DECOMPOSITION OF RELATIONS:

A NEW APPROACH TO CONSTRUCTIVE INDUCTION IN MACHINE LEARNING AND DATA MINING - AN OVERVIEW

Marek Perkowski

Portland State University

Data Mining Application for Epidemiologists

Control of a robot

FPGA

Machine Learning from Medical databases

VLSI Layout

- **This is a review paper that presents work done at Portland State University and associated groups in years 1989 - 2001 in the area of functional decomposition of multi-valued functions and relations, as well as some applications of these methods.**

Group Members

Current Students:

Anas Al-Rabadi

Faculty

Marek Perkowski

Alan Mishchenko

Collaborating Faculty

Bernd Steinbach

Lech Jozwiak

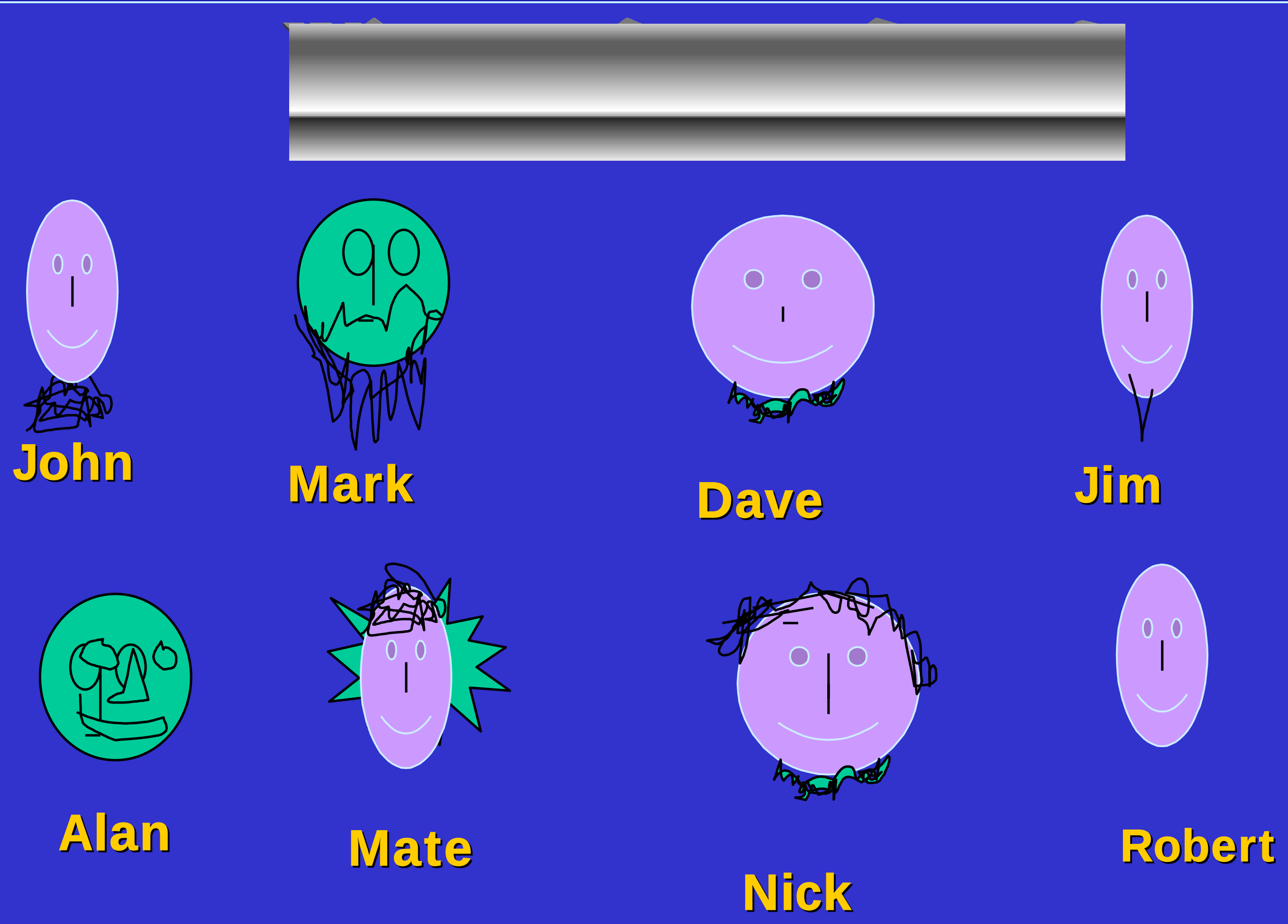
Martin Zwick

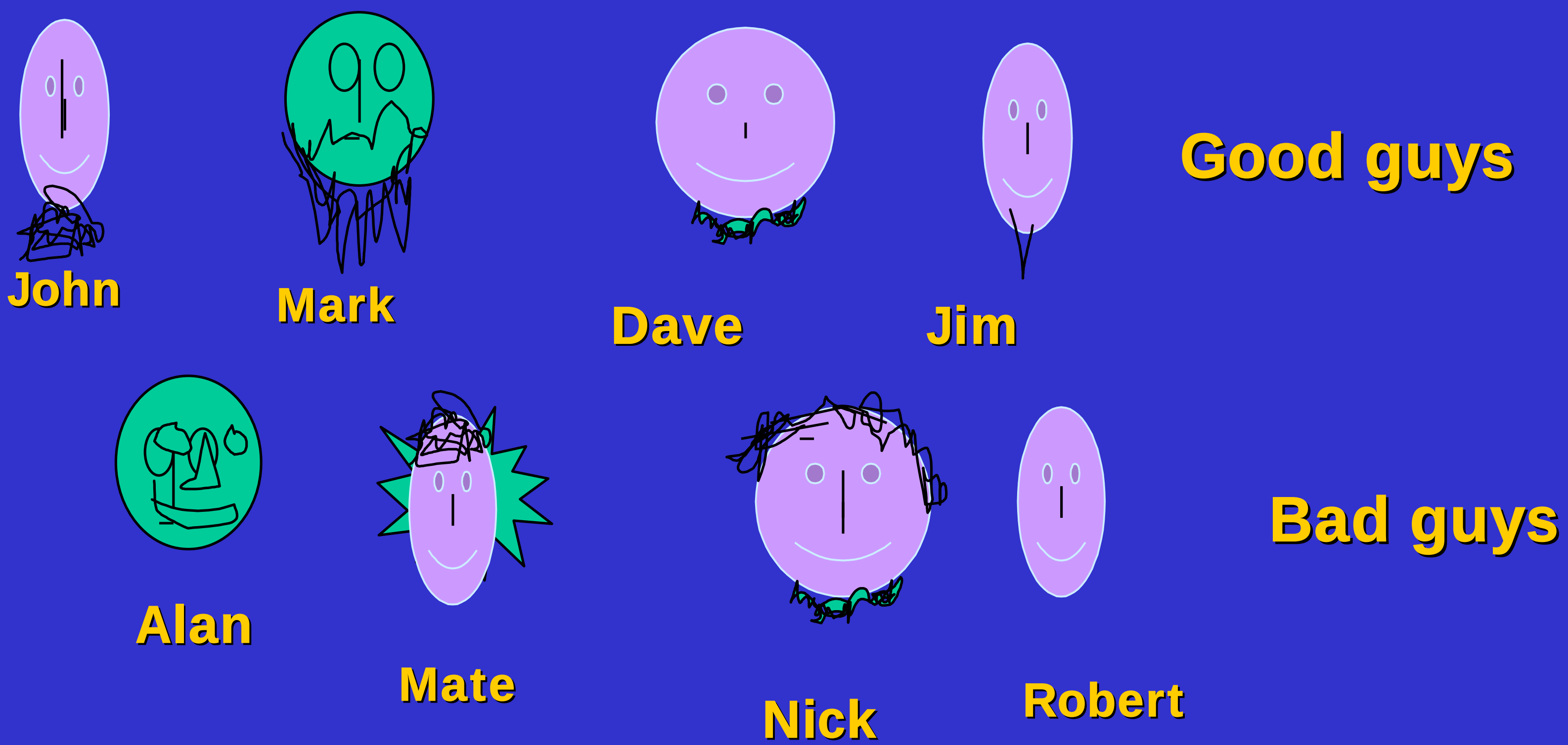
Previous Students:

Stanislaw Grygiel, Ph.D., **Intel**
Craig Files, Ph.D., **Agilent.**
Paul Burkey, **Intel**
Rahul Malvi, **Synopsys**
Michael Burns, **Vlsi logic**,
Timothy Brandis, **OrCAD**
Tu Dinh,
Michael Levy, **Georgia Tech**

