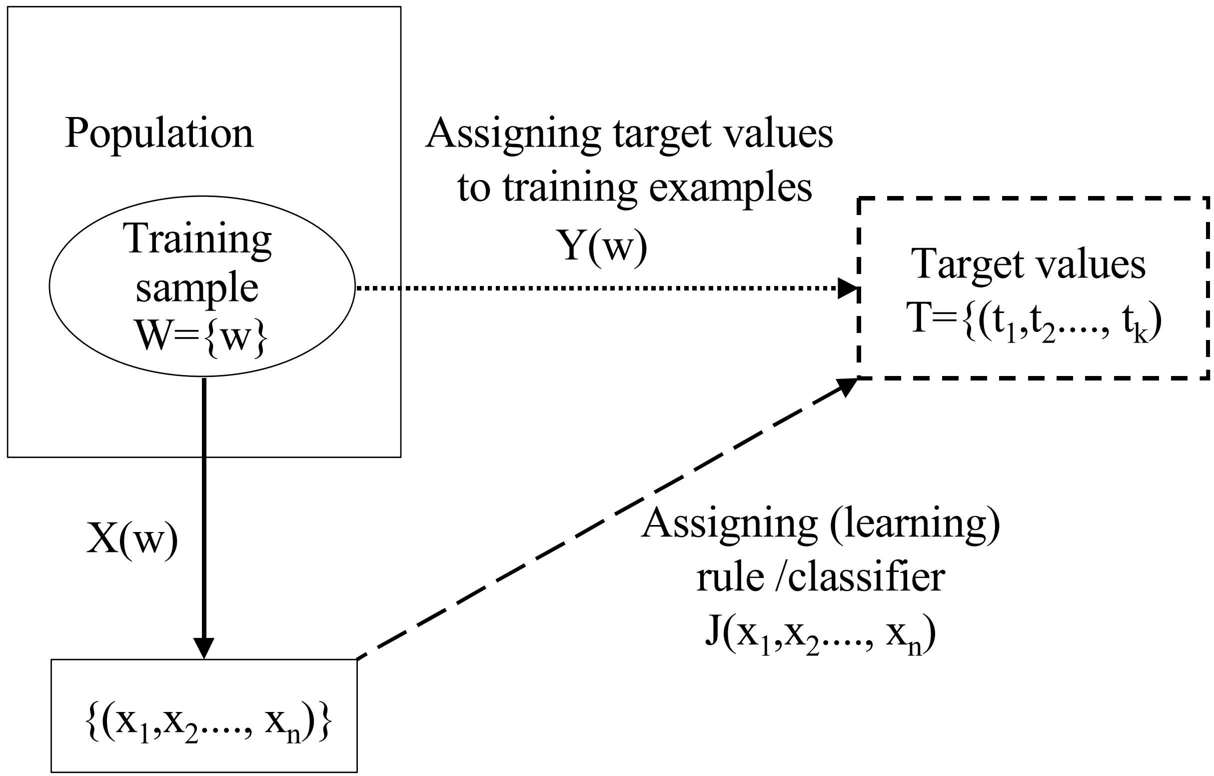


Database vs. data mining queries



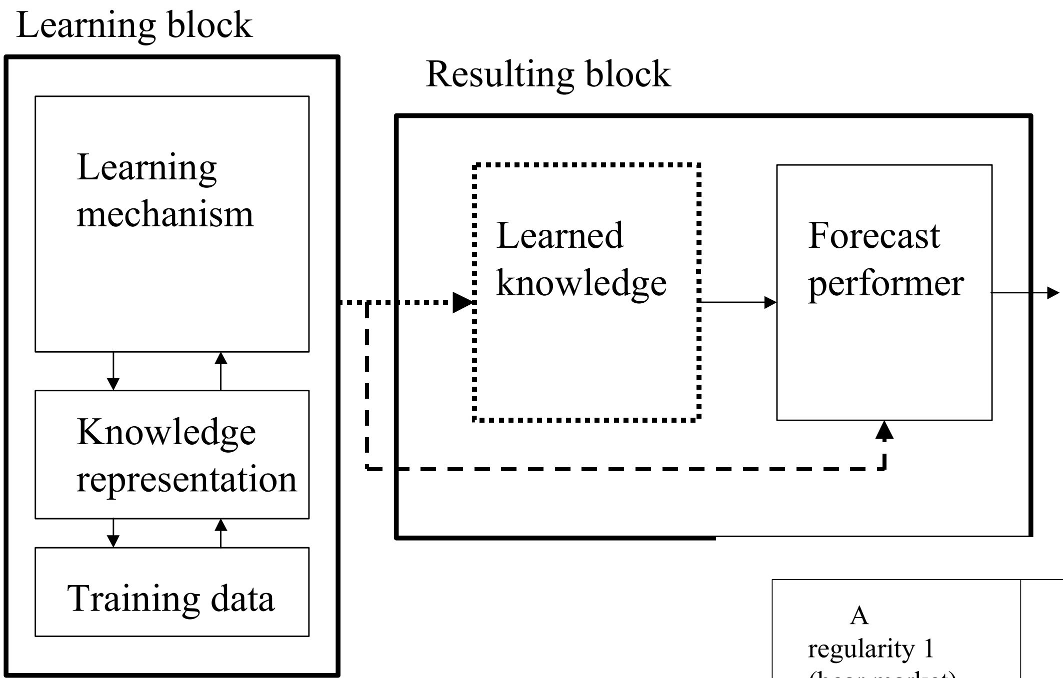
How realistic to have an adequate automatic DM process?
Visual DM is a realistic interactive DM alternative.
ROLAP- Relational Online Analytic Processing

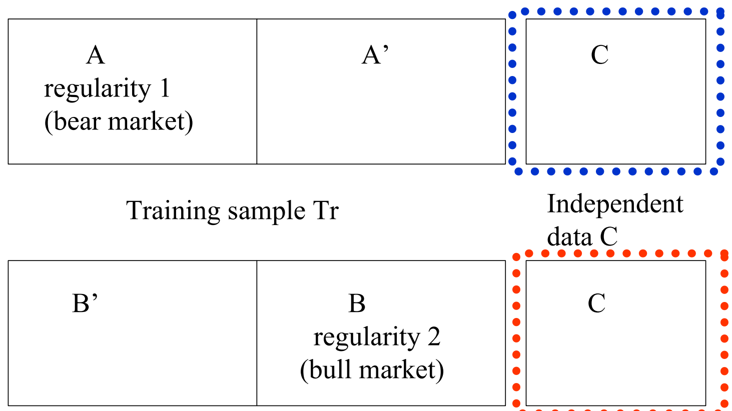
Data mining process with visual components



Representation training examples by **visual** descriptors: $(x_1,x_2\ldots, x_n)= X(w)$

Learning, testing and application process





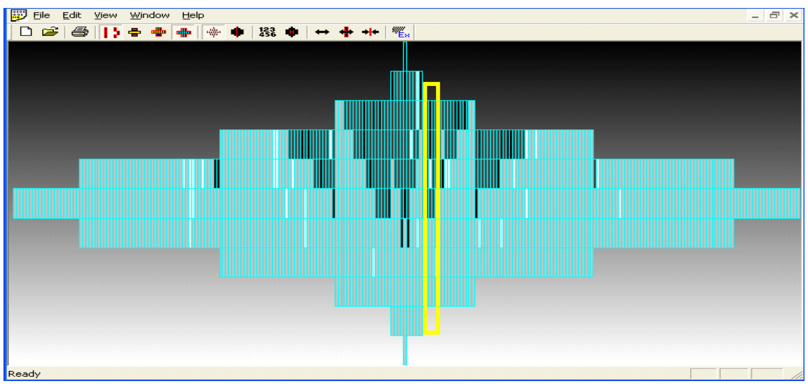
Testing data for non-stationary data

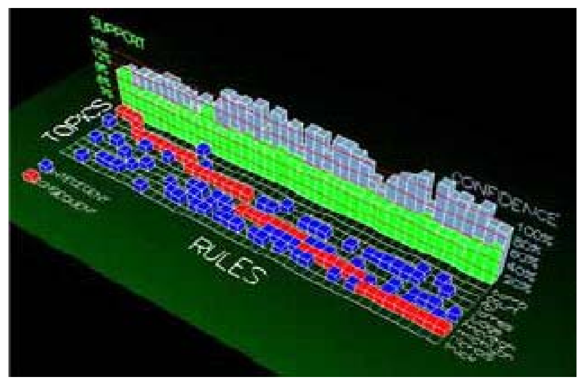
Analytical and visual learning

Data-based leaning vs. Visual knowledge-based learning

- **The** *data-based learning methods* **rely on training data as a major or sole source for discovering regularities.**
  - **Decision trees, neural networks and other general-purpose learning methods belong to this class.**
- *Visual and Knowledge-based methods* **rely on training data and** *prior knowledge* **(including human tacit knowledge) in variety of other forms, which is routinely ignored by the data-based methods.**
- **Ignoring other forms of knowledge is a major reason why it is difficult to interpret the internal structure of a classifier produced by the data-based methods**
- **Outside of physics, it is very hard to produce a explicit causal mechanism which actually generates training data**
- **This is the area where visual data mining can help.**

1.1. Visualization for Data Mining vs. Visual Data Mining





Visualization for Data Mining

A Definition of Visualization

- It is the Visual Interface to the data and mining tools
- It is a method of interacting with data and algorithms
  - supports the user through all the knowledge discovery steps
  - uses selections, queries, probes, and view transformations
- It is completely separable from the analysis methods
  - Data can be analyzed using many different algorithms
  - Each result can be viewed in a different visualization
  - Each visualization thus provides a different view of the results

Human Involvement

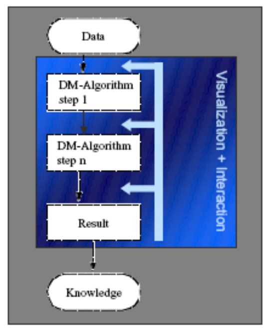
- When ?

☆ Right before the data mining step
- ⇨ Display initial data
- ⇨ Focus on/ narrow relevant search space

During the data mining step
- ⇨ Display intermediate results
- ⇨ Direct the search

After the data mining step
- ⇨ Display the result

Tightly Integrated Visualization (TIV)



- Visualization of algorithmic decisions
  - Data and patterns are better understood
  - User can make decisions based on perception
  - User can make decisions based on domain knowledge
  - Visualization of result enables user specified feedback for next algorithmic run or iteration

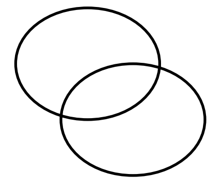
***Visual Web Mining* (VWM) is**
an application of Information Visualization techniques on results of Web Mining in order to further amplify the perception of extracted patterns and visually explore new ones in web domain.

Amir H. Youssefi, David J. Duke, Mohammed J. Zaki, Ephraim P. Glinert, **Visual Web Mining** 13th International World Wide Web Conference (poster proceedings), New York, NY, May 2004.
http://www.cs.rpi.edu/~zaki/PS/WWW04.pdf

Visual Data Mining

0 **The goal of *visual data mining* (VDM) is to help a user**

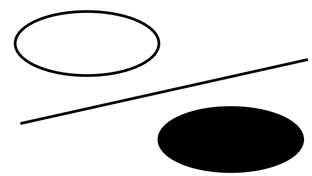
- **to get a feeling for the data,**
- **to detect interesting knowledge, and**
- **to gain a deep visual understanding of the data set** [Beilken & Spenke, 1999].



0 **One of especially important aspects of visual data mining is *visualizing the border* between patterns.**

0 **A visual result in which**

- **the border between patterns is simple and**
- **patterns are far away from each other**



matches our intuitive concept of the pattern and

serves as important support that the data mining result is robust and not accidental.

Limitations of traditional visualization

- The *subjectivity of the visual representation* can
  - cause different conclusions looking to the same data.
- The *poor scalability* of visual data analysis can
  - fail when representing hundreds of attributes.
- Humans *unable* to perceive more than six to eight dimensions on the same graph.
- The *slow speed* of manual interactive examination of the multi-dimensional, multi-color charts is a drawback.

Last & Kandel, 1999

Visualization for Data Mining vs. Visual Data Mining

Summary
Visual Data Mining Architectures

Benefits of Visualization/ Interaction

	Preceding Visualization	Subsequent Visualization	Tightly int. Visualization	Visual pattern discovery
Present/ display patterns		●	●	●
Search problem space with perception	●		●	●
Incorporate domain knowledge	●		●	●
Provide trust and understandability of patterns discovered analytically		●	●	●
Provide meaningful explainable patterns				●
Provide patterns that are too complex for analytical methods				●

Georges Grinstein, UMass Lowell – Daniel Keim, Univ. of Konstanz – Matt Ward, WPI 296

Analytical methods vs. Visual Methods or hybridization?

- **Human dimension –who is applying the method?**
  - **DM professional, CS student, Subject Matter Expert (SME)**
- **Data**
  - **highly noisy real data**
  - **the International CoIL data mining competition** [Putten, Someren, 2004]
- **CoIL participants**
  - **data mining professionals and university students**
    - **CWU and Free University in Amsterdam**
- **Surprising results with analytical methods**
  - **DM Professionals vs. undergraduate students**
- **Visual methods open the opportunity to do data mining by Subject Matter Experts directly**
  - **Winner combined parameters using subject matter knowledge, others just manipulated data as is.**

Numeric Data Mining

- CoIL surprising results [Putten, Someren, 2004].

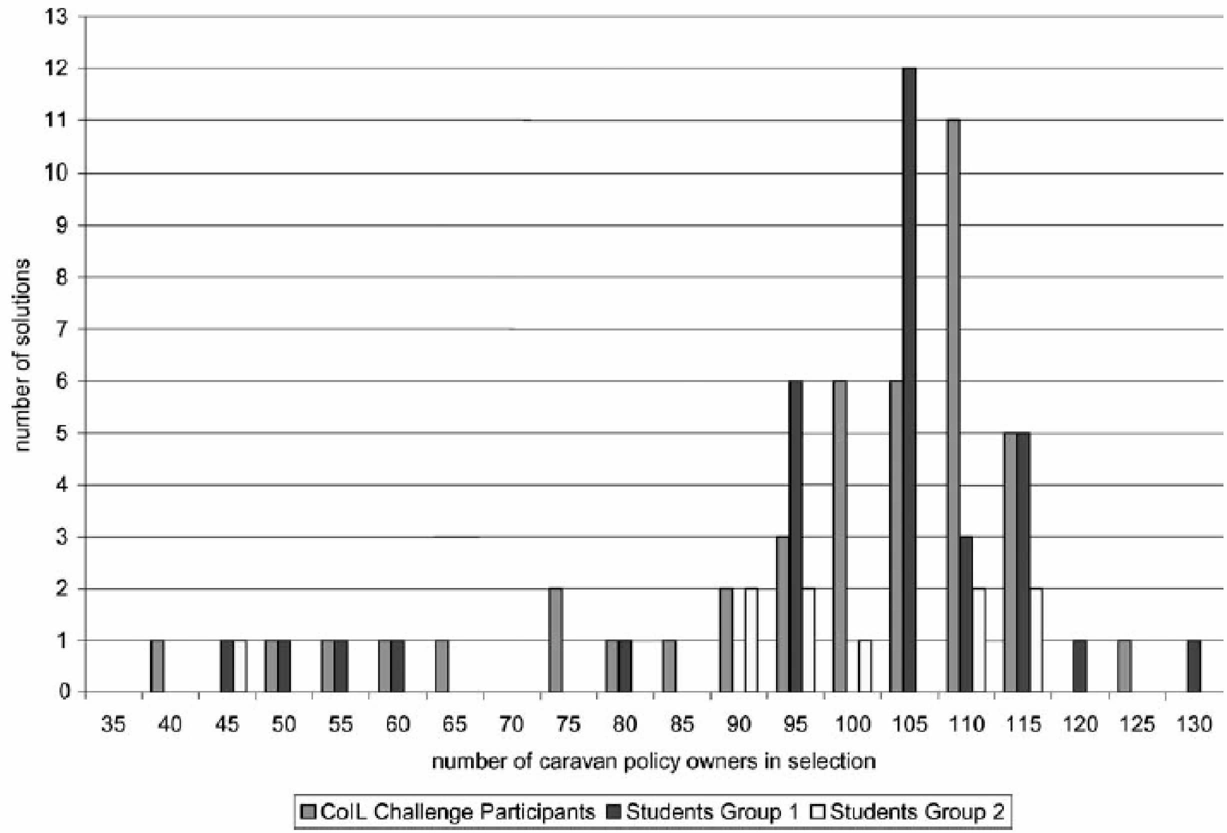
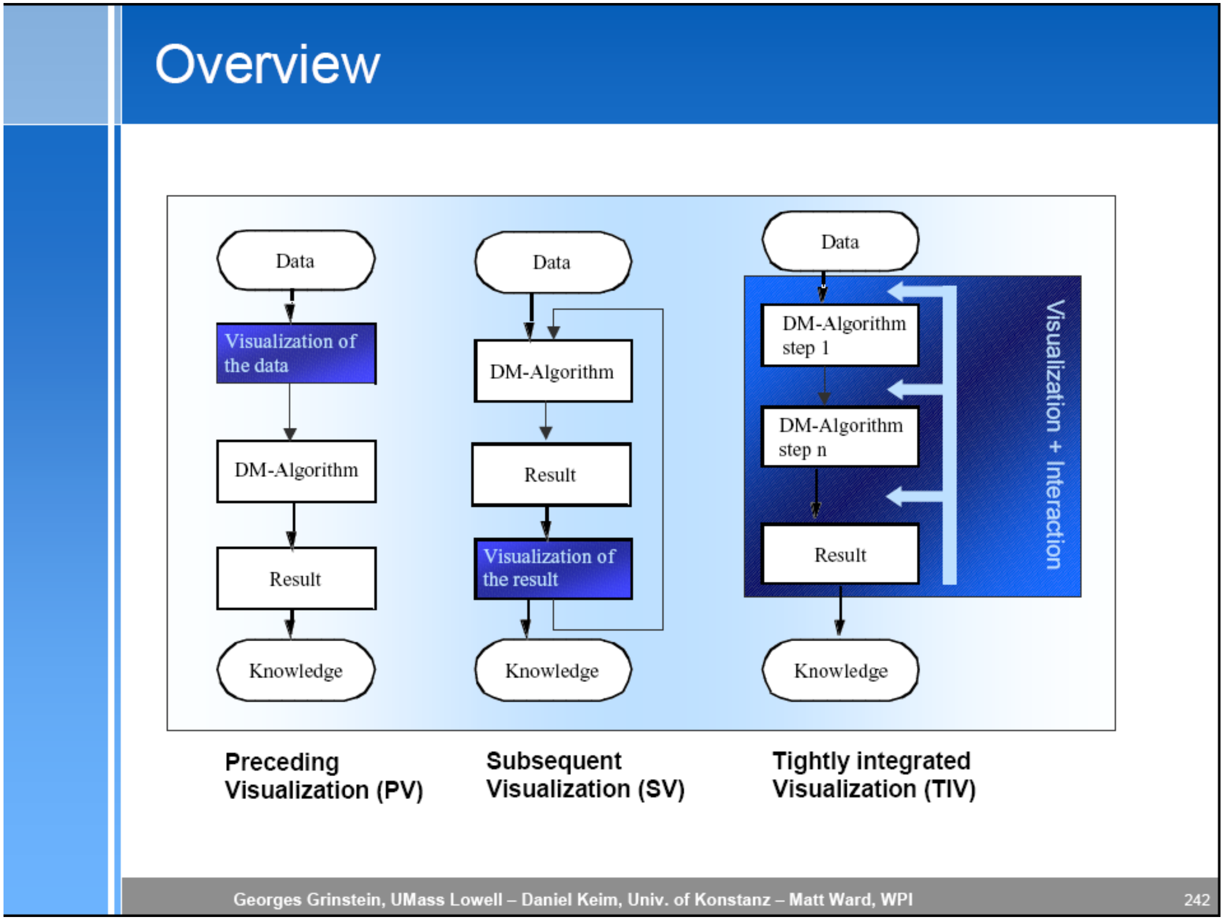
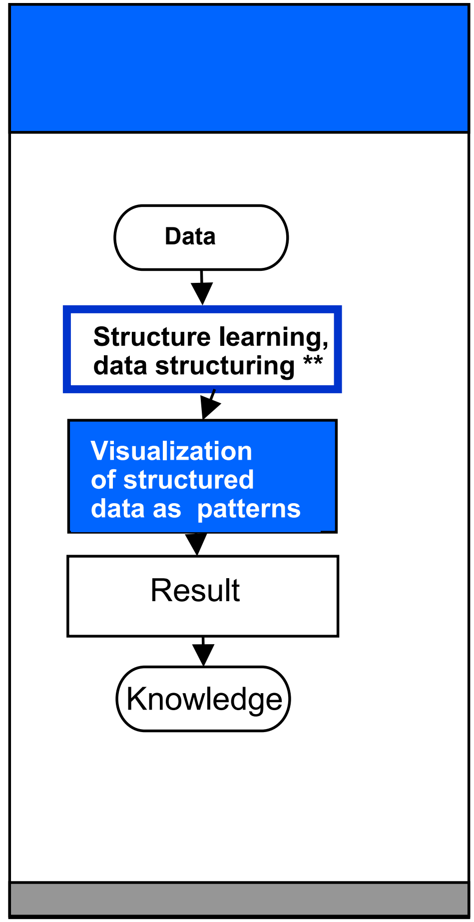


Figure 1. Histogram of prediction task performance for CoIL Challenge participants and two reference groups of students (bucket size is 5).

Visualization for Data Mining vs. Visual Data Mining as Visual Discovery



Visualization for Data Mining



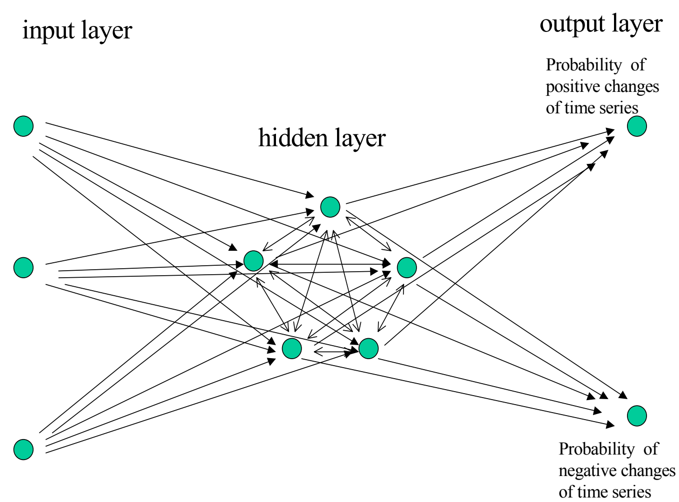
Visual Discovery

**Domain structure learning (discovering) algorithm and data structuring,

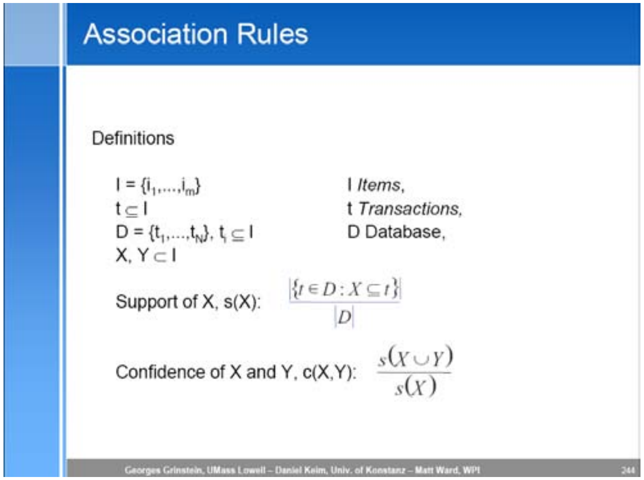
Outline

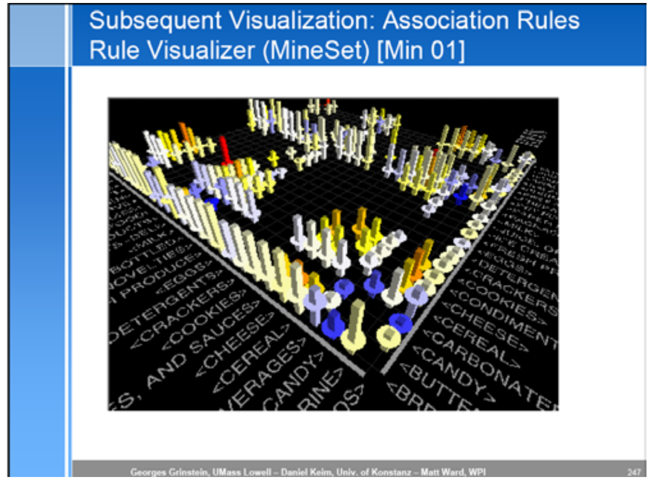
Visual Data Mining

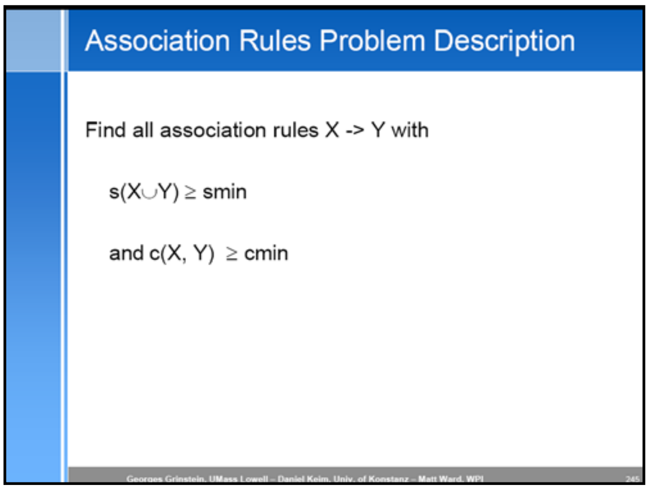
- Data Mining result visualization vs. Visual version of Data Mining methods vs.
  - Time Series analysis,
  - Neural Networks,
  - Decision Trees,
  - Discriminant Analysis, and
  - Bayesian methods
  - Association rules
- How to evaluate the results?
  - problem ID and method ID

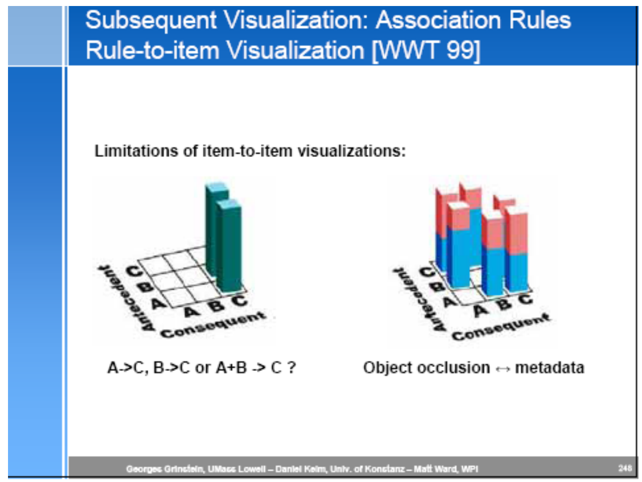


How to convert association rules to visual discovery tool?



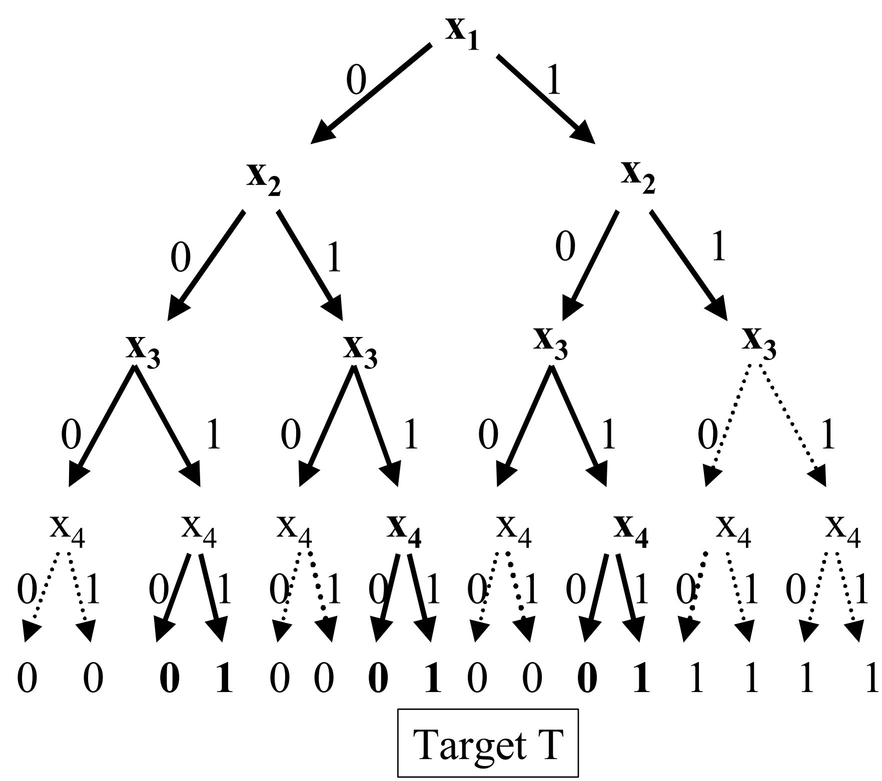






How to convert a decision tree method to a visual learning (visual discovery) tool ?

x_1	x_2	x_3	x_4	Target T
0	0	0	0	0
0	0	0	1	0
0	0	1	0	0
0	0	1	1	1
0	1	0	0	0
0	1	0	1	0
0	1	1	0	0
0	1	1	1	1
1	0	0	0	0
1	0	0	1	0
1	0	1	0	0
1	0	1	1	1
1	1	0	0	1
1	1	0	1	1
1	1	1	0	1
1	1	1	1	1



IF (x_1=1 & x_2=1) OR (x_3=1 & x_4=1) THEN T=1.

How to convert a first –order logic rule discovery methods to a visual learning (visual discovery) tool ?

First order logic rules

Rule: IF stock price today is greater than stock price yesterday and
trade volume today is greater than yesterday
THEN
tomorrow stock price will go up.

Rule:
IF Greater(StockPrice(t), StockPrice(t-1)) AND Greater(StockTradeVolume(t),StockTradeVolume(t-1))
THEN
StockPrice(t+1)>StockPrice(t)

Propositional rules

Category	Expression	Covered examples	Statistical significance
IF part for Class 1 (down stock direction)	1. v1∈[22.50,123.50) and v2≥-37.50 and v4<-80.00 and v9≥-91.50 and v10<-27.50 2. v1∈[22.50,123.50) and v2≥-37.50 and v4∈[-80.00, 4.50) and v7≥83.00 and v8≥-14.00 and v10<-27.50 3. v1∈[-107.50,22.50) and v2<-37.50 and v3≥120.50 and v7≥83.00 and v8∈[-4.50,14.00) 4. v2<-37.50 and v3≥20.50 and v4≥4.50 and v7≥83.00	9 4 3 7	0.995 0.901 0.824 0.983
IF part for Class 2 (up stock direction)	5. v1∈[9.50,22.50) and v2≥-37.50 and v4∈[-46.50,.4.50) and v5<-73.50 and v7<83.00	6	0.920

Visualization of Relational Data Mining and “Expert Mining” methods

- **Relational data mining is**
  - **the data mining technology process that uses relations between pieces of data and/or objects.**
- **Expert mining is**
  - **the technology of extracting knowledge from experts for discovering patterns**
- **Expert mining is complimentary to data mining**
  - **rules extracted using both technologies should be consistent to each other to be meaningful.**
- **Visual data mining as visual discovery approach builds a bridge between Expert mining and Data Mining.**

Outline

Part 1.Visual Data Mining Process

- 1.1. Visualization for Data Mining vs. Visual Data Mining
- 1.2. Challenges
- 1.3. Visual Relational Data Mining and Expert Mining

Part 2. Case studies

- 2.1. Fraud and Deception
- 2.2. Visual Correlation and Terrorism
- 2.3. Visual Reasoning

Part 3. Visual Data Mining for Binary Data

- 3.1. Methods
- 3.2. Case study: Breast cancer

Relational Data Mining

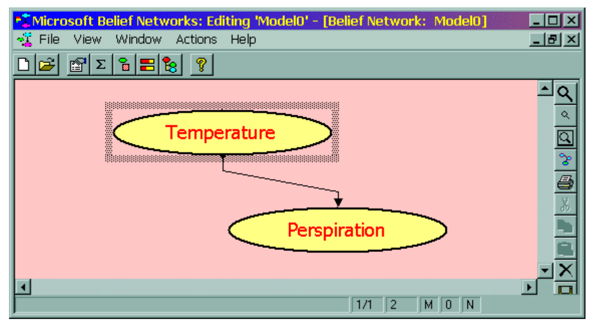
- **Ideas behind relational methods**
  - **Inductive and Stochastic Logic Programming**
  - **Propositional vs. First order logic Bayesian networks**
  - **Semantic Probabilistic Reasoning, MMDR**
  - **Propositional Decision Trees vs. First order logic Decision trees**
  - **FOIL**
  - **Propositional vs. First order logic association rules.**

Kovalerchuk,B., Vityaev, E., Data mining in finance: advances in relational and hybrid methods, Kluwer, 2000

S. Dˇzeroski and N. Lavraˇc. Relational Data Mining. Springer, 2001

How to convert *belief networks* (*Bayesian networks, BN*) to a visual learning (visual discovery) tool?