Essence of logic synthesis approach to learning

Example of Logical Synthesis



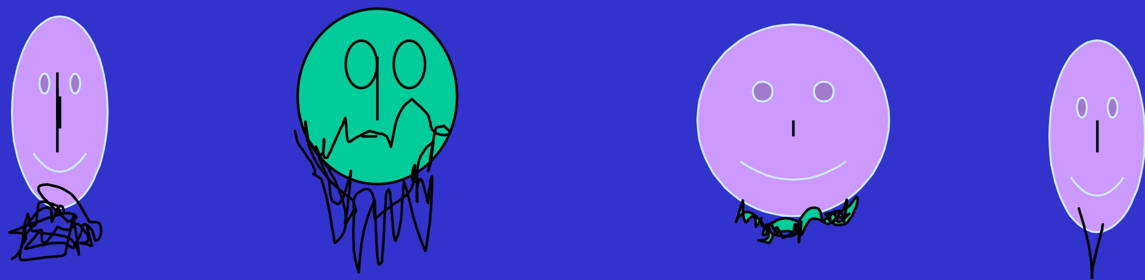


A - size of hair

B - size of nose

C - size of beard

D - color of eyes



John	Mark	Dave	Jim
A’ BCD	A’ BCD’	A’ B’CD	A’ B’CD

Good guys

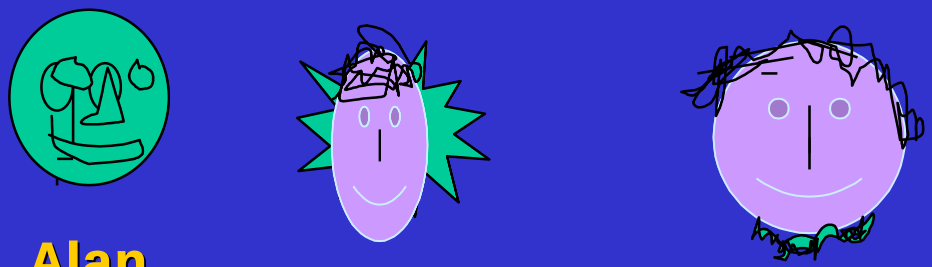
AB \ CD	00	01	11	10
00	-	-	1	-
01	—	-	1	1
11	-	—	-	-
10	-	-	-	-

A - size of hair

B - size of nose

C - size of beard

D - color of eyes



Alan | Mate | Nick | Robert

Bad guys

A' BC'D' | AB'C'D | ABCD | A' B'C'D

- A - size of hair
- B - size of nose
- C - size of beard
- D - color of eyes

AB \ CD	00	01	11	10
00	-	-	1	-
01	0	0	1	1
11	-	—	0	-
10	-	0	-	-

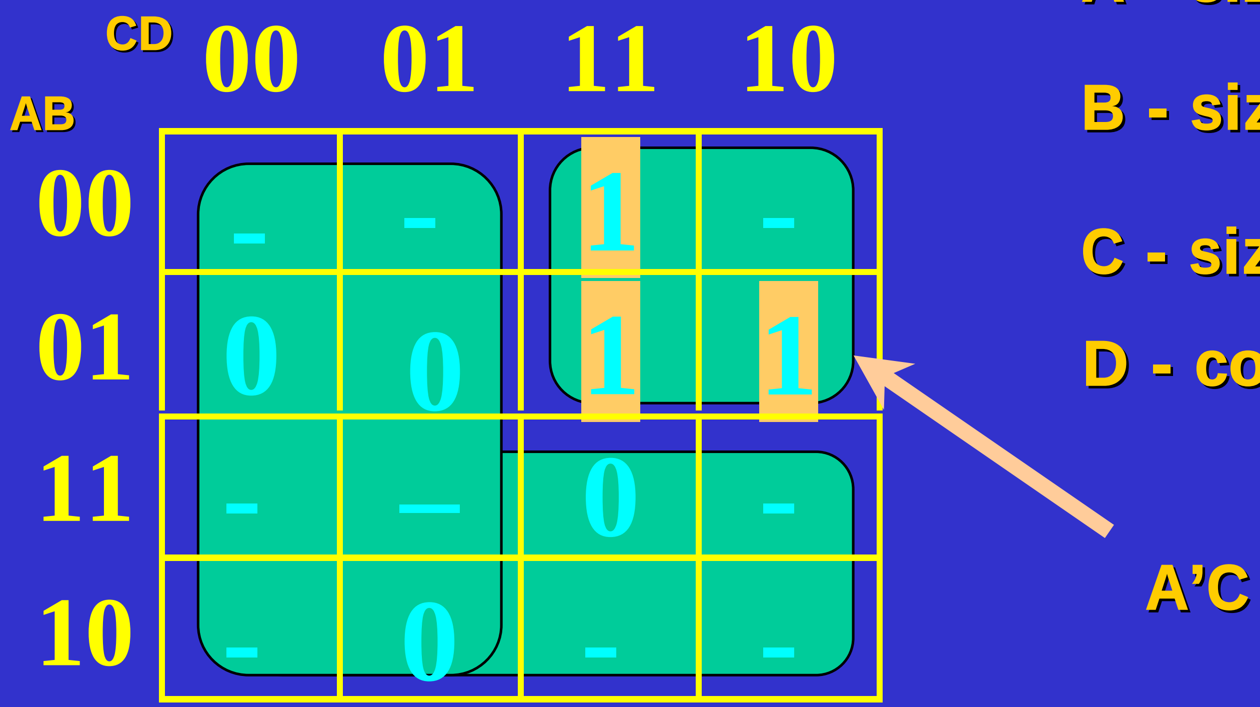
A'C

Generalization 1:

Bald guys with beards are good

Generalization 2:

All other guys are no good



A - size of hair

B - size of nose

C - size of beard

D - color of eyes

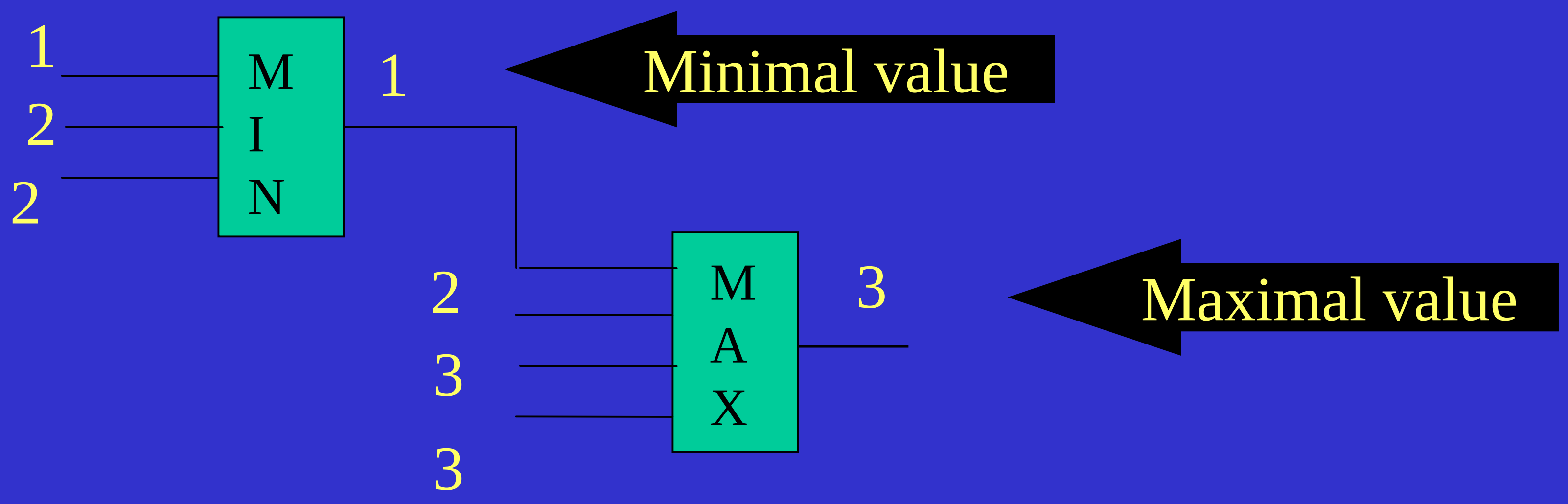
Short Introduction: multiple-valued logic

Signals can have values from some set, for instance {0,1,2}, or {0,1,2,3}

{0,1} - binary logic (a special case)