- **Main ideas:**
- **Create variables representing the distinct elements of the situation**
  - **define the set of outcomes or states that covers all possibilities for the variable and unimportant distinctions are shared between states.**
- **Establish the causal dependency relationships between the variables**
  - **create lines with arrowheads leading from the parent variable to the child variable.**
- **Assess the prior probabilities**
  - **supply the model with numeric probabilities for each variable**
- **Solve forecasting tasks using BN**



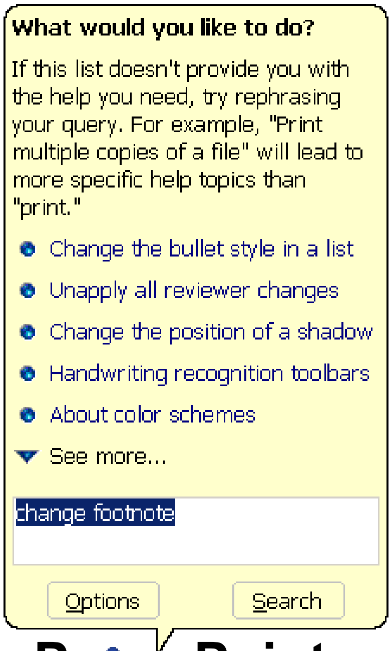
BN Editor and toolkit MSBNx http://research.microsoft.com/msbn/

Advantages of *belief networks* (*Bayesian networks, BN*)

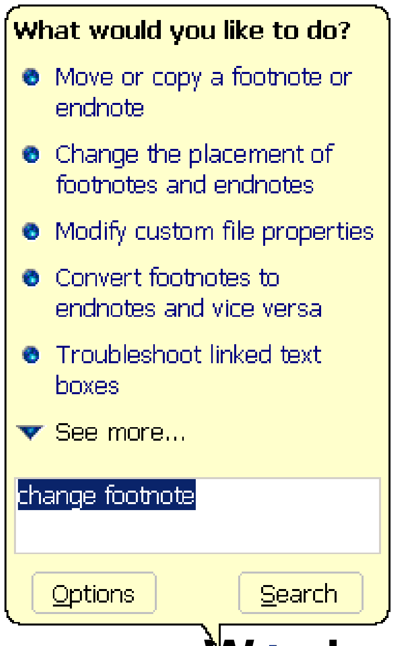
- **BNs allow to incorporate deep domain *structural knowledge* not only data**
  - **in the form of network structure.**
- **BNs allow to incorporate domain *uncertain knowledge***
  - **in the form of conditional probabilities associated with network nodes.**
- **BNs support relatively simple and efficient probabilistic inference process, e.g., for clustering.**
- **Software systems that support BN development are available**
- **Many systems have been implemented, e.g., for infectious diseases and traffic analysis**

Belief networks (Bayesian networks, BN)
Significant assumptions and limitations

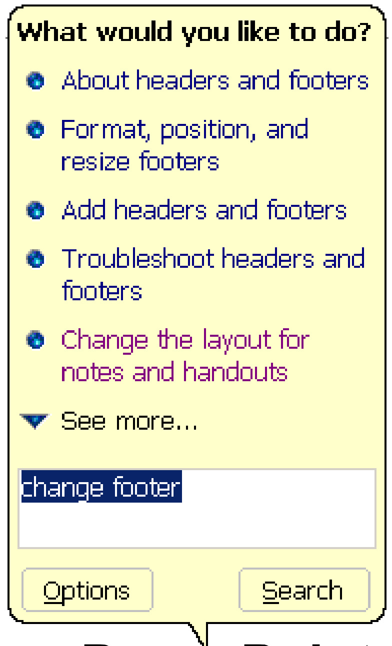
Microsoft office assistant   Footnote ? Footer



PowerPoint



Word

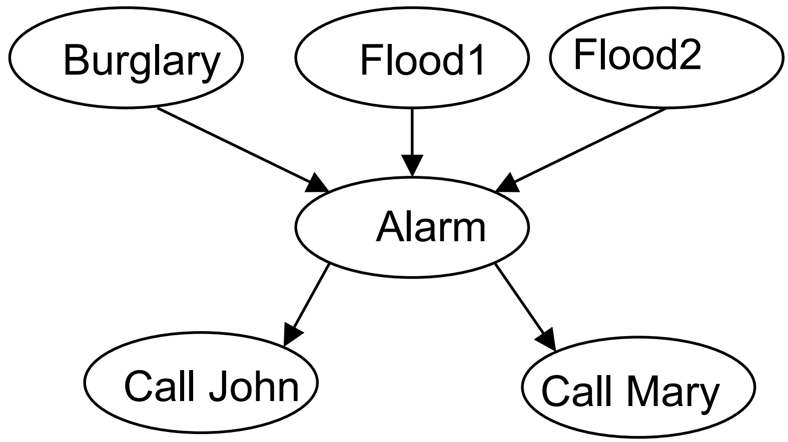


PowerPoint

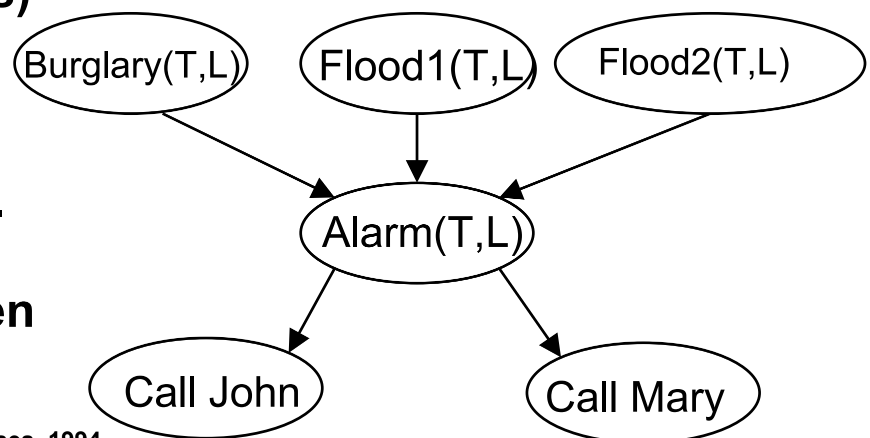
The domain and its causal structure should be stable, known, can be learned and formalized enough
Should the assistant be designed to handle such questions?
Visual discovery can help to improve BNs

How to convert First-order belief networks (FOBN) to a visual discovery tool ?

- **The structure of a traditional BN is**
  - **a set of (propositional) clauses, one clause for each node**
    - **E.g., Flood**



- **The structure of FOBN is**
  - **A set of FOL clauses, one clauses for each node.**
- **Types of clauses:**
  - **an extensional (factual clauses)**
    - **specify individual objects and situations**
    - **Flood(time, location)**
  - **intensional (non-fact clauses)- the overall regularities**
    - **If Flood(time, location) then Wet(time, location, soil)**



Haddawy. Generating Bayesian networks from probabilistic logic knowledge bases, 1994.
Koller. Probabilistic relational models.1999.
Ngo, Haddawy. Answering queries from context–sensitive probabilistic knowledge bases.1997.

Outline

Part 1.Visual Data Mining Process

- **1.1. Visualization for Data Mining vs. Visual Data Mining**
- **1.2. Challenges**
- **1.3. Visual Relational Data Mining and Expert Mining**

Part 2. Case studies

- **2.1. Fraud and Deception**
- **2.2. Visual Correlation and Terrorism**
- **2.3. Visual Reasoning**

Part 3. Visual Data Mining for Binary Data

- **3.1. Methods**
- **3.2. Case study: Breast cancer**

Detecting deception

Counter-deception business process hypothesis generation.

= automated course of action generation for tactical situations

= Manual hypothesis elicitation from analysts for strategic situations

- a Bayesian belief network generation
- Sensitivity analysis based on Bayesian classification

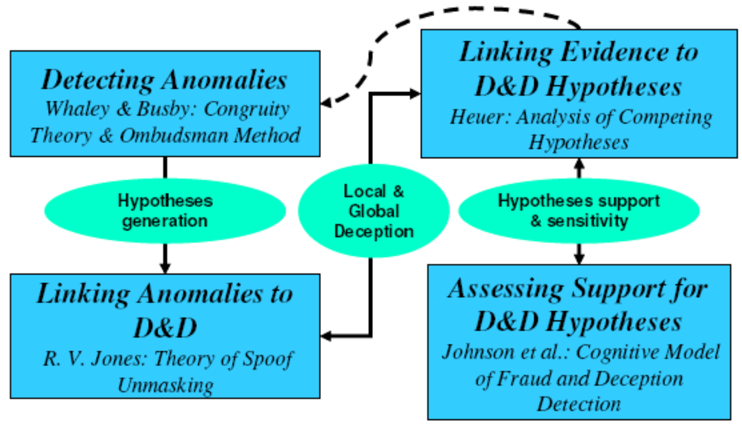
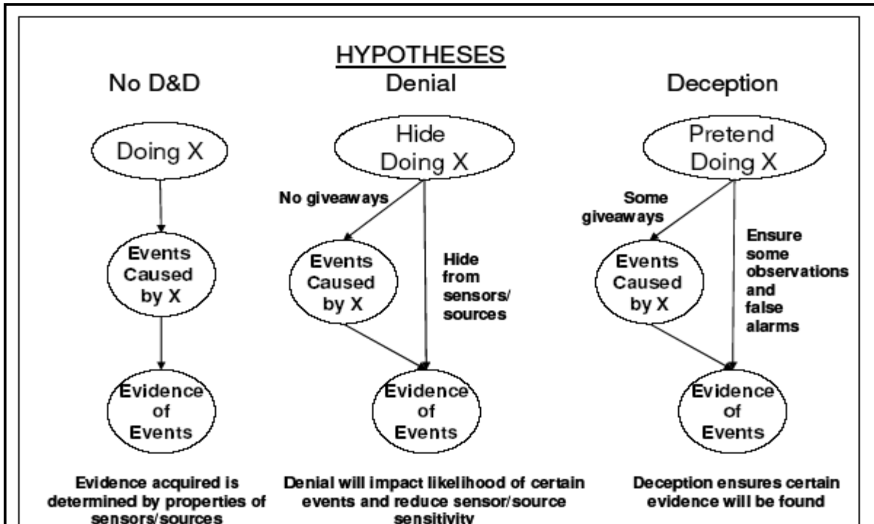


Figure 1: Counter-deception business process

[Stech, Elsässer, 2004]

Relationships among events and evidence with no denial or deception, global denial, and global deception



Bayesian network to model denial or deception

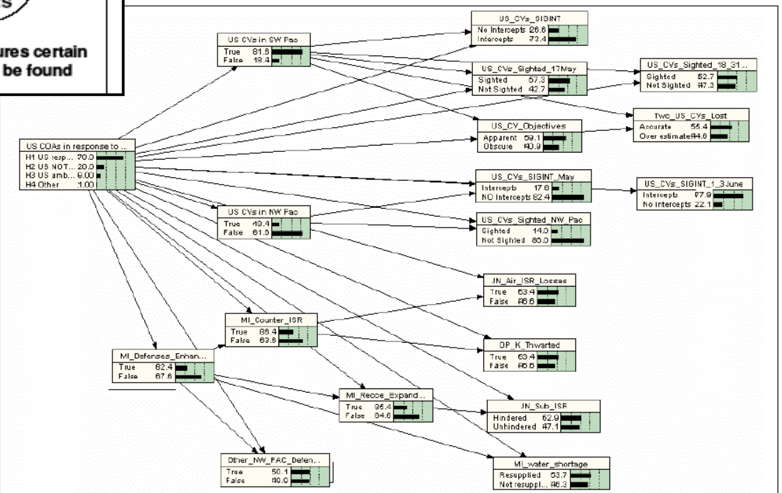


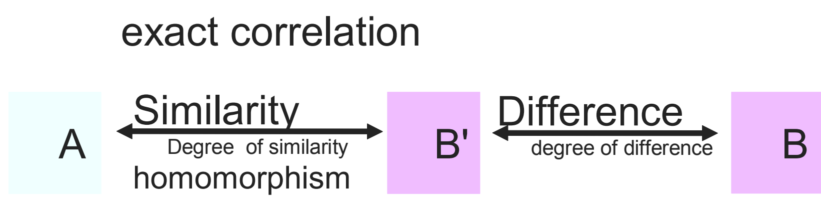
Figure 3: Bayesian belief network reflecting events (center) related to Japanese hypotheses (left, enemy courses of action, COAs) in response to "Operation MI" and available intelligence evidence (right) of possible American denial and deception tactics linked to those COAs and events.

Stech, Elsässer, 2004

Outline

Visual Correlation and Terrorism

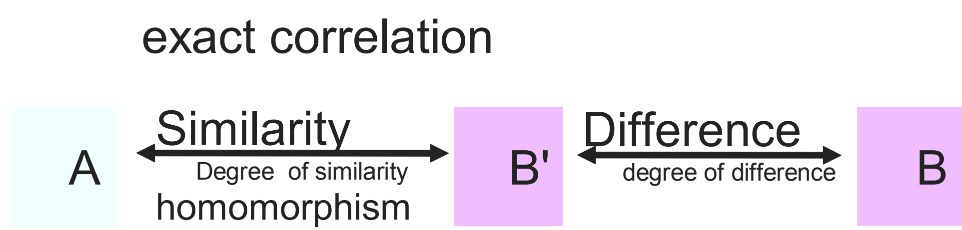
- **Visual analysis, data mining, link discovery and correlation are gaining momentum in IA**
  - **ARDA/NGA GI2VIs program.**
- **How visual relational methods contribute to this work?**
- **Data -- MUC3/4 competition data on terrorism**
- **A visual correlation system**
  - **Demo**
- **Visual Reasoning**



Visual correlation using intermediate object *B'*

Visual reasoning, data mining, link discovery and correlation are gaining momentum

- The goal is
  - to help to reveal relations in non-numeric data
  - to reveal a structure for building forecasting model
- Visual co-relation (correlation) has a goal of
  - Relating pieces of data visually
  - Revealing relations between objects.



Visual correlation using intermediate object *B'*

Visual Correlation and Terrorism

MUC raw text corpus description

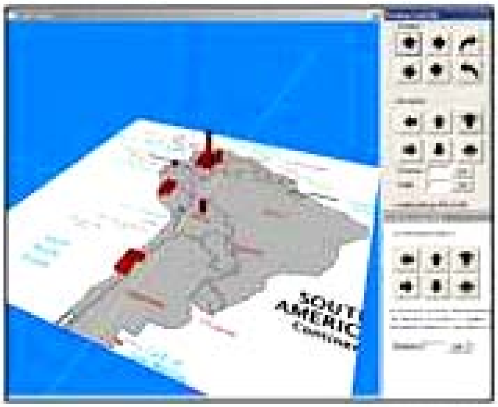
Characteristic	*Description*
1. Data sources	The Foreign Broadcast Information Service.
2. Text types	Newspaper and newswire stories, radio and TV broadcasts, interviews, and rebel communiqués summary reports, transcripts from speeches and interviews
3. Location	Latin America
4. Original language	Spanish
5. Text grammar	Well-formed sentences, all are in upper case
6. Number of texts s	1300 texts
7. Individual text size	Average size is 12 sentences (~0.5 a page), smallest text -one paragraph, largest text -two pages
8. Number of sentences	15,600 sentences
9.Number of unique lexical items	18,240 lexical units
10.Number of words	400,000 words
11.Number of events in a text	1-5 events per single text source
12.Average sentence length	27 words
13.Timeframe	1980s
14.Terrorism relevant texts, %	50%

Visual Correlation and Terrorism

A sample of raw text [MUC Data Sets, NIST, 2001]

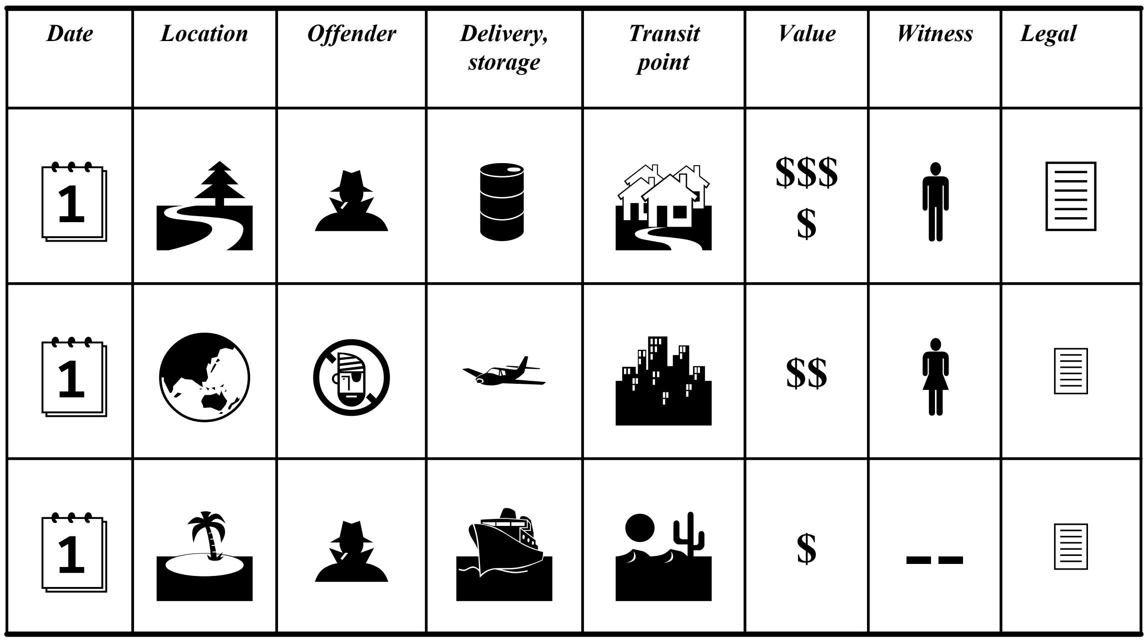
DEV-MUC3-0008 (NOSC)
BOGOTA, 9 JAN 90 (EFE) -- [TEXT] RICARDO ALFONSO CASTELLAR, MAYOR OF ACHI, IN THE NORTHERN DEPARTMENT OF BOLIVAR, WHO WAS KIDNAPPED ON 5 JANUARY, APPARENTLY BY ARMY OF NATIONAL LIBERATION (ELN) GUERRILLAS, WAS FOUND DEAD TODAY, ACCORDING TO AUTHORITIES.CASTELLAR WAS KIDNAPPED ON 5 JANUARY ON THE OUTSKIRTS OF ACHI, ABOUT 850 KM NORTH OF BOGOTA, BY A GROUP OF ARMED MEN, WHO FORCED HIM TO ACCOMPANY THEM TO AN UNDISCLOSED LOCATION. POLICE SOURCES IN CARTAGENA REPORTED THAT CASTELLAR'S BODY SHOWED SIGNS OF TORTURE AND SEVERAL BULLET WOUNDS. CASTELLAR WAS KIDNAPPED BY ELN GUERRILLAS WHILE HE WAS TRAVELING IN A BOAT DOWN THE CAUCA RIVER TO THE TENCHE AREA, A REGION WITHIN HIS JURISDICTION. IN CARTAGENA IT WAS REPORTED THAT CASTELLAR FACED A ''REVOLUTIONARY TRIAL'' BY THE ELN AND THAT HE WAS FOUND GUILTY AND EXECUTED. CASTELLAR IS THE SECOND MAYOR THAT HAS BEEN MURDERED IN COLOMBIA IN THE LAST 3 DAYS. ON 5 JANUARY, CARLOS JULIO TORRADO, MAYOR OF ABREGO IN THE NORTHEASTERN DEPARTMENT OF SANTANDER, WAS KILLED APPARENTLY BY ANOTHER GUERILLA COLUMN, ALSO BELONGING TO THE ELN. TORRADO'S SON, WILLIAM; GUSTAVO JACOME QUINTERO, THE DEPARTMENTAL GOVERNMENT SECRETARY; AND BODYGUARD JAIRO ORTEGA, WERE ALSO KILLED. THE GROUP WAS TRAVELING IN A 4-WHEEL DRIVE VEHICLE BETWEEN CUCUTA AND THE RURAL AREA KNOWN AS CAMPANARIO WHEN THEIR VEHICLE WAS BLOWN UP BY FOUR EXPLOSIVE CHARGES THAT DETONATED ON THE HIGHWAY.

Visual Discovery of Correlation and Terrorism



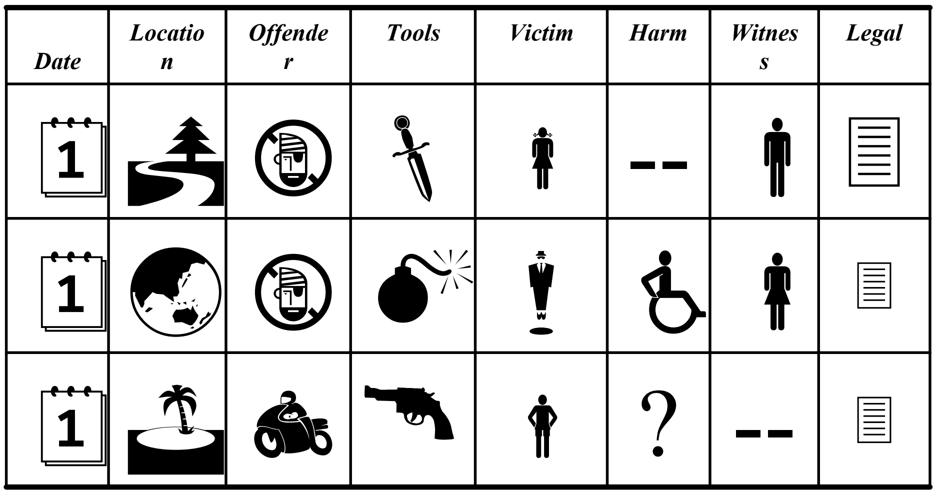
Bruegel System case studies.

Visual Correlation



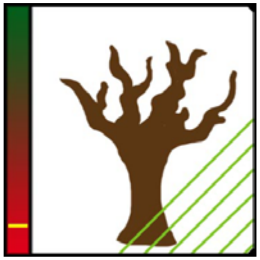
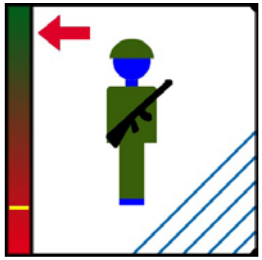
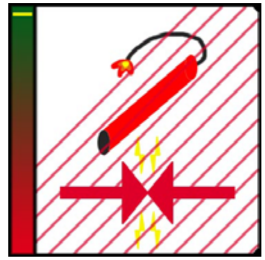
Iconic sentences that summarize drag trafficking records and
Allow quickly co-relate events

Visual Correlation

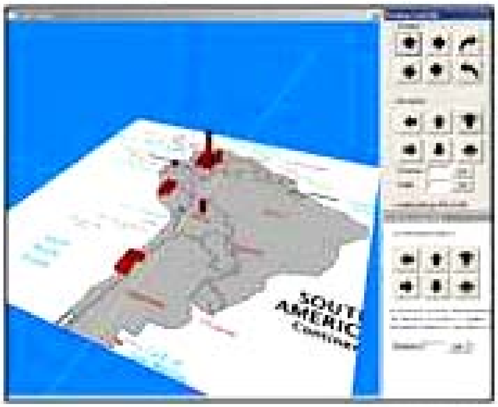


Iconic sentences that summarize criminal records

Visual Correlation and Terrorism

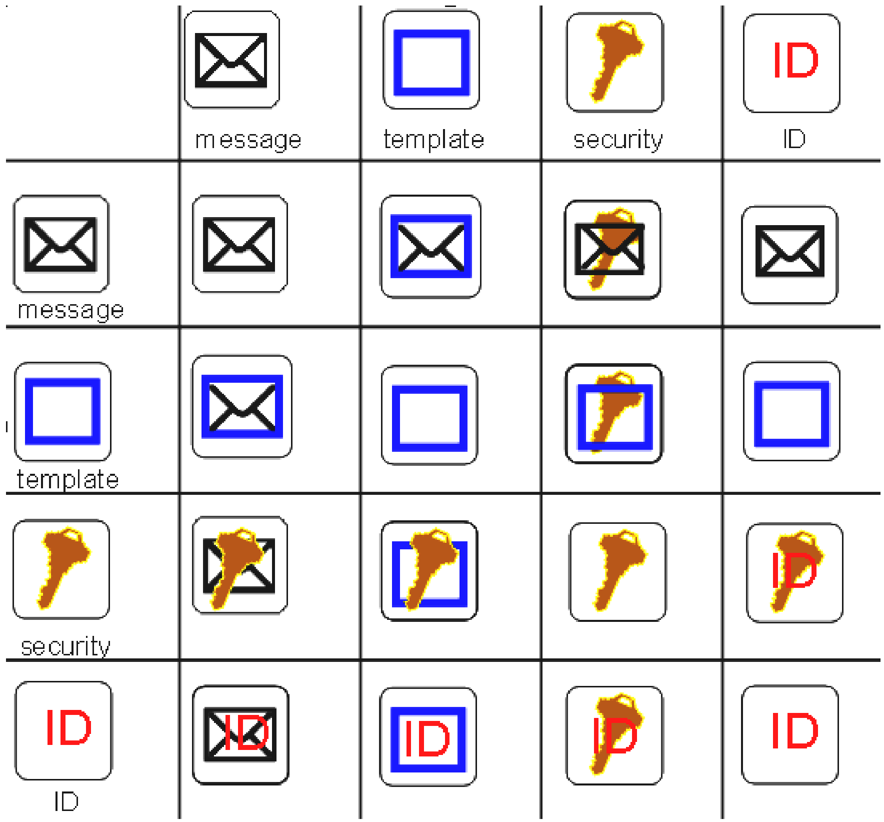
		
organization (a tree icon) of a medium size (encoded by green lines) and relatively high confidence (encoded by a yellow mark on red)	Several soldiers perpetrators encoded by the soldier icon, red modifier for perpetrators; 5 blue lines for several and a yellow mark for medium in the red confidence scale.	terrorist act with dynamite and significant damage (red lines)

Visual Correlation and Terrorism



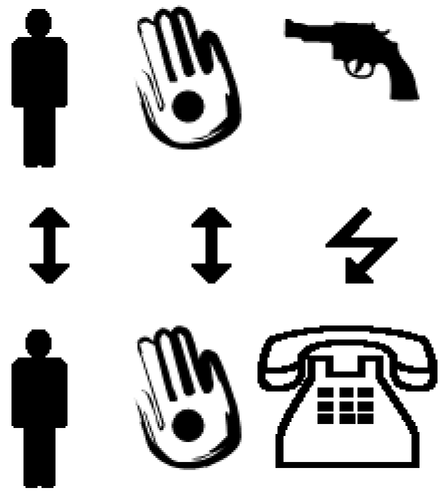
Bruegel System case studies.

Visual Correlation and Terrorism



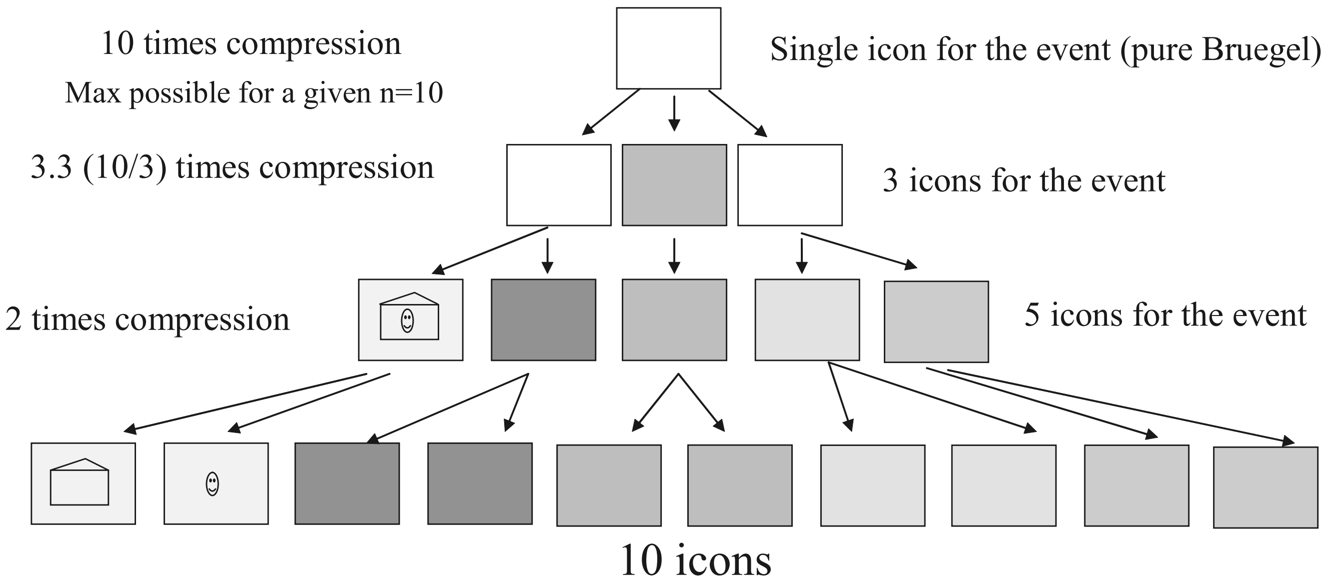
Composite icon generation.

Visual Correlation and Terrorism

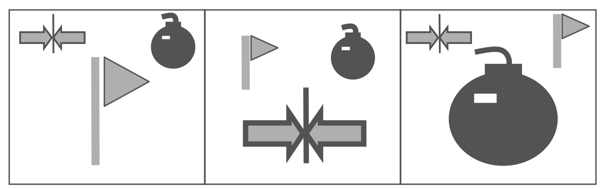
Visual queries	*Comment*
	Standard query: Find an adult male using his hand to operate a gun OR Exception: "Find an adult male using his *hand* to operate a telephone". 

Visual Correlation and Terrorism

Dynamics of compression of iconic sentence.



Icons with user defined weighting.
[More detail]



Outline

Part 1.Visual Data Mining Process

- **1.1. Visualization for Data Mining vs. Visual Data Mining**
- **1.2. Challenges**
- **1.3. Visual Relational Data Mining and Expert Mining**

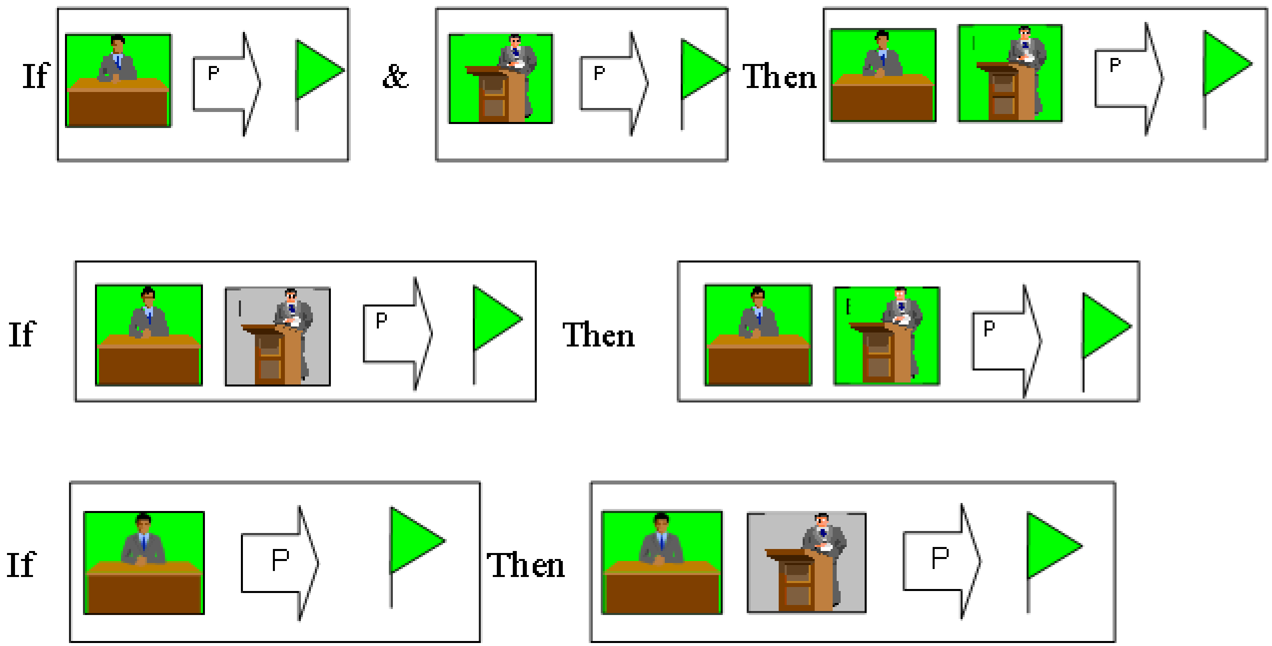
Part 2. Case studies

- **2.1. Fraud and Deception**
- **2.2. Visual Correlation and Terrorism**
- **2.3. Visual Reasoning**

Part 3. Visual Data Mining for Binary Data

- **3.1. Methods**
- **3.2. Case study: Breast cancer**

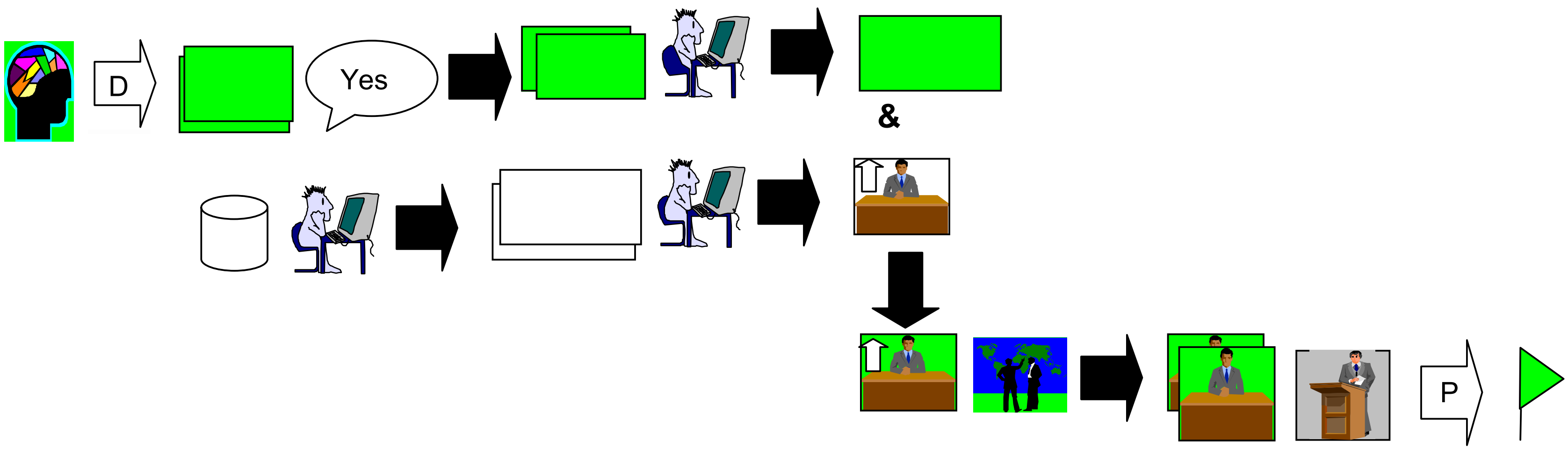
Visual Reasoning



Visual reasoning rules

Idea and motivation of visual reasoning
Geometry visual art

Visual Correlation and Reasoning



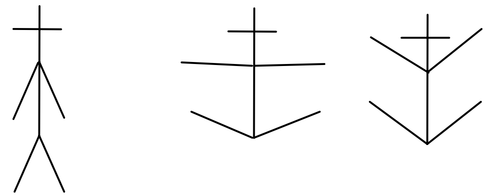
Integrated visual evidentiary reasoning scheme

Outline

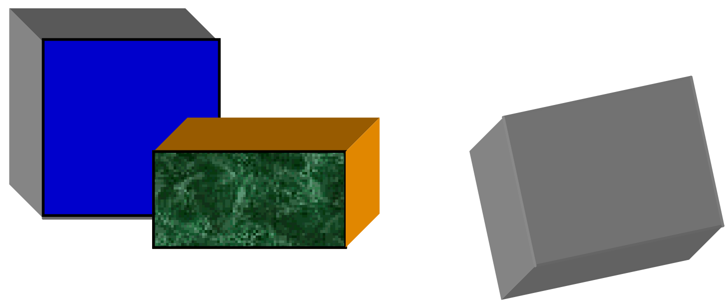
Challenges

- **Available VDM methods do not address the <u>specifics of binary data</u>**
  - **little variability in the visual representation**
- **We need to preserve <u>data richness</u> without the excessive aggregation, e.g., presentation graphics such as bar charts** [Keim, Hao, Dayal, & Hsu, 2002].
- **Often data <u>lack natural 3-D space and time dimensions</u> and require the visualization of an abstract feature.**

Glyphs



- A *glyph* is a 2-D or 3-D object (icon, cube, or more complex *"Lego-type"* object).
- Glyph Visualization or iconic visualization is an attempt to encode multidimensional data within the parameters of the icons, such as the
  - Shape,
  - Color,
  - Transparency,
  - Orientation
  - Length, height, width
  - The number of glyphs that can be visualized is relatively limited because of possible glyph overlap and occlusion



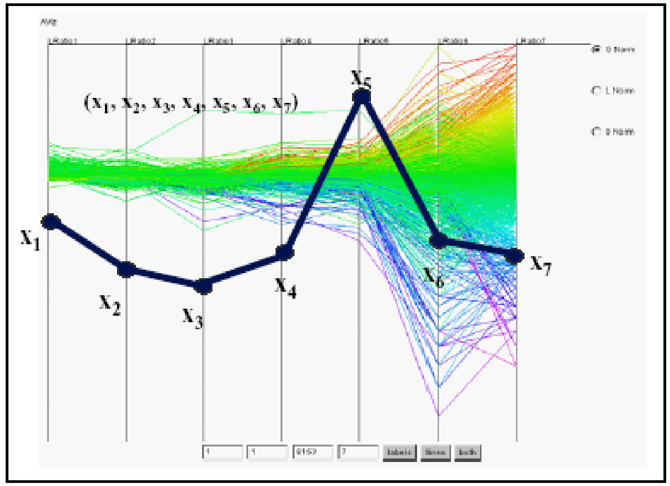
Ebert, Shaw, Zwa, Miller & Roberts, 1996; Post, van Walsum, Post & Silver, 1995; Ribarsky, Ayers, Eble & Mukherja, 1994.