{0,1,2} - ***a ternary logic***

{0,1,2,3} ***- a quaternary logic,*** **<u>etc</u>**



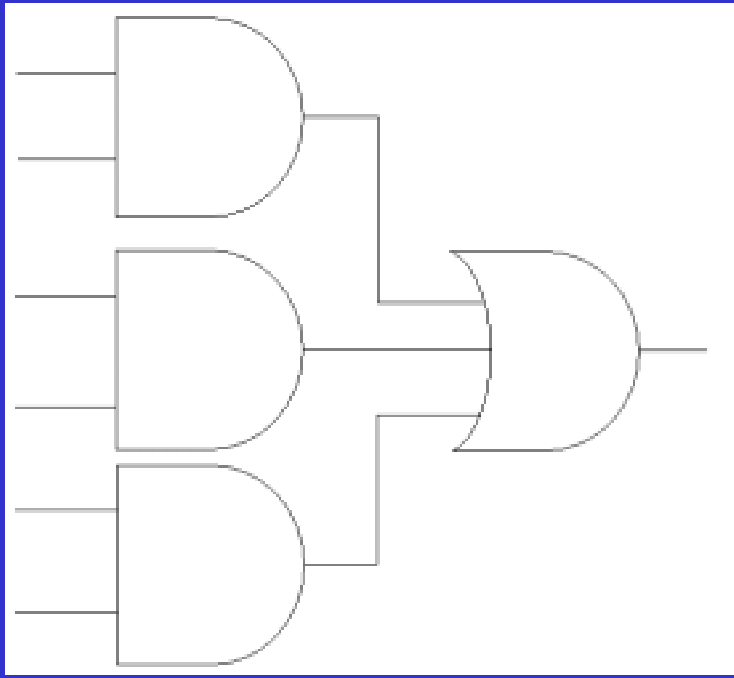
Types of Logical Synthesis

- Sum of Products
- Decision Diagrams
- Functional Decomposition

The method we are using

Sum of Products

AND gates, followed by an OR gate that produces the output. (Also, use Inverters as needed.)



Decision Diagrams

A Decision diagram breaks down a Karnaugh map into set of decision trees.

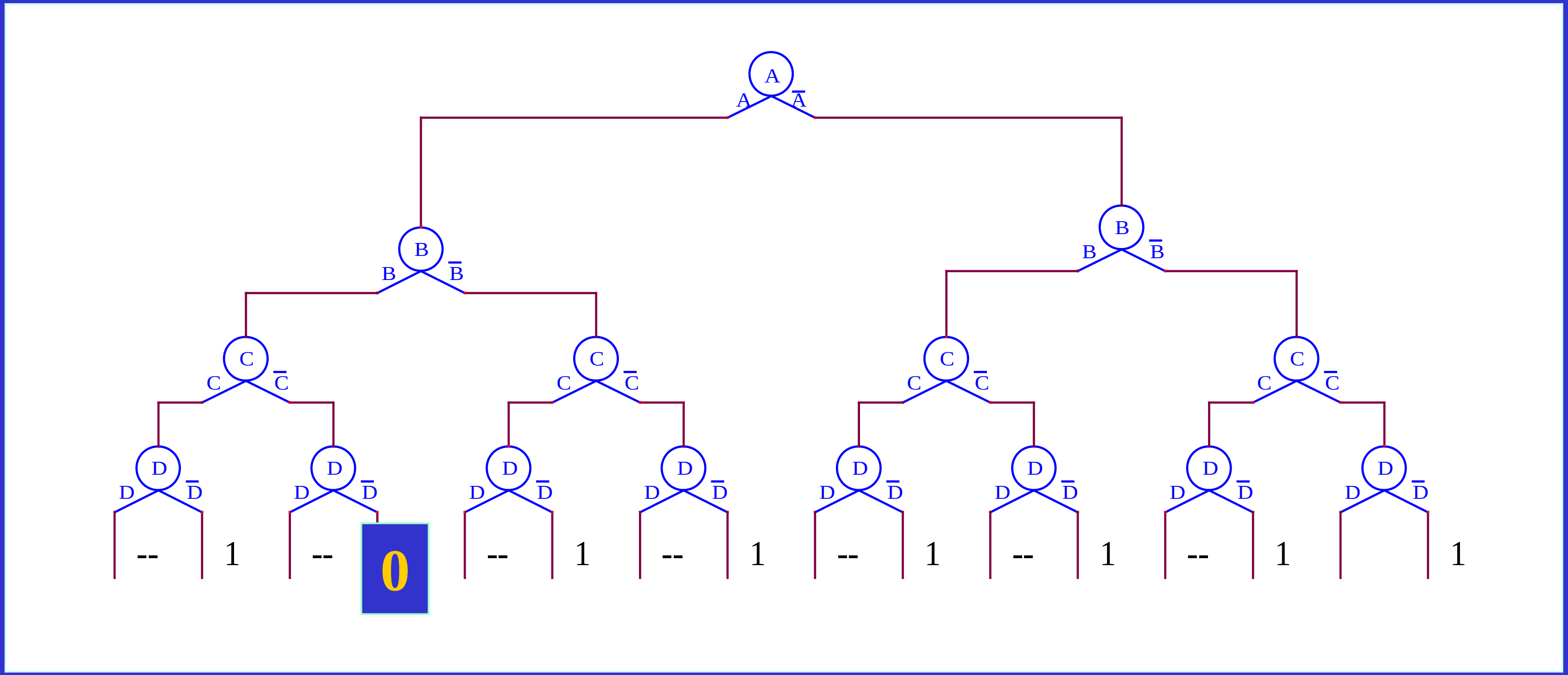
A decision diagram ends when all of branches have a yes, no, or do not care solution.

This diagram can become quite complex if the data is spread out as in the following example.

Example Karnaugh Map

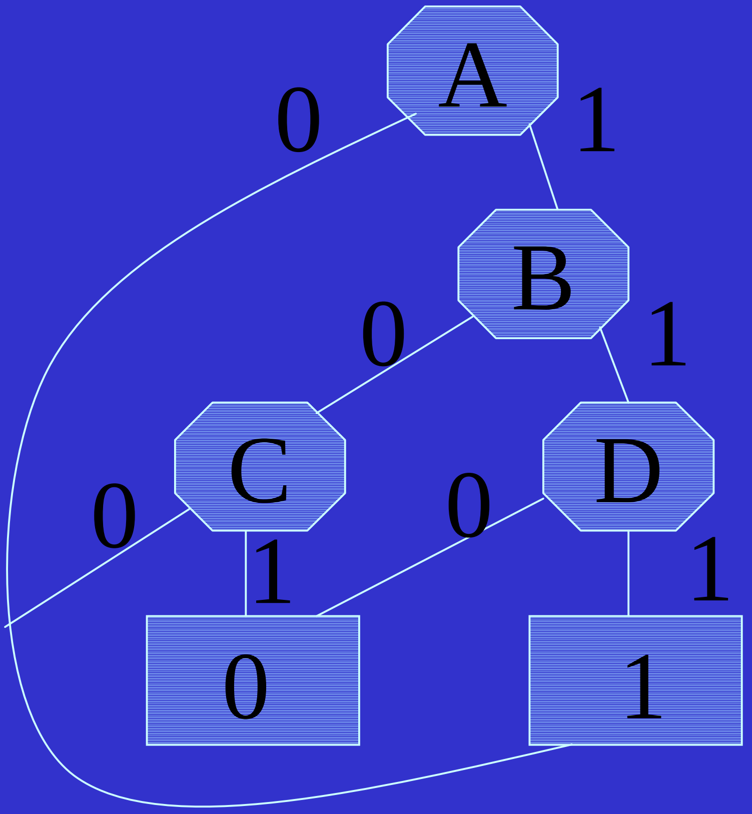
AB\CD	00	01	10	11
00	1	-	1	-
01	-	1	-	1
10	1	-	1	-
11	0	1	-	1