Spiral Bar and others techniques

- **Alternative techniques such as Generalized Spiral and Pixel Bar Chart are developed in [Keim, Hao, Dayal & Hsu, 2002].**
- **These techniques work with large data sets without overlapping, but only with a few attributes (these range from a *single attribute* to *perhaps four to six attributes*).**
- **Another set of visualization methods, known as Scatter, Splat, Map, Tree, and Evidence Visualizer, are implemented in MineSet (Silicon Graphics), which permits up to *eight dimensions* to be shown on the same plot by using color, size, and animation of different objects [Last & Kandel, 1999].**

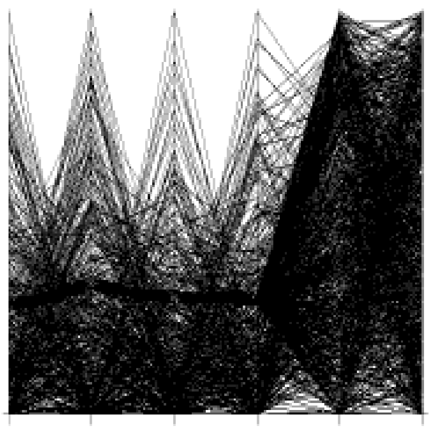
Parallel coordinate techniques

- In parallel coordinates,
  - each vertical axis corresponds to a data attribute (x_i) and
  - a line connecting points on each parallel coordinate corresponds to a record.

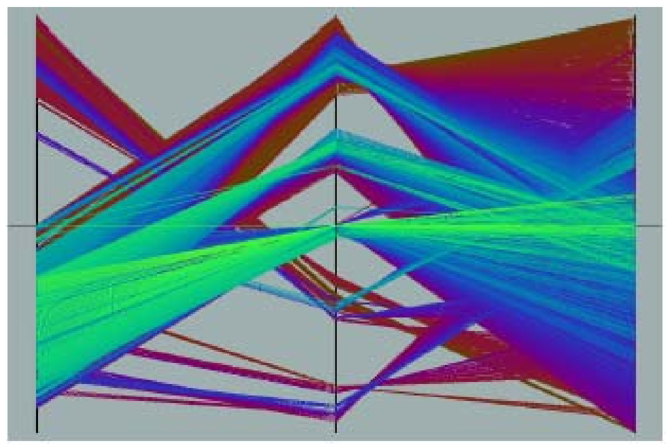


Inselberg & Dimsdale, 1990

15,000 data items



5 % of the data

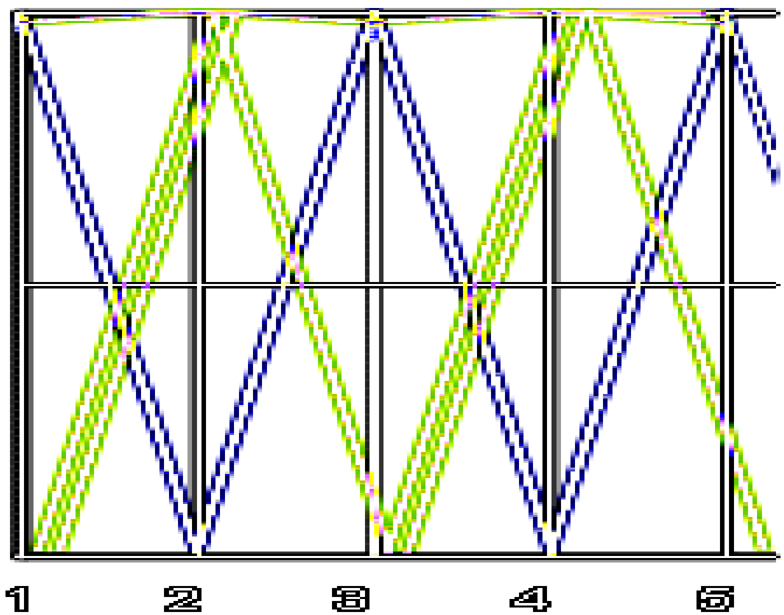


Georges Grinstein, Daniel Keim, Matthew Ward, Information Visualization and Visual Data Mining, IEEE Visualization 2004 Conference, Seattle, fusion.cs.uni-magdeburg.de/pubs/TVCG02.pdf

Parallel coordinate techniques

- This visualization can
  - work with ten or more attributes, but
  - suffers from record overlap and thus
  - is limited to tasks with well-distinguished cluster records.

01010;11010;01110;01011;01111;
11011;11111;**10101**;11101;10111



Can we discover a regularity that governs the dataset using parallel coordinates?

01010;11010;01110;01011;01111;

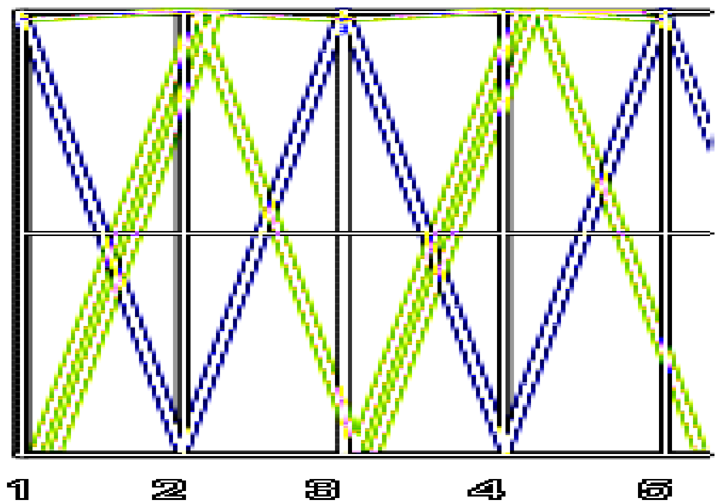
11011;11111;10101;11101;10111

0 **It is difficult, but the regularity is a simple monotone Boolean function (x2 & x4) ∨ (x1 & x3 & x5).**

0 **This function is true for these vectors.**

0 **Technically, the number of dimensions is limited only by the screen resolution.**

0 **In reality this is typically too overwhelming for data understanding.**



Ten Boolean records overlapped in parallel coordinates

Glyphs

- **Typically, glyphs can visualize up to *nine attributes***
  - **three positions *x*, *y*, and *z*;**
  - **three size dimensions;**
  - **color; opacity; and shape.**
- **Texture can add more dimensions.**
- **About 22 separate shapes can be distinguished on the average** [Shaw, Hall, Blahut, Ebert & Roberts, 1999].
- **Some glyph methods use data dimensions as positional attributes to place glyphs; other methods place glyphs using *implicit or explicit structure* within the data set.**
- **An overview of multivariate glyphs is presented in** [Ward, 2002]. **This overview includes a taxonomy of glyph placement strategies and guidelines for developing such a visualization.**

Boolean Visual Data Mining

- Many data mining problems can be encoded using Boolean vectors, where each record is a set of binary values {0; 1} and each record belongs to one of two classes (categories) that are also encoded as 0 and 1.
  - A patient can be represented as a Boolean vector of symptoms along with an indication of the diagnostic class (e.g., benign or malignant tumor)
- For *n*-dimensional Boolean attributes, traditional glyph-based visualizations are useful but somewhat limited.
- Attributes of a Boolean vector can be encoded in glyph lengths, widths, heights, and other parameters.
  - There are only two values for the length, width, and other parameters for each Boolean vector.
- When plotted as nodes in a 3-D binary cube, many objects will not be visually separated.

Glyph Placement on a Data Structure

0 **Glyph placement based on the use of data structure is a promising approach.**

0 **We call this the GPDS approach (Glyph Placement on a Data Structure).**

0 **It increases the number of attributes that can be visualized and decreases an occlusion.**

0 **In this approach, some attributes are**

- **implicitly encoded in the data structure while others are**
- **explicitly encoded in the glyph.**

Glyph Placement on a Data Structure

- **The approach and methods described below do not follow the traditional glyph approaches that would**
  - **put *n*-dimensional Boolean vectors ($n > 3$) into 3-D space, making them barely distinguishable.**
- **The methods rely on monotone structural relations between Boolean vectors in the *n*-dimensional binary cube, E^n.**
  - **Data are visualized in 2-D as chains of Boolean vectors.**
- **Currently, the system supports two visual forms:**
  - **the Multiple Disk Form (MDF) and**
  - **the *"*Yin Yang*"* Form (YYF).**

Method: the numeric order of data layout

- **Consider a set of *n*-dimensional Boolean vectors *V* such as**

  01010;11010;01110;01011;01111;

  11011;11111;10101;11101;10111

- **Every n-dimensional Boolean data set *V* can be encoded as a Boolean function in a disjunctive normal form (DNF) or conjunctive normal form (CNF).**
  - **Thus, *visualization of a Boolean data set is equivalent to visualization of a Boolean function*.**

- **Every Boolean function can be decomposed into a set of monotone Boolean functions [Kovalerchuk et al., 1996].**

The monotone structure is important for the data mining tasks

- **Most of data mining methods are based on the *hypothesis of local compactness*:**
  - **if two objects have similar features, then they belong to the same class.**
- **Assume that the data satisfy the property of monotonicity, that is,**
  - **if vector *a* belongs to class 1, then a vector *b* that is greater than or equal to *a* also belongs to class 1.**
- **Example: a= 01010, b= 11010**

  01010 < 11010, Each component a_i of 01010 is no greater than b_i of 01010.
- **Informally, monotonicity means that if a patient with symptoms *a* has a malignant tumor then another person with symptoms *b* that include all *a* symptoms and some additional symptoms most likely also has a malignant tumor.**

Representation Structure: level hierarchy of Boolean vectors

Boolean vectors are ordered vertically by their Boolean norm (sum of "1"s) with the largest vectors are rendered on the top, starting from (11111).
Each vector is first placed in the view and then drawn as colored bar: white for the 0 class, black for the 1 class.

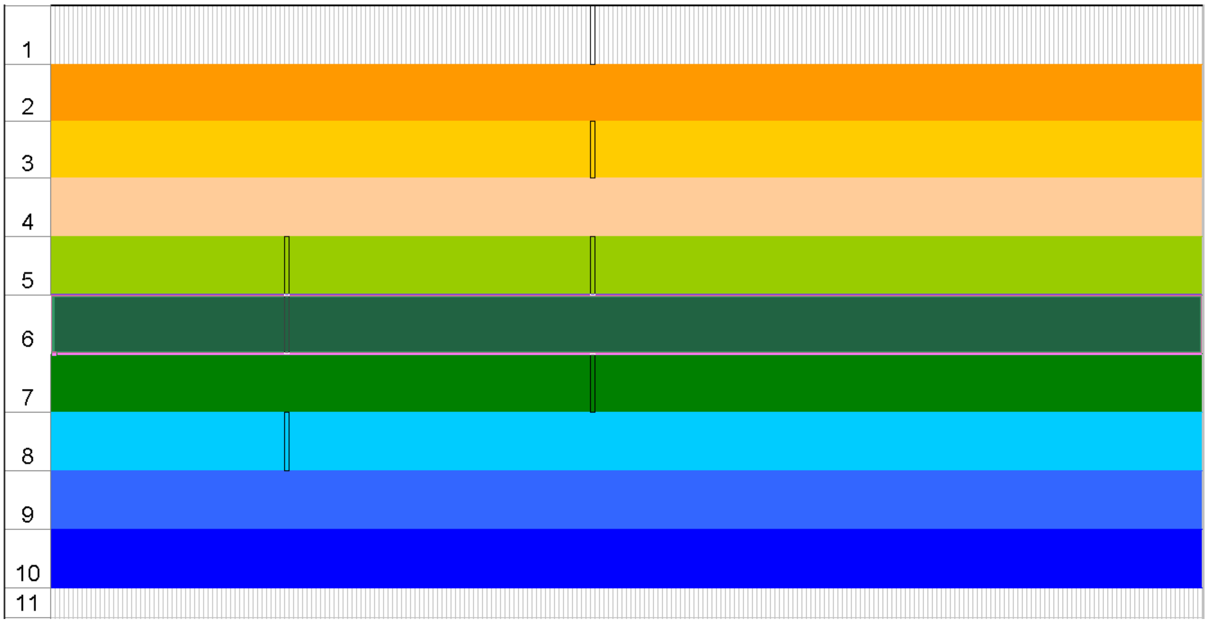
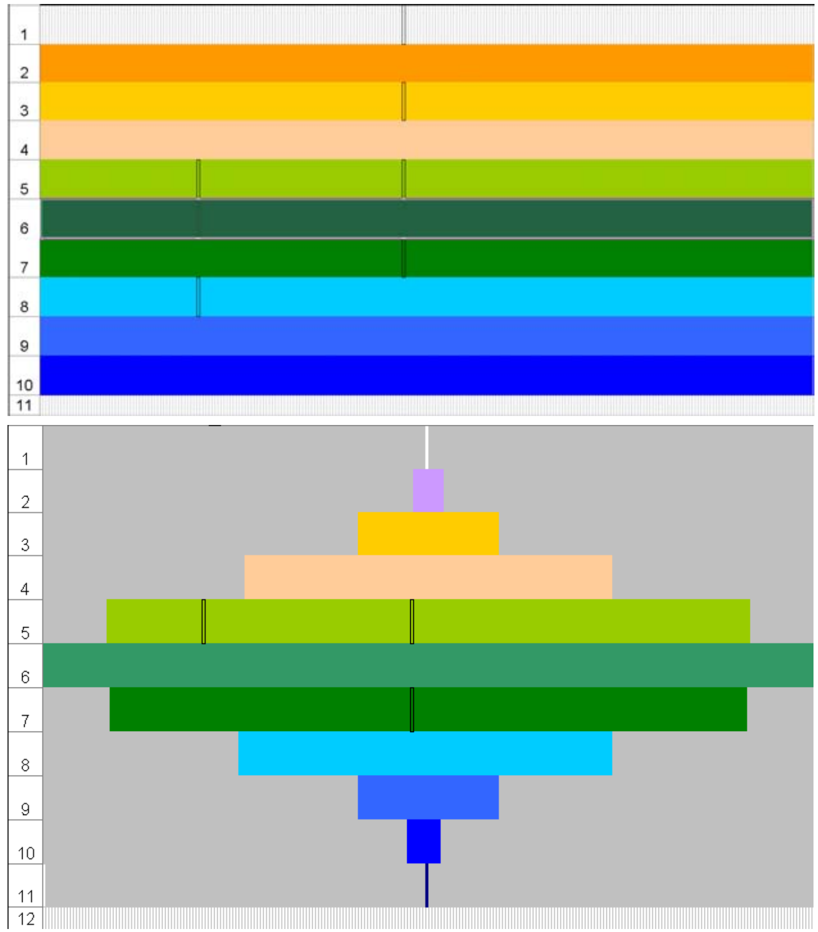


Table Form Visualization (TFV)- Vectors occupy fixed positions based on their numeric value (binary/decimal) where

N=10, level 0 – (00000 00000) and level 10 – (11111 11111),

Multiple disk form (MDF) for visualization

- In MDF all repeating vectors are deleted.
- Each level is called a disk and the entire visualization is called the multiple disk form (MDF).
- In the MDF form, every vector has a fixed horizontal and vertical position.



Vector placement procedure P_1

- **Each binary vector is converted to its decimal equivalent. For instance, the decimal equivalent of the vector 0000000010 would be 2. Each vector is then placed horizontally in its disk based its value with value 0 being on the right side. We call this procedure *P1*.**

- **The advantage of this procedure is to allow the user to compare more than one binary dataset or Boolean function at a time. This is possible because each vector always is assigned the same fixed location of the disks based on its value. If two data sets or Boolean functions are equal then they have exactly the same layout.**

Comparison of patterns

- **Procedure P1**
  - **the same border scheme for different Boolean functions and datasets,**
  - **direct comparison of different functions.**

- **Procedure P_1 can only be used for comparing two data sets or functions**

- **Procedure P_1 does not visualize any border or structure of the Boolean function;**

Chain-based border visualization procedure P_2

- Procedure *P2* is based on the decomposition of the binary cube, E^n (set of all n-dimensional binary vectors) into chains.

Table 1. Boolean data chains

Chain 1	Chain 2	Chain 3	Chain 4	Chain 5
01010	01010			
11010	01011	01110	10101	
11011	01111	01111	11101	10111
11111	11111	11111	11111	11111

Chains of Boolean vectors

While vectors on the same chain are ordered, vectors on the different chains may not be ordered. There is only a partial order on Boolean vectors.

The **partial order** is defined as follows: vector $\boldsymbol{a} = (a_1, a_2, \ldots, a_n)$ is greater or equal to vector $\boldsymbol{b} = (b_1, b_2, \ldots, b_n)$ if for every $i = 1, 2, \ldots, n$; $a_i \geq b_i$. This partial order means that chains may overlap as shown in Table 1. We will use the notation $\boldsymbol{a} \geq \boldsymbol{b}$ if Boolean vector $\boldsymbol{a}$ is greater than or equal to Boolean vector $\boldsymbol{b}$. A set of vectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n$ is called a **chain** if

$$\boldsymbol{v}_1 \geq \boldsymbol{v}_2 \geq \ldots \geq \boldsymbol{v}_n .$$

Chain 1	Chain 2	Chain 3	Chain 4	Chain 5
01010	01010			
11010	01011	01110	10101	
11011	01111	01111	11101	10111
11111	11111	11111	11111	11111

Hansel chains for data structuring

- **We focus on special chains of vectors called Hansel chains.**
- **The Hansel chains are computed and then aligned vertically on the multiple disk structure (MDF).**
- **Hansel chains have different lengths with possible values from 1 to *n* elements.**
  - **To keep the integrity of the MDF structure, we have to place these chains so that no elements fall out of the disks.**
- **The longest chain is placed on the center of the disk and the others chains are placed alternatively to the right and left of the first chain.**

A Method For Visualizing Pattern Borders

0 **The goal of this VDM method is to show patterns in a simple visual form for SME to explore.**

0 **If cases of two classes (say, benign and malignant) are separated in the visual space and the border between classes is simple, then the goal of VDM is reached.**

0 **We can reveal such border visually using the technique of monotone Boolean functions.**

Pattern border with procedures P_1 and P_2

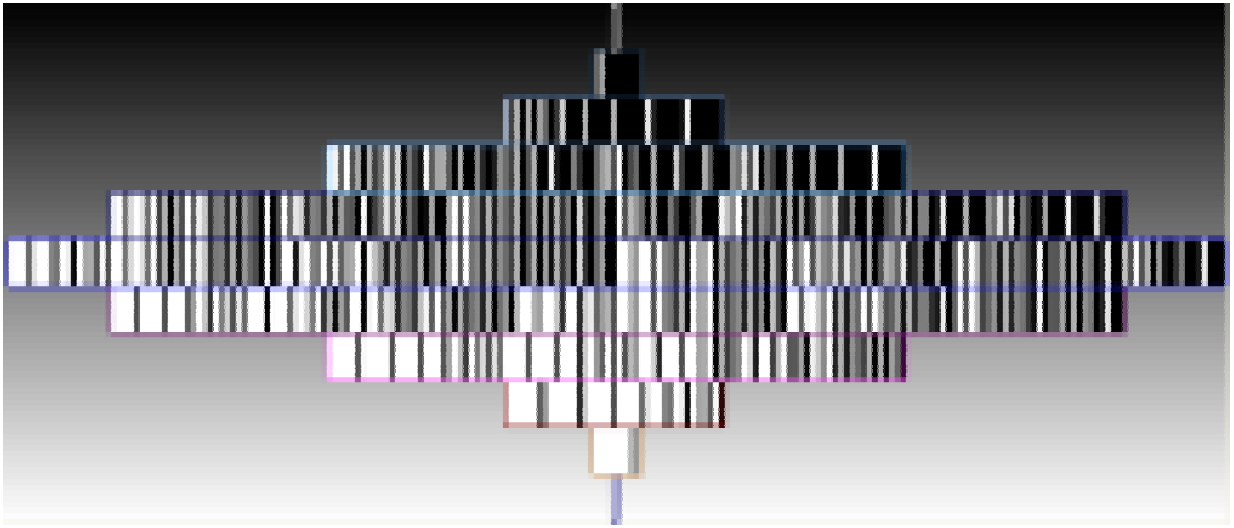


Figure 4. A MDF using procedure P_1 with $f(x_1, x_2, \ldots, x_{10}) = x_1$

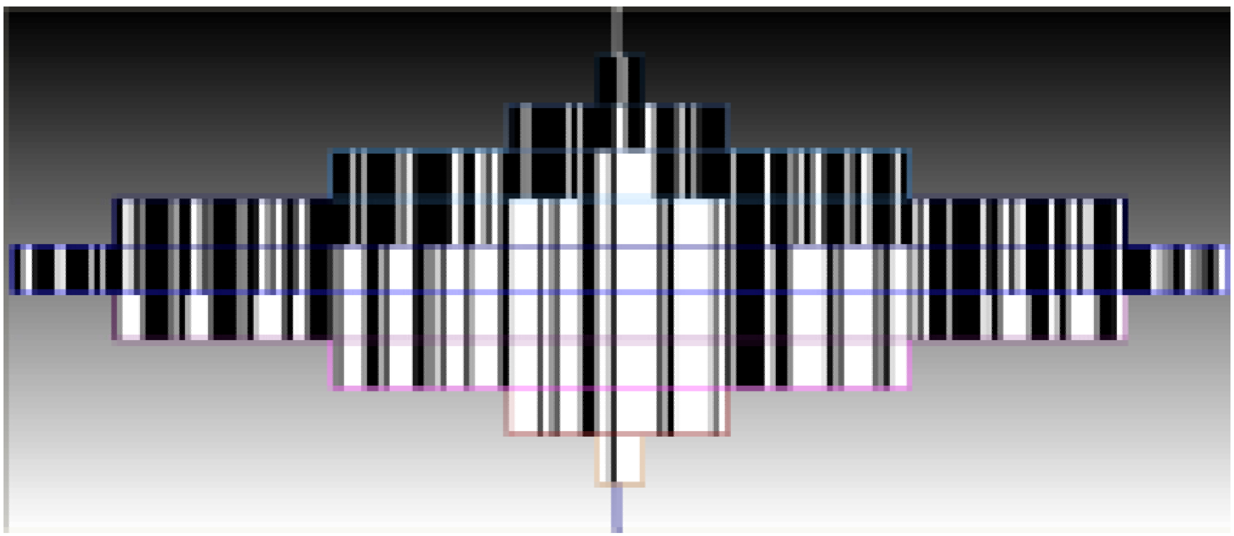


Figure 5. A MDF using procedure P_2 with $f(x_1, x_2, \ldots, x_{10}) = x_1$

Procedures *P1* and *P2* produce a border that can be very complex

Procedure P_3 and its pattern border

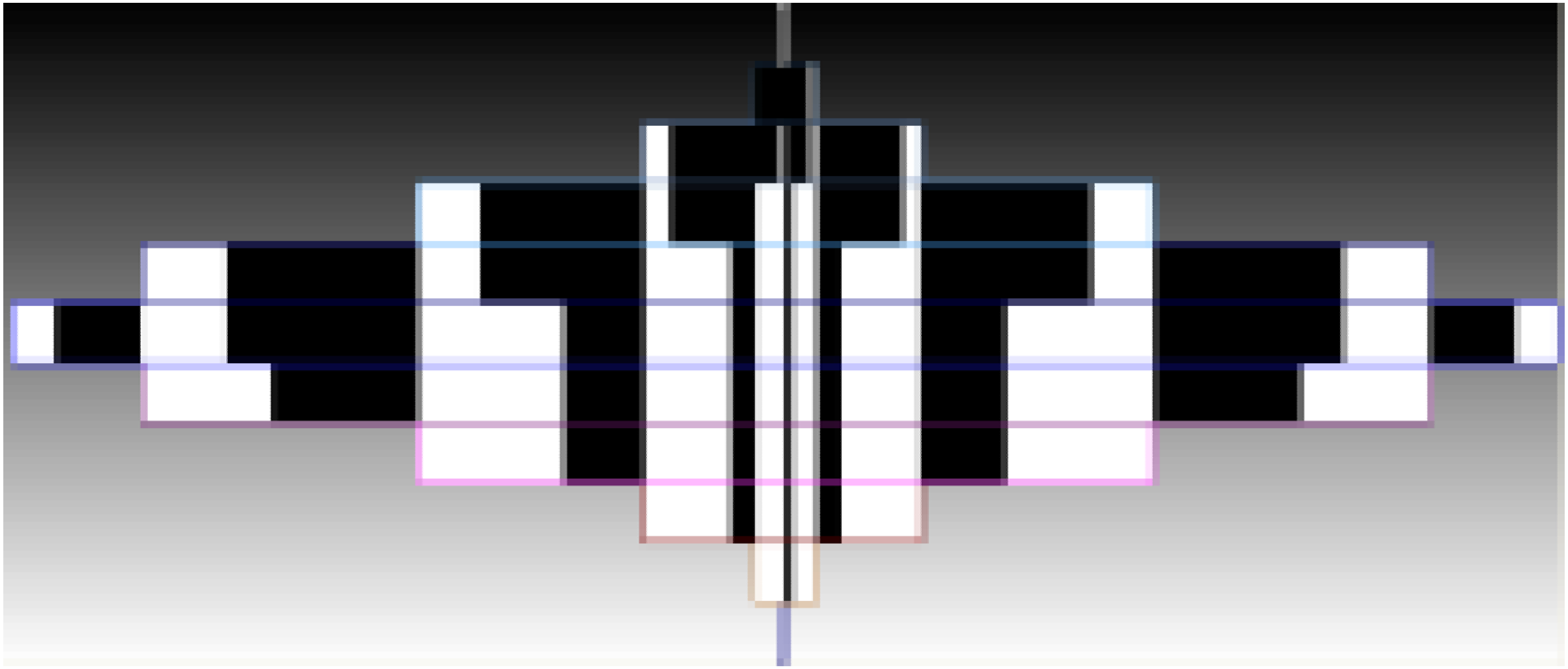


Figure 6. A MDF using procedure P_3 with $f(x_1,x_2,\ldots, x_{10}) = x_1$

P3 **moves all Hansel chains to the center of the disk by ordering chains using the level of the first "1" value in each chain. and checking that the disk architecture is preserved. In this way, two different functions will produce distinct visualizations.**

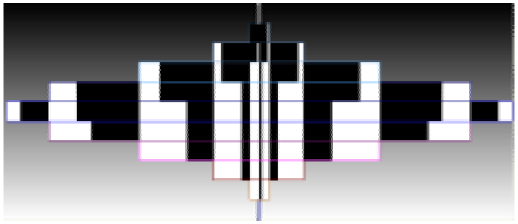
Procedure P_3

0 **Procedure *P3* is a derivative of *P2*.**

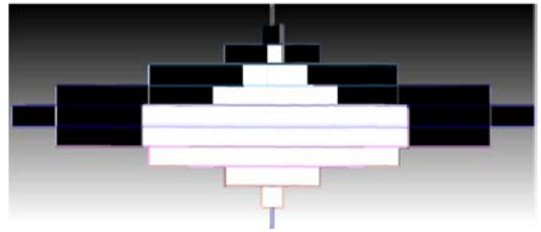
0 **After computing and placing the vectors using *P2*, every Hansel chain is given a value *L* equal to the level of the first 1 value present within the chain.**

0 **Next, every Hansel chain is moved so that the chain with the highest *L* value is located in the center of the disk so that the MDF structure is kept.**

0 **Using this procedure, classes are grouped within the MDF. Nevertheless to keep the MDF structure, the chains have to be placed in a position related to their length. This potentially introduces a complex border between classes because of a possible *gap* between groups of vectors within the same class.**



Procedure P_4



- The borders produced by P_3 can still be complex.
- New data structure YYF and procedure P_4.
- In the YYF, the movement of all chains is based on only the level of the first 1 in each chain. Additionally, this structure allows filling the gaps between the groups.
- The set of chains is sorted from left to right according to the indicated level and providing a clear, simple border between the two classes of Boolean vectors.
- Procedure *P4* extends every Hansel chains created before placing them according to the same method used in *P3*.
- The first step consists in *extending the Hansel chain* with elements extended in relation to the edge elements both up and down.

Final simple pattern border by using P_4, YYF

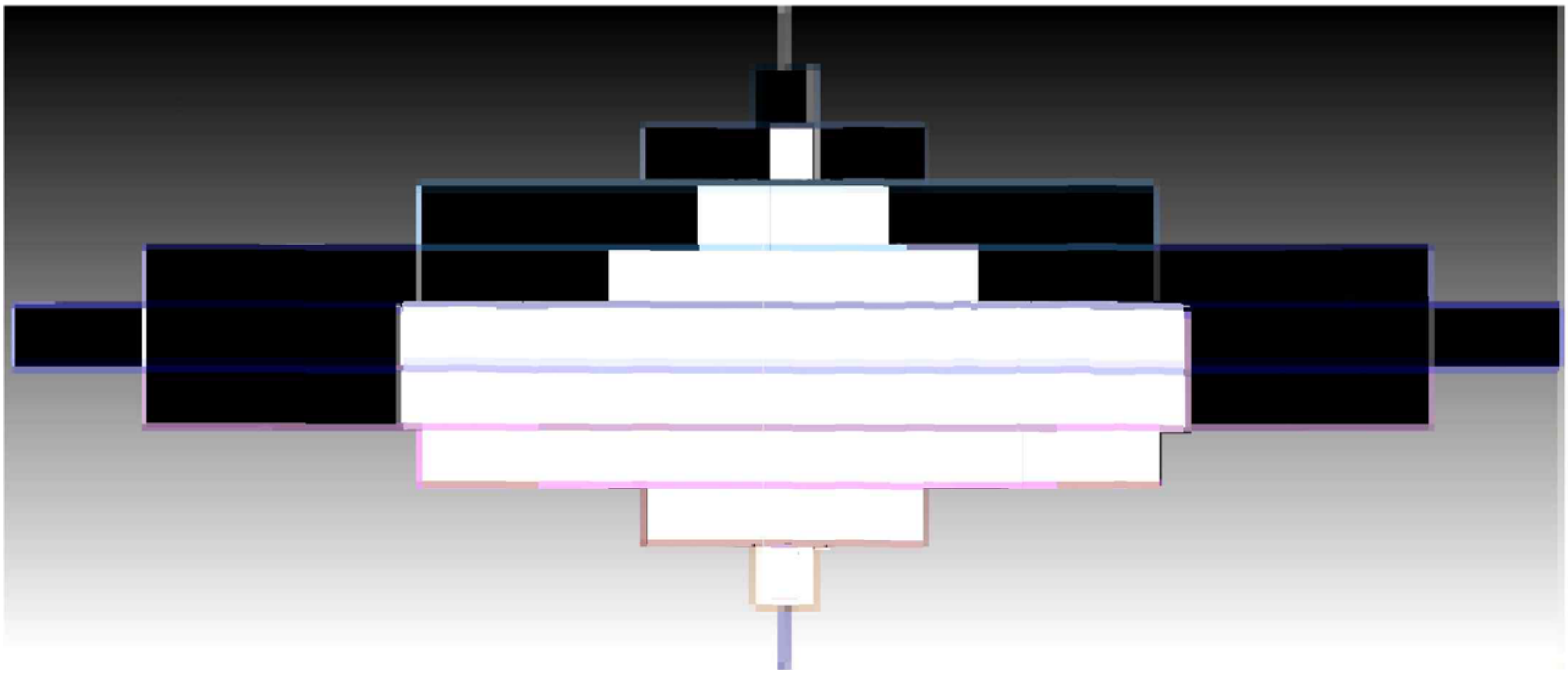


Figure 7. YYF using procedure P_4 with $f(x_1, x_2, \ldots, x_{10}) = x_1$

Edge elements and chain expansion

- ***Edge elements*** **are Boolean vectors that form the border between two classes on the chain.**
- **To extend a chain up, we try to find the first vector belonging to class 1 above the edge element.**
  - **Having the edge vector x we look for a vector y, $y \geq x$, if no such y is found on the level just above the x level, we then add a vector z from class 0 so that $z \geq x$ and so that the path to the first vector $y \geq z$ of class 1 is minimized.**
- **We repeat these steps until we find a vector y from the 1 class and add it to the chain.**
  - **To expand a chain down, we apply the same steps reversing the relation $y \geq x$ and swapping the classes 0 and 1.**
  - **Using this procedure, we duplicate some of the vectors which would display them more than once. This is justified because we keep a consistent relative relationship between the vectors.**

Yin Yang form

- **Once the chains are expanded, a value *L* is assigned to each chain. Then, the chains are sorted with regard to this value.**
- **This approach visualizes a border between classes 0 and 1.**
  - **To visualize the borders between the two classes of elements, we moved the data out of the MDF structure.**
  - **In the YYF vectors are ordered vertically in the same way as in the MDF but they are not centered anymore.**
- **All vectors are moved with regard to the data in order to visualize to the border between the two classes.**
- **In this way, a clear border will appear, class 1 being above class 0 thus giving the YYF the *"Yin Yang"-like* shape □, that responsible for its name.**

Outline

Part 1.Visual Data Mining Process

1.1. Visualization for Data Mining vs. Visual Data Mining

1.2. Challenges

1.3. Visual Relational Data Mining and Expert Mining

Part 2. Case studies

2.1. Fraud and Deception

2.2. Visual Correlation and Terrorism

2.3. Visual Reasoning

Part 3. Visual Data Mining for Binary Data

3.1. Methods

3.2. Case study: Breast cancer

Experiment With A Boolean Data Set

- **Breast cancer data, about 100 cases with an almost equal number of benign and malignant results.**
  - **Each case was described by 10 binary characteristics retrieved from mammographic X-ray images [Kovalerchuk, Vityaev & Ruiz, 2001].**
- **The goal of experiment was to check the monotonicity of this data set which is important from both radiological and visualization viewpoints.**
- **This procedure permits the comparison of multiple functions and data sets.**
- **All cases in the same visuallayer have the same number of cancer positive symptoms, but the symptoms themselves can be different. Light grey areas indicate monotonic expansion of benign cases to lower layers for each benign case and dark grey areas indicate monotonic expansion of malignant cases to upper layers for each malignant case.**

Breast cancer visual data mining

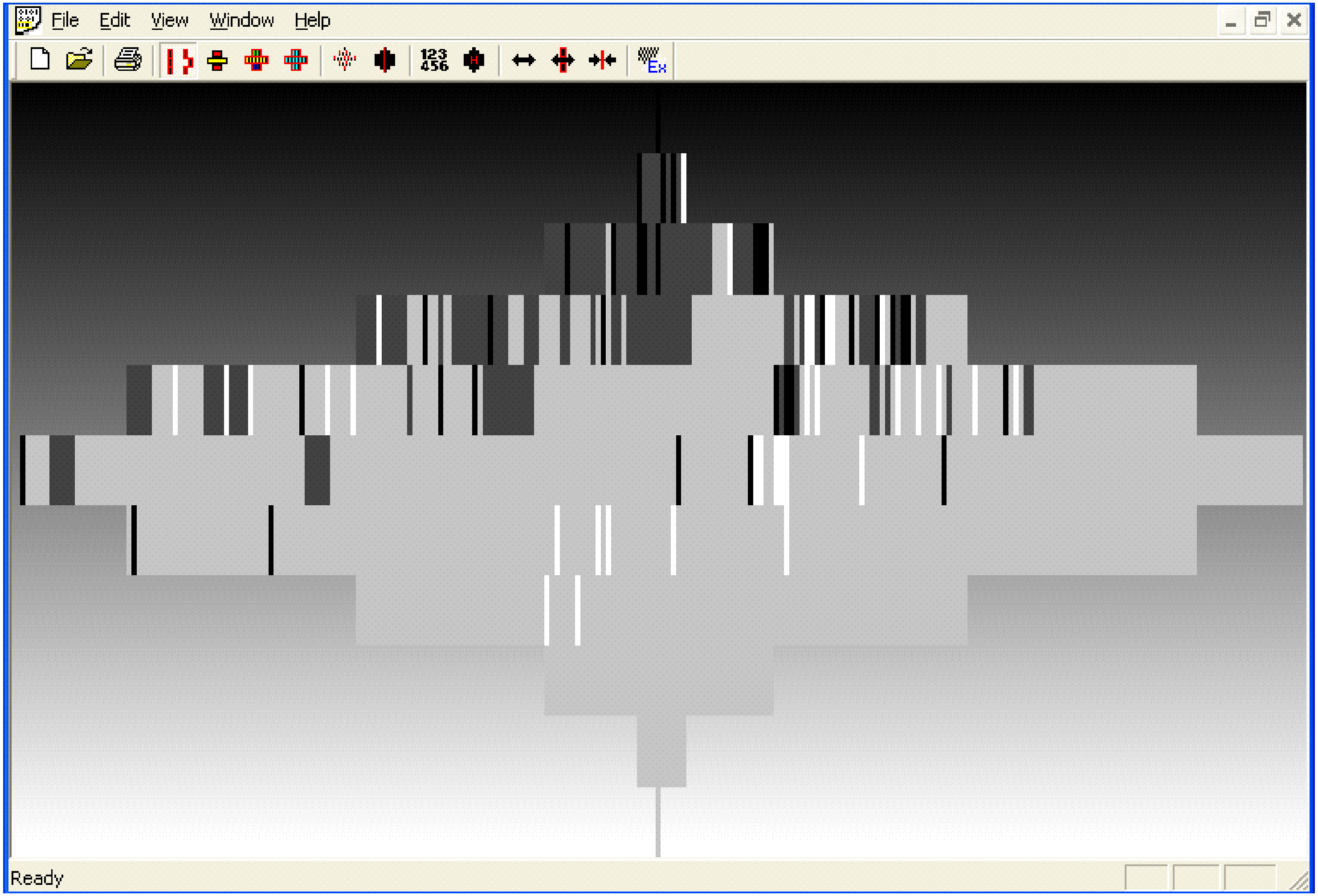


Figure 8. Breast cancer cases based on characteristics of X-ray images visualized using fixed location procedure P_1