Decision Tree for Example Karnaugh Map



BDD Representation of function

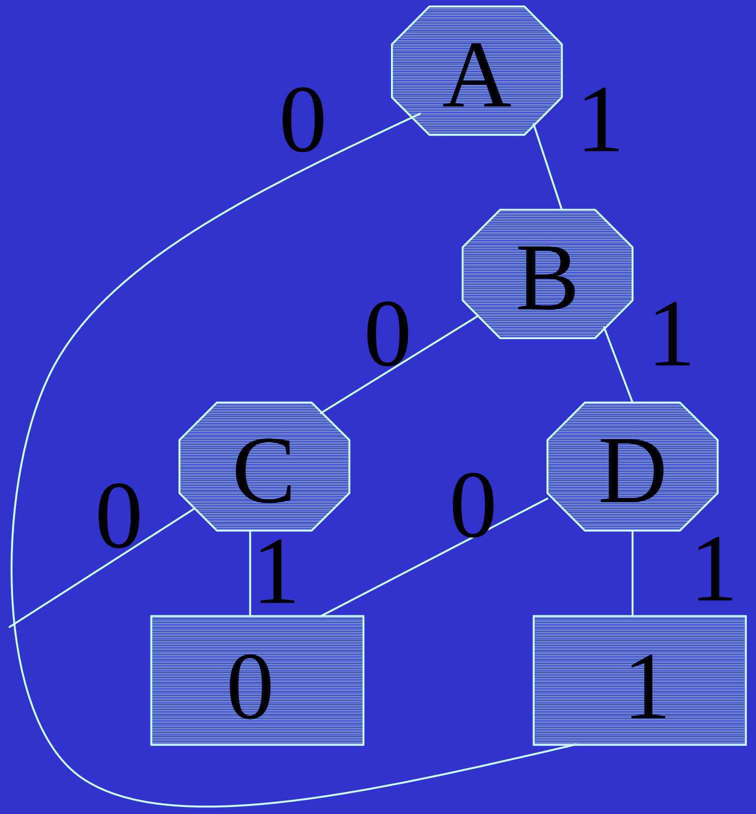
AB \ CD	00	01	11	10
00	-	1	1	-
01	1	-	1	1
11	0	1	-	0
10	-	1	0	-



Incompletely specified function

BDD Representation of function

AB \ CD	00	01	11	10
00	1	1	1	1
01	1	1	1	1
11	0	1	1	0
10	1	1	0	0

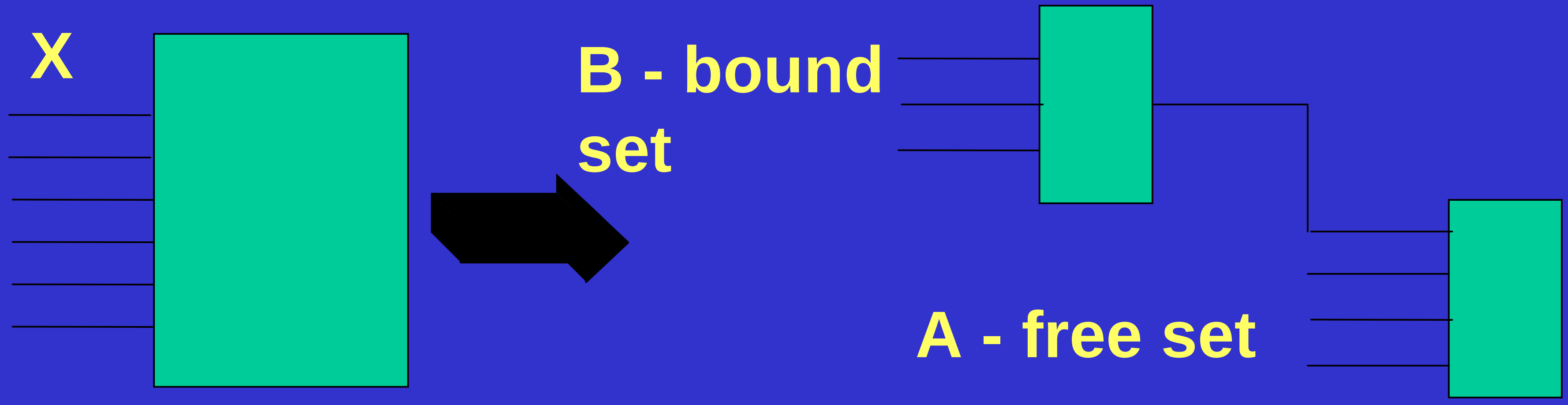


Completely specified function

Functional Decomposition

Evaluates the data function and attempts to decompose into simpler functions.

$$F(X) = H(\ G(B), A\ ),\quad X = A \cup B$$



if $A \cap B = \varnothing$, it is *disjoint decomposition*

if $A \cap B \neq \varnothing$, it is *non-disjoint decomposition*

Pros and cons

In generating the final combinational network, BDD decomposition, based on multiplexers, and SOP decomposition, trade flexibility in circuit topology for time efficiency

Generalized functional decomposition sacrifices speed for a higher likelihood of minimizing the complexity of the final network

Overview of data mining

What is Data Mining?

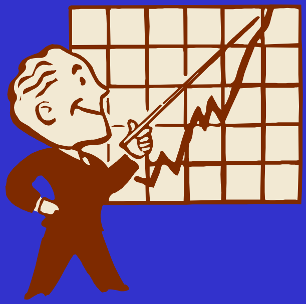
Databases with millions of records and thousands of fields are now common in business, medicine, engineering, and the sciences.

To extract useful information from such data sets is an important practical problem.

Data Mining is the study of methods to find useful information from the database and use data to make predictions about the people or events the data was developed from.

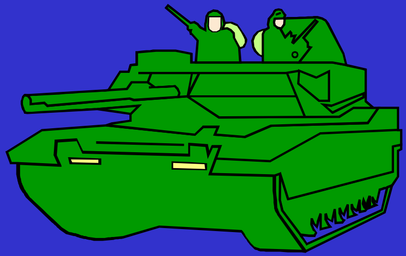
Some Examples of Data Mining

1) Stock Market Predictions



2) Large companies tracking sales

3) Military and intelligence applications



Data Mining in Epidemiology

Epidemiologists track the spread of infectious disease and try to determines the diseases original source

Often times Epidemiologist only have an initial suspicions about what is causing an illness. They interview people to find out what those people that got sick have in common.

Currently they have to sort through this data by hand to try and determine the initial source of the disease.

A data mining application would speed up this process and allow them to quickly track the source of an infectious diseases

Types of Data Mining

Data Mining applications use, among others, three methods to process data

1) Neural Nets

2) Statistical Analysis

3) Logical Synthesis ←

The method we are using

A Standard Map of function 'z'

Bound Set

Free Set

a b \ c	0	1	2
0 0	-	-	-
0 1	-	-	-
0 2	1	0 , 1	-
1 0	-	-	2
1 1	-	1	2
1 2	-	1	-
2 0	-	-	-
2 1	-	-	0
2 2	-	2 , 3	-