Breast cancer visual data mining

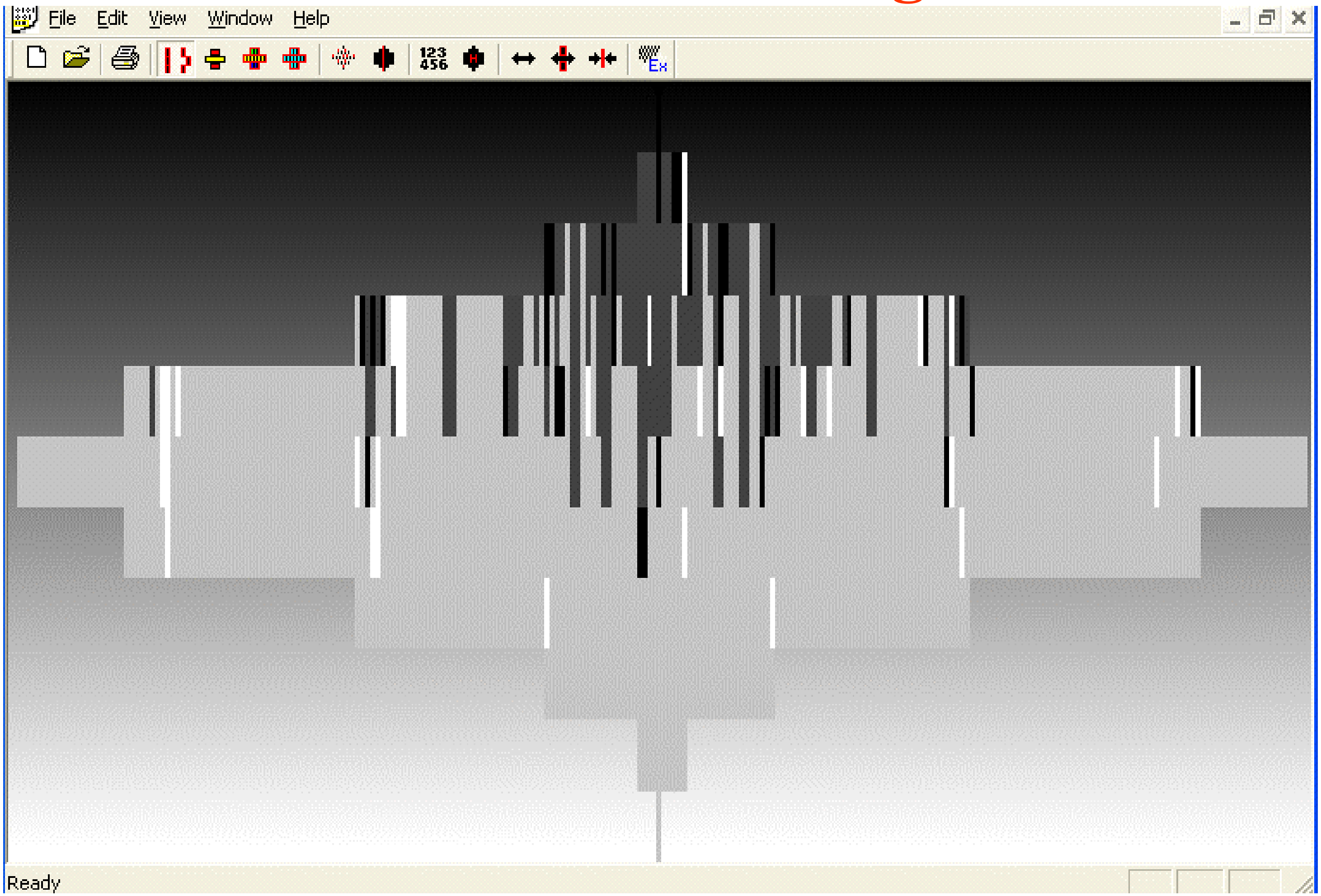


Figure 9. Breast cancer cases visualized using procedure *P2*

Comparison of Processes P_1 and P_2

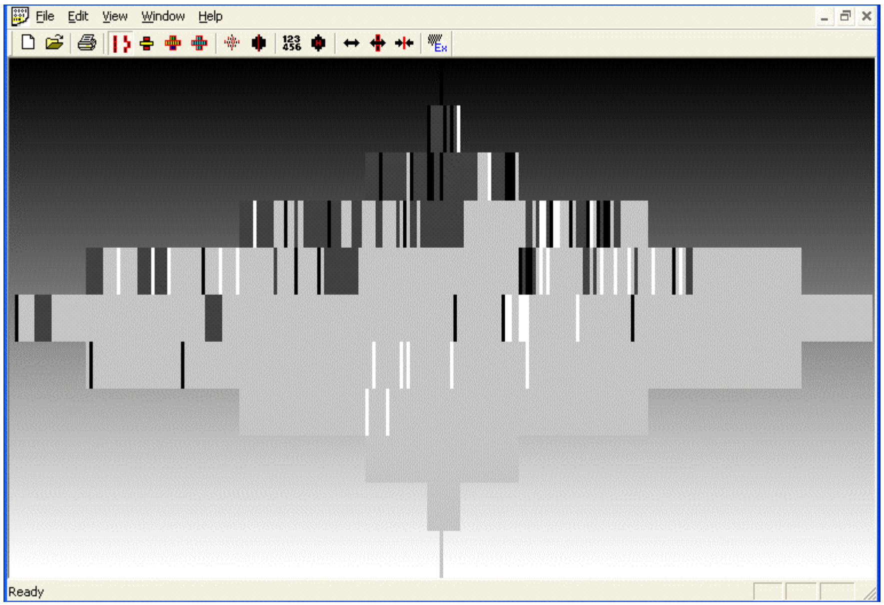


Figure 8. Breast cancer cases based on characteristics of X-ray images visualized using fixed location procedure P_1

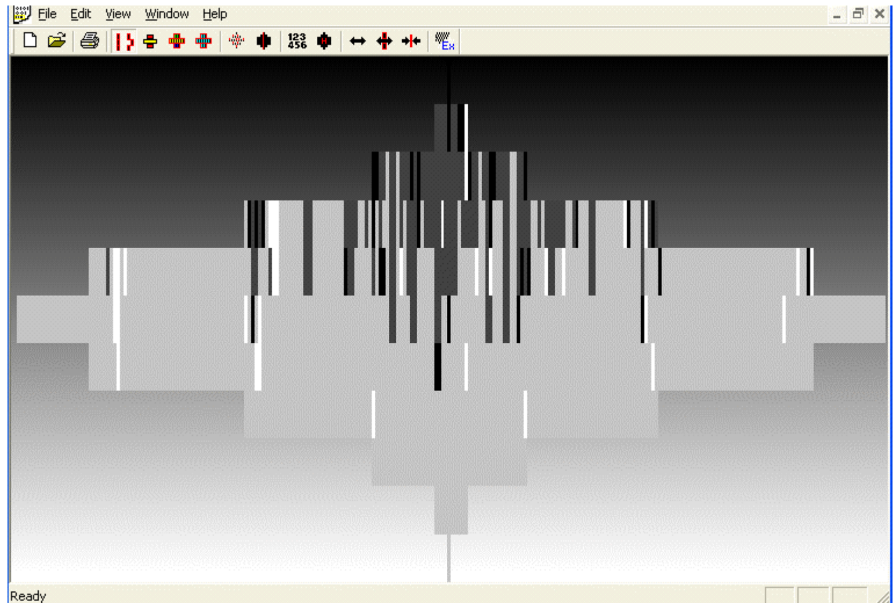


Figure 9. Breast cancer cases visualized using procedure P_2

Breast cancer visual data mining

- Benign cases are lined up monotonically. That is, each benign case below a given benign case contains only a part of its positive cancer characteristics.
- Similarly malignant cases (bars) are also lined up monotonically.
  - Thus, a malignant case above a given malignant case contains more positive cancer characteristics than the given malignant case.
  - The vertical lines (chains) that contain both benign and malignant cases are most interesting for further analysis.

Breast cancer visual data mining

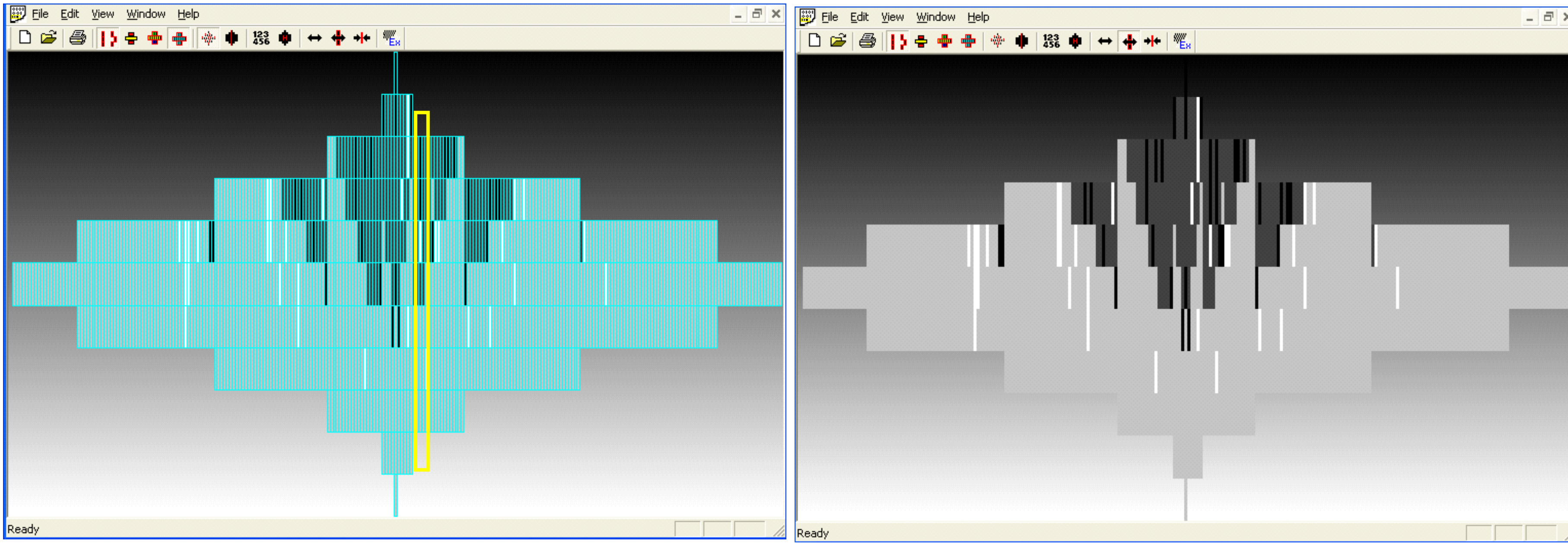


Figure 11. Breast cancer cases visualized using procedure P_3 with cases shown as bars with frames. See also color plates.

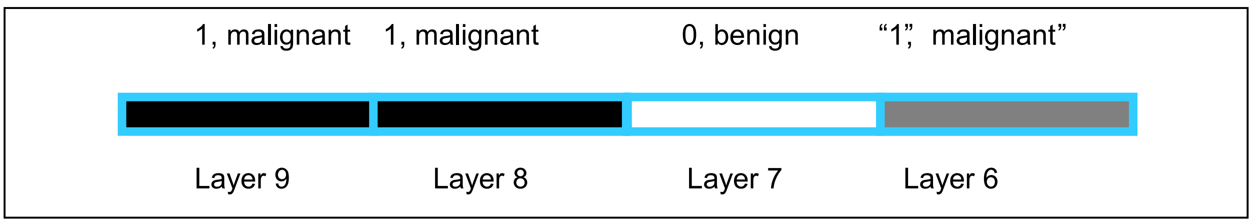
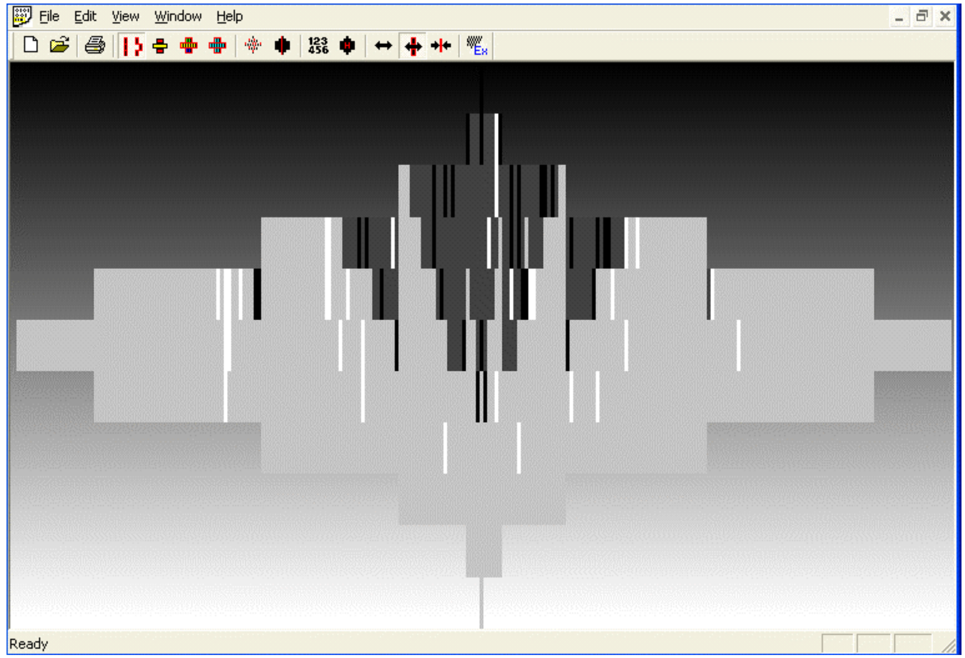


Figure 12. Fragment of individual chain with violated monotonicity

Visual discovery of structural inconsistencies

- **Figures analyzed reveal that there are *inconsistencies* with monotonicity for several of the cases.**
- **The “white” case is an *inconsistent case* if there are black and dark grey bars (areas) above and below it.**
- **The “black” case is inconsistent if there are white and light grey cases above and below it.**



Cases are shown without frames

Monotonization: building consistent meaningful pattern

- **This visualization permits us building different monotone Boolean functions interactively and visually for situations with inconsistencies.**
- **The first way to do this is to find all white inconsistencies and convert all elements below them to white bars.**
- **This process is called a *white precedence monotonization.***
- **Similarly, we can use a *black precedence monotonization* that converts all white and light grey elements above inconsistent black cases to black.**

3-D versions of the Monotone Boolean Visual Discovery

- **The first version uses only vertical surfaces and is quite similar to the 2-D versions.**
- **The second version uses both vertical and horizontal surfaces. 3-D versions have several advantages over 2-D versions.**
- **The first one is the ability to increase the dimensionality *n* that can be visualized. It is done by using front, back and horizontal surfaces of disks and by grouping similar Hansel chains and by showing only "representative" chains in the global disk views.**
- **More detail can be provided by changing camera location, which permits one to see the back side of the disks combined with the semantic zoom that permits one to see all chains not only the "representative" ones when the camera closes up on the disk.**