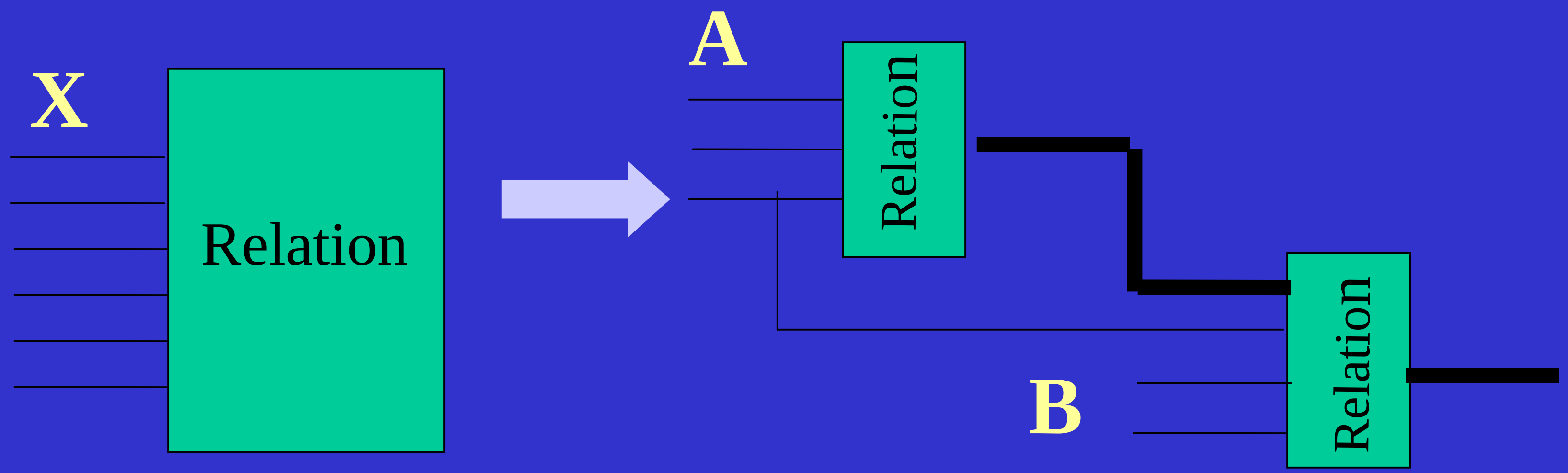
z

Columns 0 and 1 and columns 0 and 2 are compatible

column compatibility = 2

Decomposition of Multi-Valued Relations

$F(X) = H(G(B), A)$, $X = A \cup B$



if $A \cap B = \varnothing$, it is ***disjoint decomposition***

if $A \cap B \neq \varnothing$, it is ***non-disjoint decomposition***

Forming a CCG from a K-Map

Bound Set

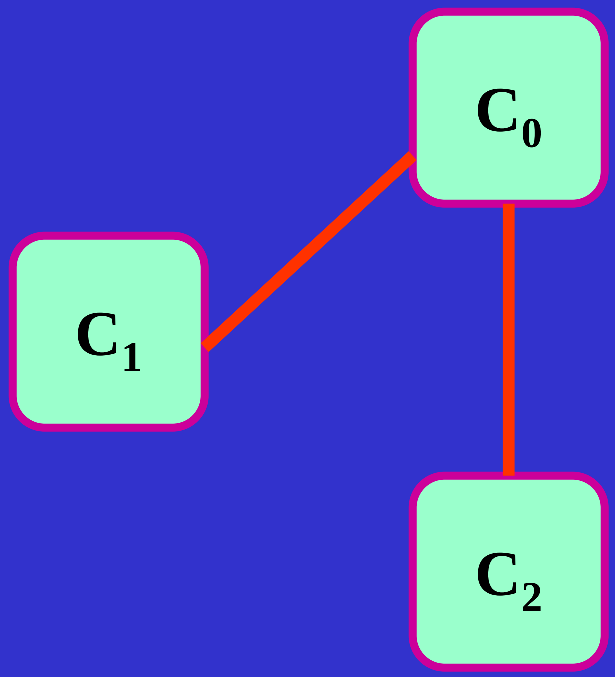
a b \ c	0	1	2
0 0	-	-	-
0 1	-	-	-
0 2	1	0 , 1	-
1 0	-	-	2
1 1	-	1	2
1 2	-	1	-
2 0	-	-	-
2 1	-	-	0
2 2	-	2 , 3	-

Free Set

z

Columns 0 and 1 and columns 0 and 2 are compatible

column compatibility index = 2



Column Compatibility Graph

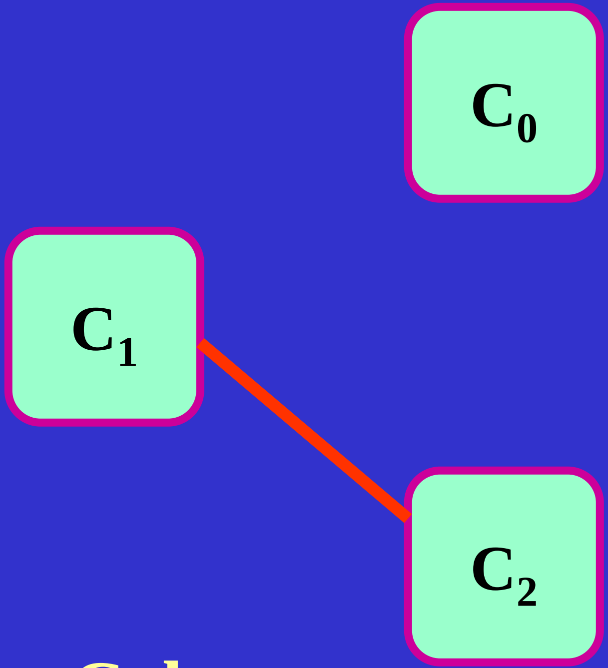
Forming a CIG from a K-Map

a b \ c	0	1	2
0 0	-	-	-
0 1	-	-	-
0 2	1	0 , 1	-
1 0	-	-	2
1 1	-	1	2
1 2	-	1	-
2 0	-	-	-
2 1	-	-	0
2 2	-	2 , 3	-

z

Columns 1 and 2 are incompatible

chromatic number = 2

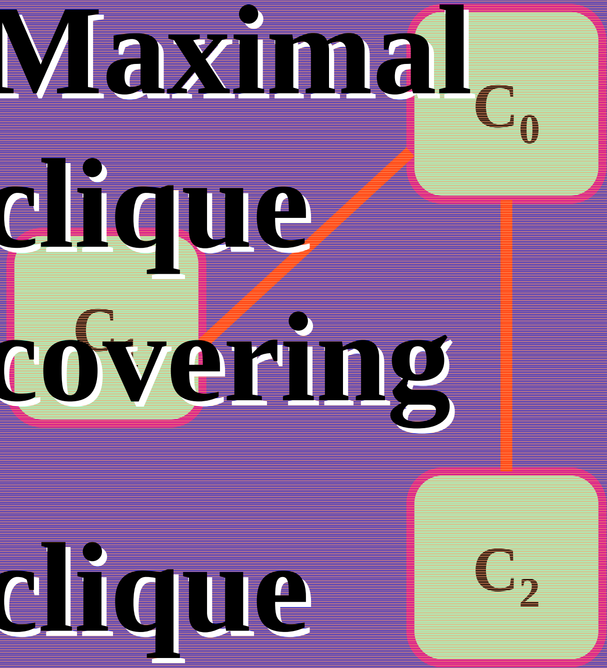


Column Incompatibility Graph

CCG and CIG are complementary

Maximal clique covering

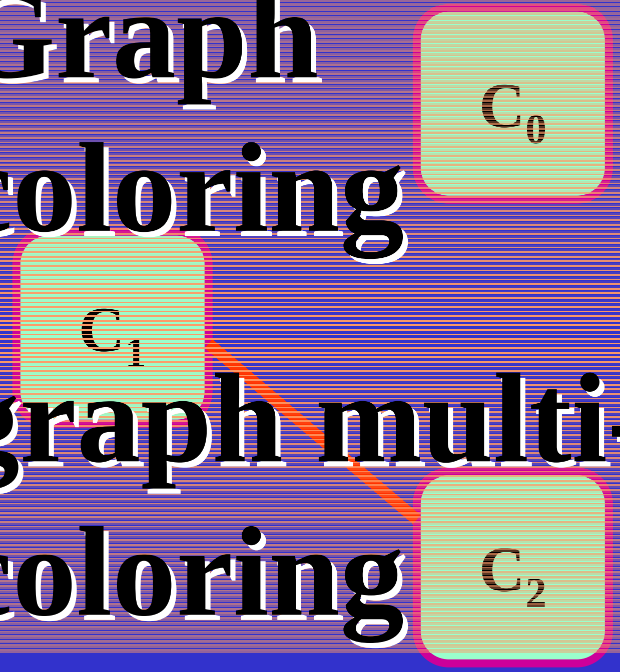
clique partitioning



Column Compatibility Graph

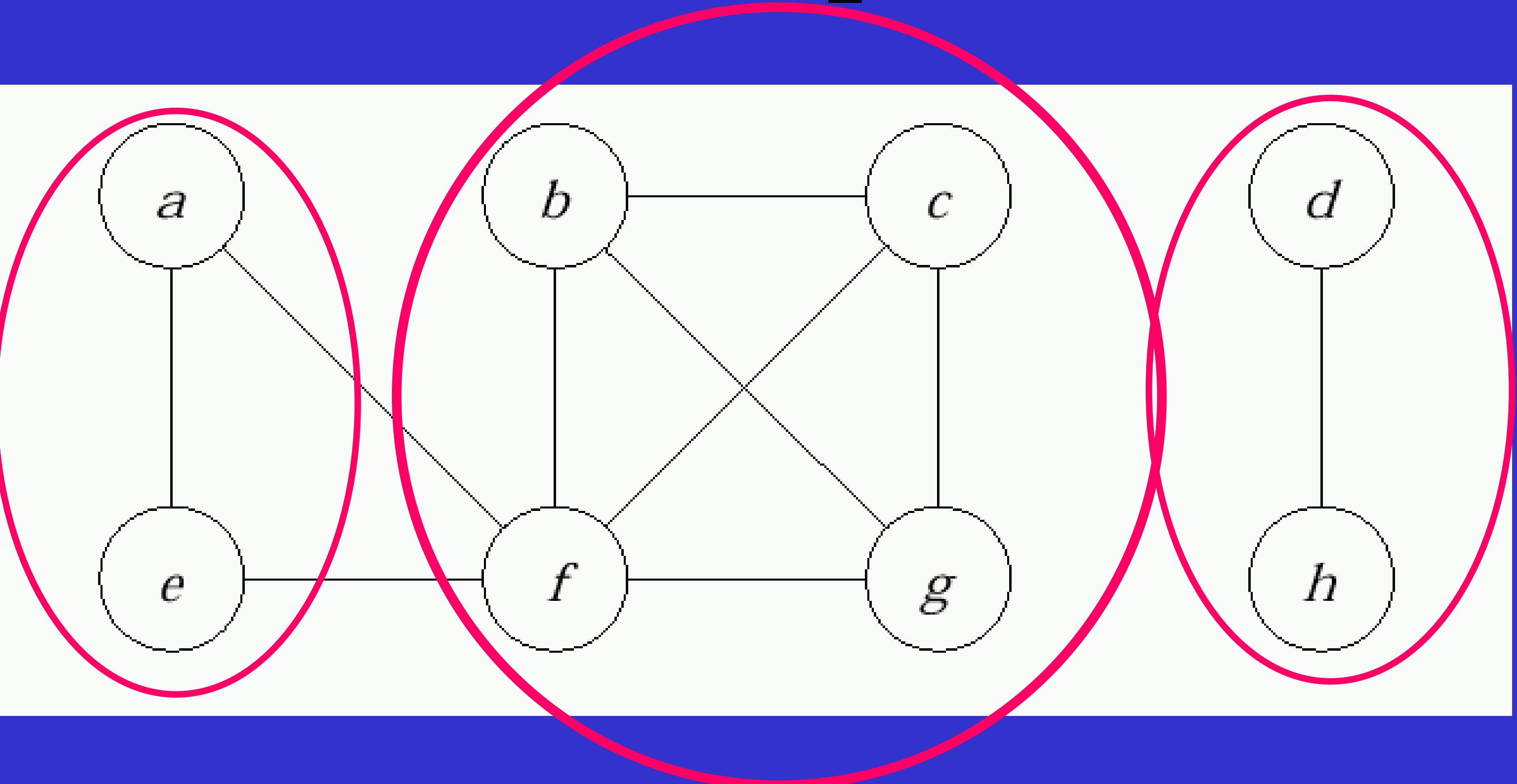
Graph coloring

graph multi-coloring

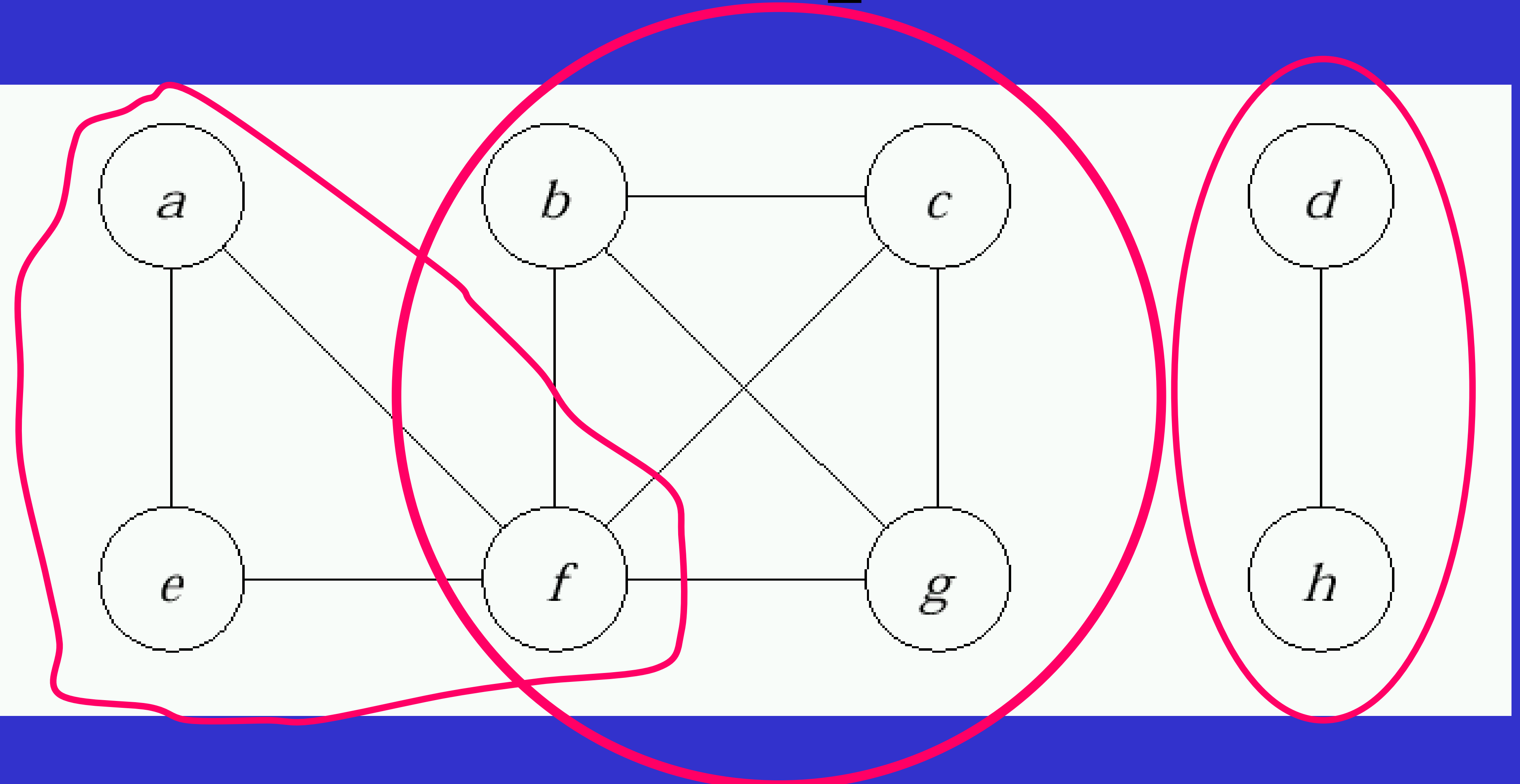


Column Incompatibility Graph

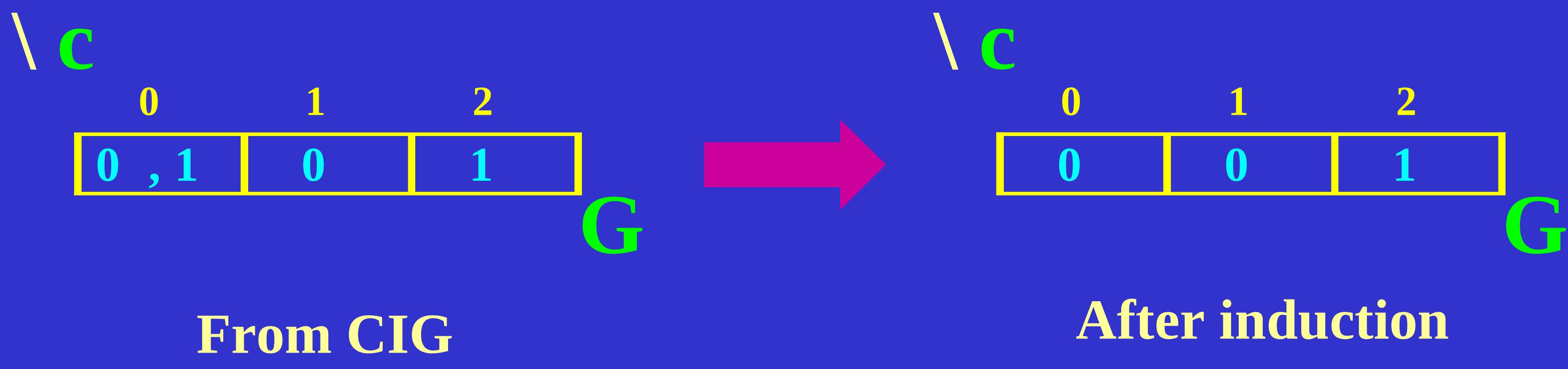
clique partitioning example.



Maximal clique covering example.