3-D versions of the Monotone Boolean Visual Discovery

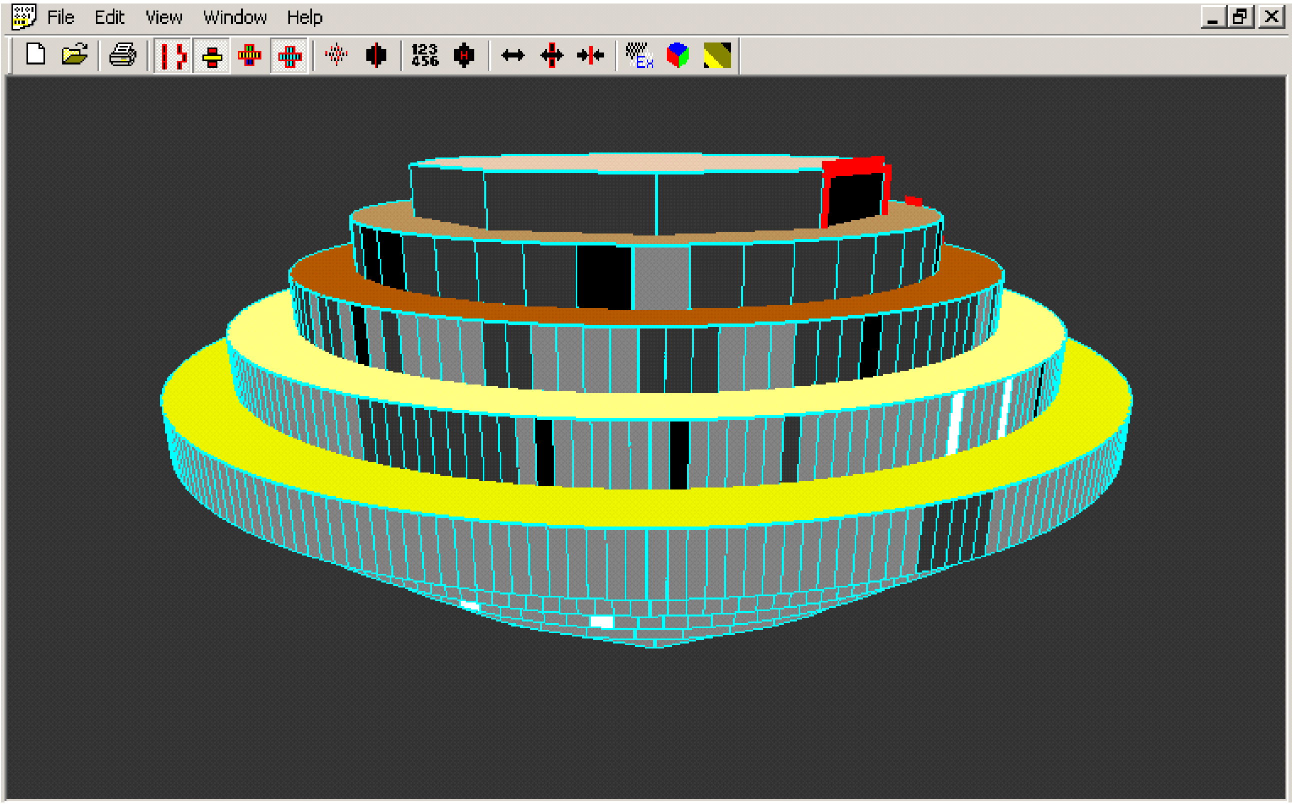


Figure 13. A 3-D version of Monotone Boolean Visual Discovery with only vertical surface used

3-D versions of the Monotone Boolean Visual Discovery

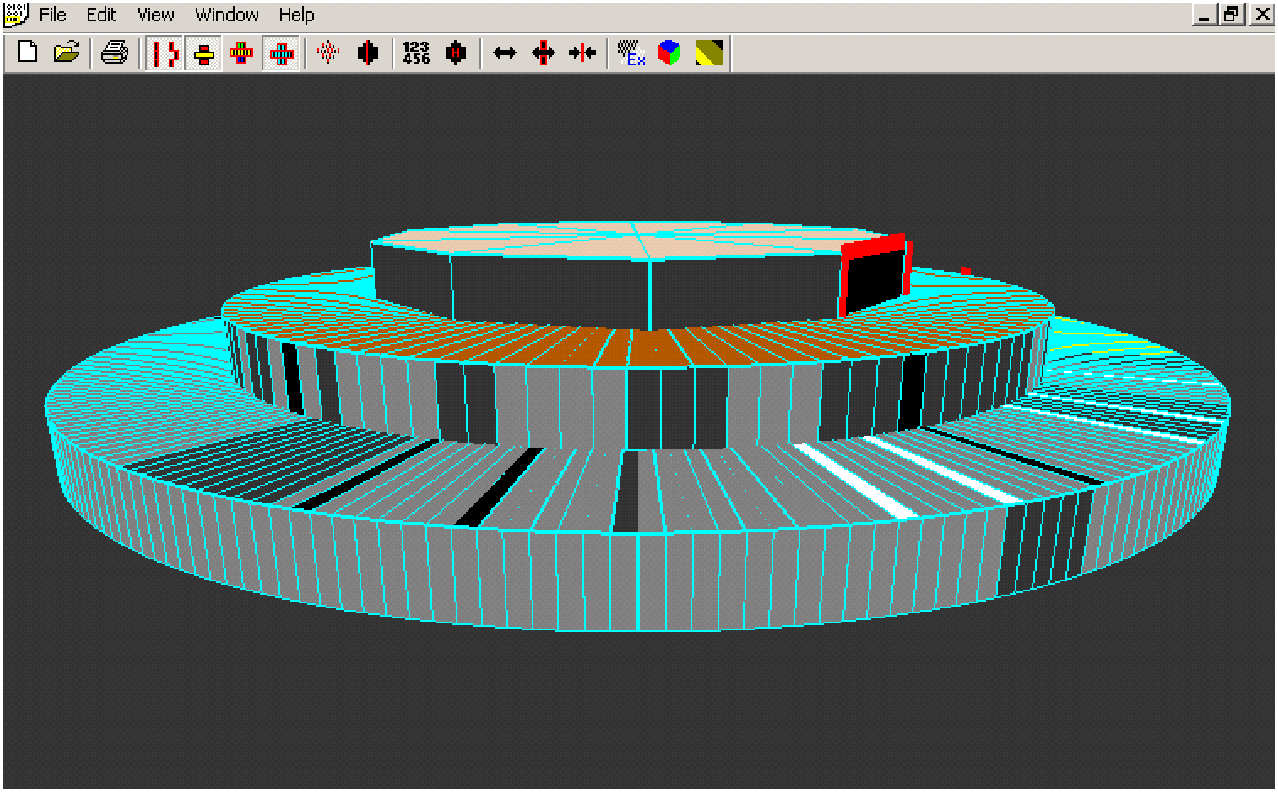


Figure 14. A 3-D version of Monotone Boolean Visual Discovery with vertical and horizontal surfaces used. See also color plates.

3-D versions of the Monotone Boolean Visual Discovery

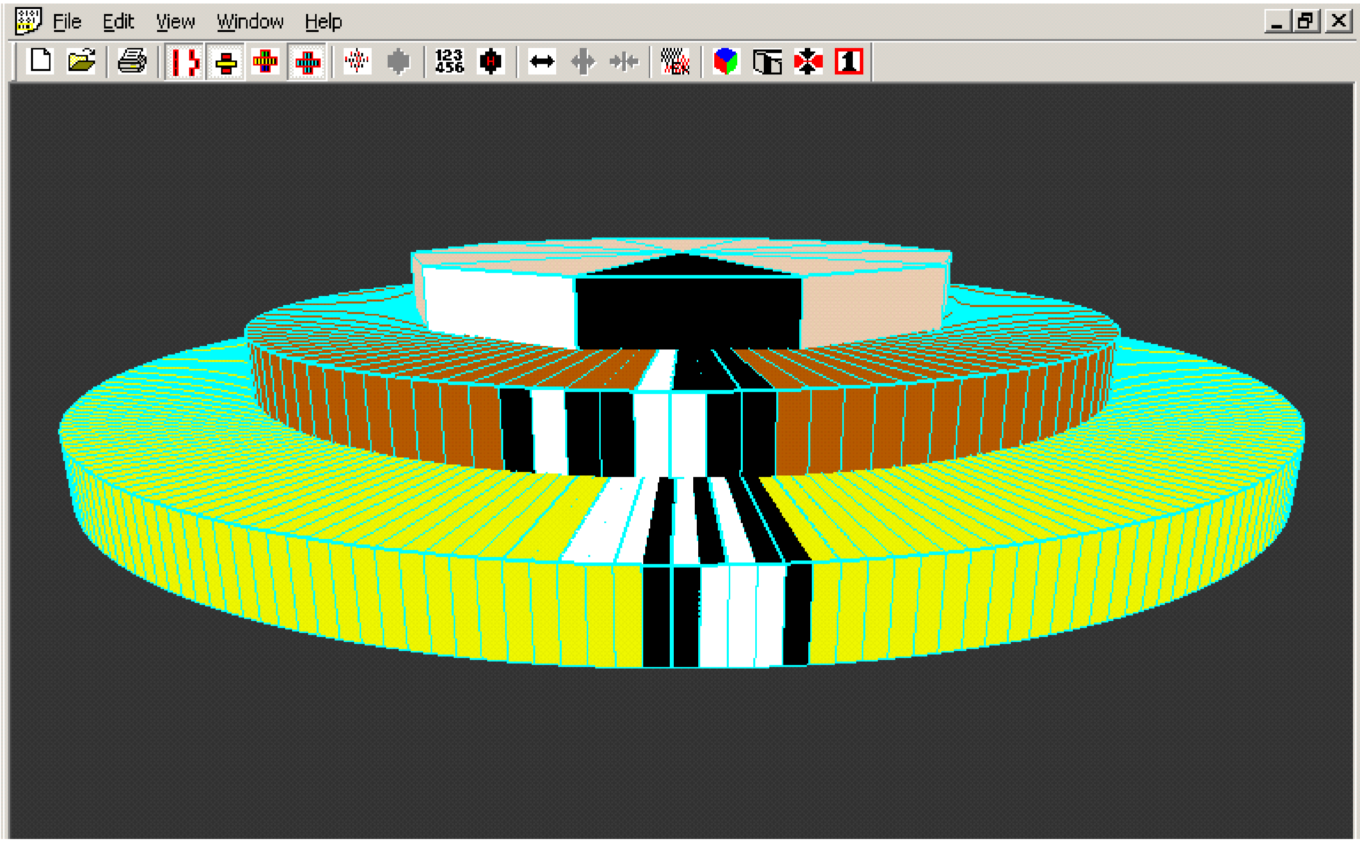


Figure 15. A 3-D version of Monotone Boolean Visual Discovery
with grouping Hansel chains. See also color plates.

Appendix: Data Structures And Formal Definitions

0 **A Boolean vector is an ordered set of *n* Boolean values 0 and 1.**

- **1011100100 -- 10-dimensional Boolean vector.**

0 **We can assign a set of properties to a Boolean vector such as its size (dimension *n*) and its norm.**

0 **The level of a Boolean vector (Boolean norm)**

- **the sum of the components of the Boolean vector.**
- **||0000000000||=0; ||1111111111||=10.**

0 **These norms are used splitting the set of vectors into *n* + 1 levels.**

Appendix: Data Structures and Definitions

- **Boolean vectors can be represented as vertices of a cube (or a hypercube E^n if n > 3).**
  - E^1 ={0; 1}; E^2= {00; 01; 10; 11}.

- **A Boolean function f is *monotone* if for any two Boolean vectors $x = (x_1, x_2, \ldots x_n)$ and $y = (y_1, y_2, \ldots, y_n)$ such that x precedes y, $f(y) \geq f(x)$, that is**

$$\forall\, i \in \{1, \ldots, n\} \quad x_i \geq y_i \Rightarrow f(y) \geq f(x).$$

- **Monotone Boolean functions divide the set of Boolean vectors into two classes:**
  - **vectors assigned to the value 0 and vectors assigned to the value 1, thus forming a border between the two classes.**

Hansel chains

- **Hansel chains provide a way to browse a Binary hypercube without overlapping** [Hansel, 1966; Kovalerchuk et al.,1996].
- **Recursive process of building Hansel chains**

The Hansel chain for n=1 is (0; 1).

The Hansel chains for *n*=2 are {(00; 01; 11), (10)}.

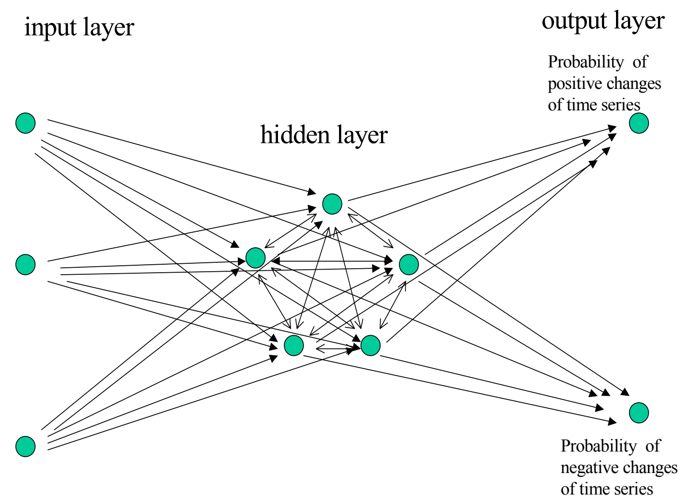
- **To obtain the Hansel chains for the level 2, we first duplicate the Hansel chains of the level 1 by generating two identical sets *G*=(0;1), *G*=(0; 1) and then by adding the prefix of 0 to the first set and then by adding the prefix 1 to the second set. This results in two sets Emin = (00; 01) and E_{max} = (10; 11). We then cut the maximum element of E_{max} and add it to E_{min}.**

By repeating those operations of duplicating and cutting for n=3, n=4 and so on, the Hansel chains are built for any size vectors.

Visual Data Mining: Research Topics

- How to convert Data Mining methods
  - Time Series analysis methods,
  - Neural Networks,
  - Decision Trees,
  - Discriminant Analysis,
  - Bayesian methods
  - Association rules

  to visual discovery tools?
- How to evaluate the results of visual discovery ?
- Methodology and architecture of visual data mining



Conclusion

- Visual data mining and reasoning have significant value and great potential for many areas.
- Visual data mining methods may play a special role in complex analysis because of their abilities to deal with patterns that are difficult to capture by analytical data mining methods.
- Visual methods may play special role in efficient integration of text based data mining with images and geospatial data because of their abilities to represent data in a visual form.
- In this tutorial we had shows how to discover and analyze binary data patterns visually. This can serve as prototype for making a similar progress with other data types. Special design of a monotone data structure is a key approach.
- Visual correlation and reasoning are promissing emerging areas of data mining.

References

0 Grinstein,G., Keim, D., Ward, M., Information Visualization and Visual Data Mining, IEEE Visualization 2004 Conference, Seattle, fusion.cs.uni-magdeburg.de/pubs/TVCG02.pdf (extensive references)

0 International Symposium of Visual Data Mining , July 6-8, 2005, London, http://www.graphicslink.demon.co.uk/IV05/

Beilken, C., Spenke, M., Visual interactive data mining with InfoZoom -the Medical Data Set. The 3rd European Conference on Principles and Practice of Knowledge Discovery in Databases, PKDD ’99, Sept. 15-18, 1999, Prague, Czech Republic. http://citeseer.nj.nec.com/473660.html

0 Ebert, D., Shaw, C., Zwa, A., Miller, E., Roberts, D. Two-handed interactive stereoscopic visualization. IEEE Visualization ’96 Conference.1996, pp. 205-210.

0 Goel, A., Visualization in Problem Solving Environments. Virginia Polytechnic Institute and State University, 1999.MS. Thesis, 64 p. www.amitgoel.com/vizcraft/docs/ masters_thesis.pdf

0 Groth, D., Robertson, E., Architectural support for database visualization, Workshop on New Paradigms in Information Visualization and Manipulation, 53-55,1998, .pp. 53-55.

0 Hansel, G., Sur le nombre des functions Bool´eenes monotones de n variables. C.R. Acad. Sci., Paris (in French), 262(20), pp.1088–1090, 1966

References

- Inselberg, A., Dimsdale, B., Parallel coordinates: A tool for visualizing multidimensional Geometry. Proceedings of IEEE Visualization ’90, Los Alamitos, CA, IEEE Computer Society Press, 1990, pp. 360–375.
- Keim,D., Ming C. Hao, Dayal, U., Meichun Hsu. Pixel bar charts: a visualization technique for very large multiattributes data sets. Information Visualization, March 2002, Vol. 1, N. 1, pp. 20–34.
- Kovalerchuk, B., Triantaphyllou, E., Despande, A.,Vityaev, E., Interactive Learning of Monotone Boolean Functions. Information Sciences, Vol. 94, issue 1-4, pp. 87–118, 1996.
- Kovalerchuk, B., Vityaev, E., Ruiz, J., Consistent knowledge discovery in medical diagnosis. IEEE Engineering in Medicine and Biology, (Special issue on Data Mining and Knowledge Discovery), v. 19, n. 426–37, 2000.
- Kovalerchuk, B., Vityaev, E., Ruiz, J., Consistent and complete data and “expert” mining in medicine. Medical Data Mining and Knowledge Discovery, Springer, 2001:238–280.
- Last, M., Kandel, A., Automated perceptions in data mining, invited paper. 1999, IEEE International Fuzzy Systems Conference Proceedings, Part I, Seoul, Korea, Aug 1999, pp.190–197.
- Post, F., T. van Walsum, Post, F., Silver, D., Iconic techniques for feature visualization. In Proceedings Visualization ’95, pp. 288–295, 1995.

References

0 Ribarsky, W., Ayers, E., Eble, J., Mukherja, S., Glyphmaker: creating customized visualizations of complex data. IEEE Computer, 27(7), pp. 57–64, 994.

0 Shaw, C., Hall, J., Blahut, C., Ebert, D., Roberts. A., Using shape to visualize multivariate data. CIKM'99 Workshop on New Paradigms in Information Visualization and Manipulation, ACM Press, 1999, pp. 17-20.

0 Ward, M., A taxonomy of glyph placement strategies for multidimensional data visualization. Information Visualization 1, 2002, pp. 194–210.

0 Kovalerchuk, B., and Vityaev E., 2000. Data Mining in Finance: Advances in Relational and Hybrid Methods, Kluwer Acad. Publ., Boston

0 Kovalerchuk B., and Schwing J. (eds), 2005. Visual and Spatial analysis: Advances in Data Mining, Reasoning, and Problem Solving, Springer

0 Geospatial Intelligence Information Visualization Program (GI2VIS) http://www.ic-arda.org/InfoExploit/ aquaint/gi2vis/ projects/cffvcaia.html

0 Van Der Putten, P., Van Someren, M., 2004, A Bias-Variance Analysis of a Real World Learning Problem: The CoIL Challenge 2000, Machine Learning, 57, 177–195

0 Kovalerchuk, B., Vityaev, E., 2003, Detecting patterns of fraudulent behavior in forensic accounting, In 7th Intern. Conf. "Knowledge-based Intelligent Information and Engineering Systems", Oxford, UK, Springer, part 1, pp. 502-509

Contact information

- borisk@cwu.edu
- http://www.cwu.edu/~borisk
- http://www.cwu.edu/~lmaglab

- **Kovalerchuk B., and Schwing J. (eds), 2005. Visual and Spatial analysis: Advances in Data Mining, Reasoning, and Problem Solving, Springer**
- http://www.cwu.edu/~borisk/bookVis/