Map of relation G



g = a high pass filter whose acceptance threshold begins at

c > 1

Cost Function

Decomposed Function Cardinality **is the total cost of all blocks.**

Cost is defined for a single block in terms of the block's n inputs and m outputs

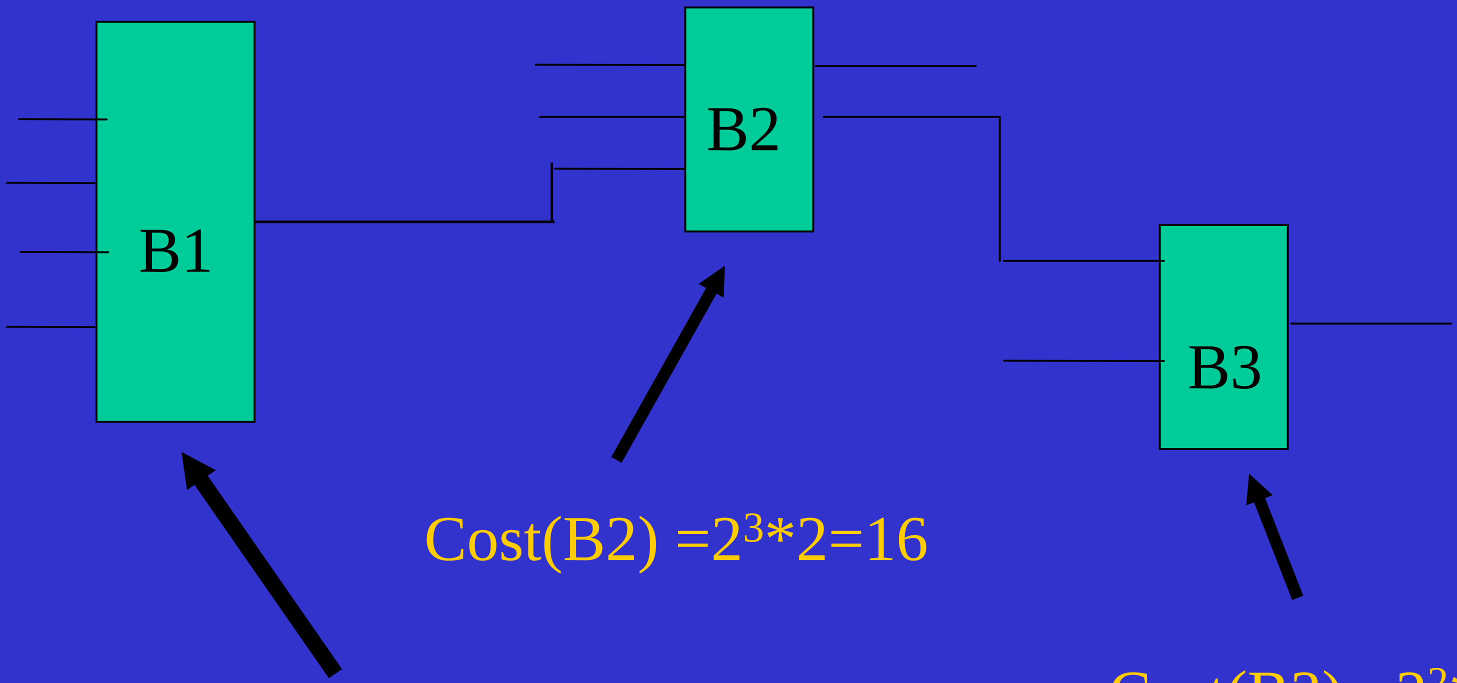
$$\textbf{Cost} := m * 2^n$$

DFC = Decomposed Function Cardinality

$$C_x(f) = \log_2 \min\{cost\ of\ \ \Gamma : \Gamma\ \ simulates\ f\}$$

$$cost(f) = 2^{|X|}|Y|$$

Example of DFC calculation



Cost(B2) $=2^3*2=16$

Cost(B3) $=2^2*1=4$

Cost(B1) $=2^4*1=16$

Total DFC = 16 + 16 + 4 = 36

Other cost functions

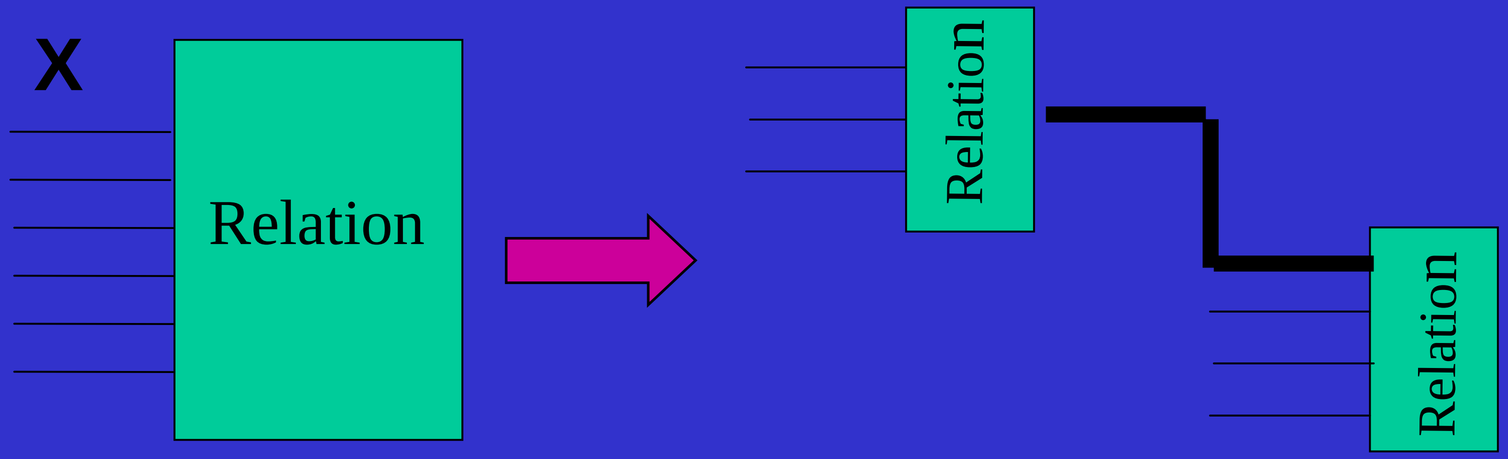
New Complexity Measures

$$C_x = \log_2 \left(\prod_{x_i \in X} |x_i| \log_2 \prod_{y_j \in Y} |y_j| \right)$$

where: $|x_i|$ is cardinality of variable $x_i \in X$,
$|y_j|$ is cardinality of variable $y_j \in Y$.

$$C_x = \log_2 \left(\prod_{y_j \in Y} |y_j| \right)^{\prod_{x_i \in X} |x_i|} = \prod_{x_i \in X} |x_i| \log_2 \prod_{y_j \in Y} |y_j|$$

Comparison of RC before and after decomposition



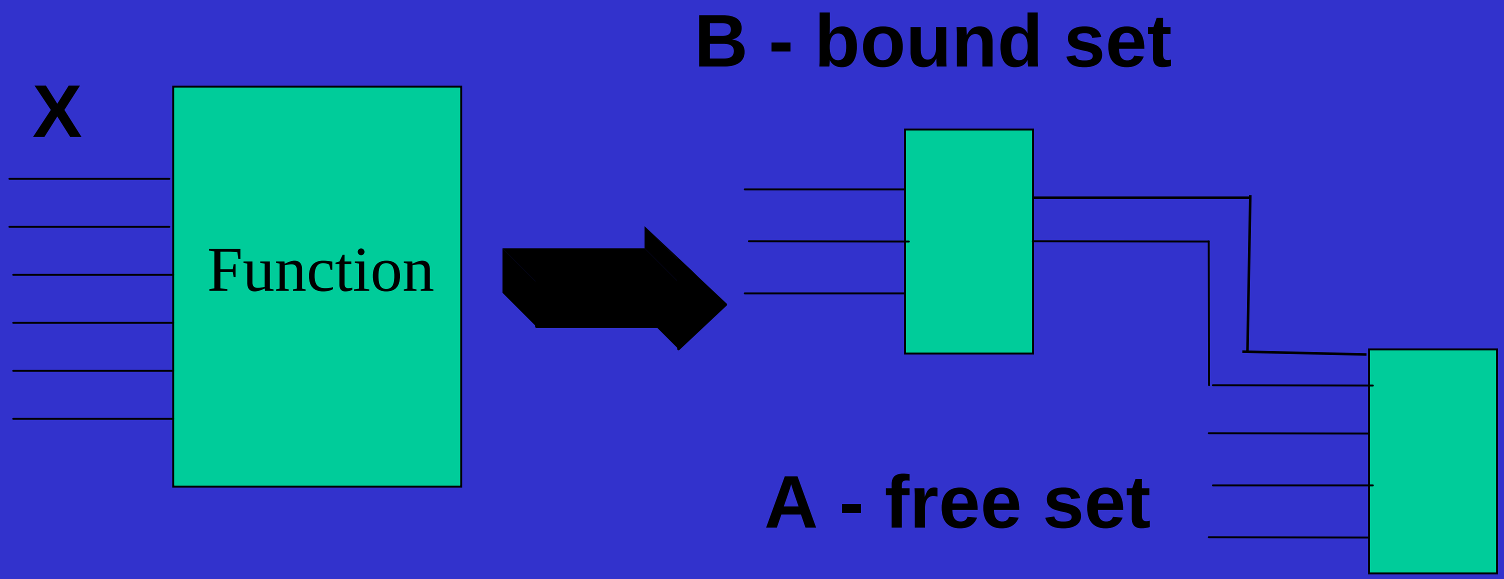
$RC_{before} = (3*3*3)*(\log_2 4) = 54$

$RC_{after} = [(3)*(\log_2 2)] +$

$[(2*3*3)*(\log_2 4)] = 3 + 36 = 39$

Two-Level Curtis Decomposition

F(X) = H(G(B), A), X = A ∪ B



if A ∩ B = ∅, it is *disjoint decomposition*

if A ∩ B ≠ ∅, it is *non-disjoint decomposition*

Decomposition Algorithm

- Find a set of partitions (A_i, B_i) of input variables (X) into free variables (A) and bound variables (B)
- For each partitioning, find decomposition $F(X) = H_i(G_i(B_i), A_i)$ such that column multiplicity is minimal, and calculate DFC
- Repeat the process for all partitioning until the decomposition with minimum DFC is found.

Algorithm Requirements

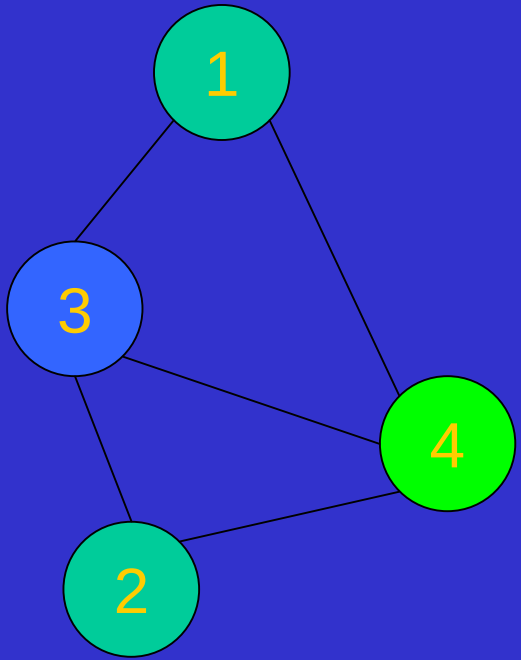
- Since the process is iterative, it is of high importance that minimization of the column multiplicity index is done as fast as possible.
- At the same time, for a given partitioning, it is important that the value of the column multiplicity is as close to the **absolute minimum** value

Column Multiplicity

Bound Set

Free Set	00	01	11	10
00	0	0	–	1
01	–	1	0	0
11	1	–	1	0
10	1	1	0	0

1 2 3 4

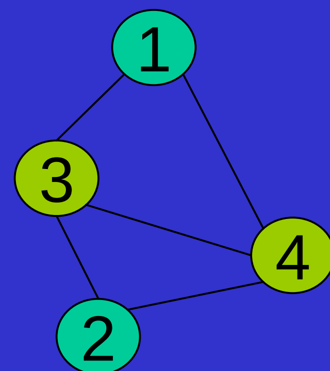


Column Multiplicity-other example

Bound Set

AB \ CD	00	01	11	10
00	0	0	–	1
01	–	1	0	0
11	1	–	1	-
10	1	1	0	0
	1	2	3	4

Free Set



C \ D	0	1
0	0	0
1	1	1

X

X=G(C,D)

X=C in this case

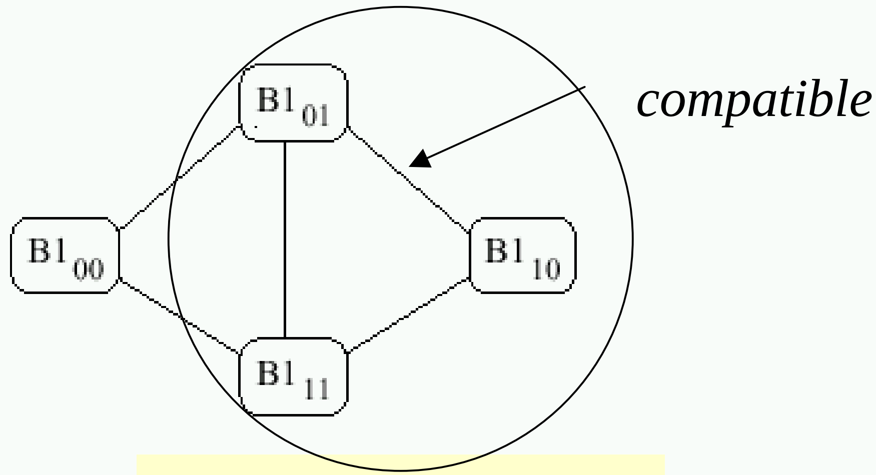
But how to calculate function H?

Decomposition of multiple-valued relation

ab \ cd	00 ($B1_{00}$)	01 ($B1_{01}$)	11 ($B1_{11}$)	10 ($B1_{10}$)
00	0,3 [0]	0,3 [4]	1,3 [7]	2,3 [9]
01	1,2 [1]	-	0,1 [8]	1,3 [10]
11	0 [2]	0,3 [5]	-	-
10	0,3 [3]	0,4 [6]	-	1,4 [11]

f

Karnaugh Map



Compatibility Graph for columns

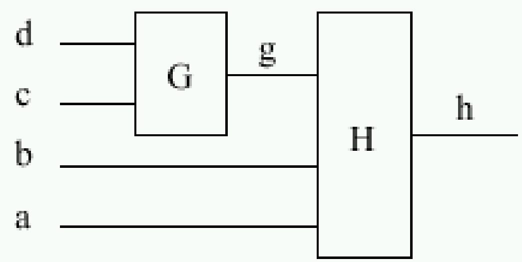
c \ d	0	1
0	0	0,1
1	1	0,1

Kmap of block G

ab \ g	0	1
00	3	3
01	1	1
11	0	0,3
10	0	4

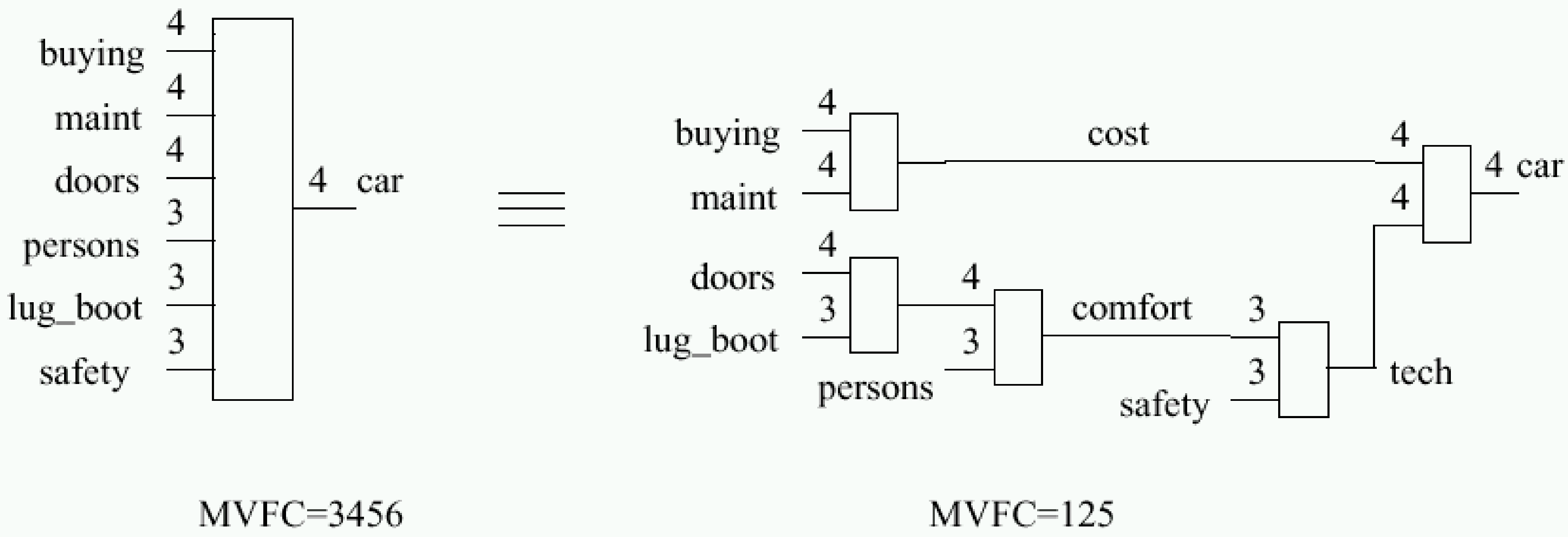
h

Kmap of block H



One level of decomposition

Discovering new concepts



- Discovering concepts useful for **purchasing a car**

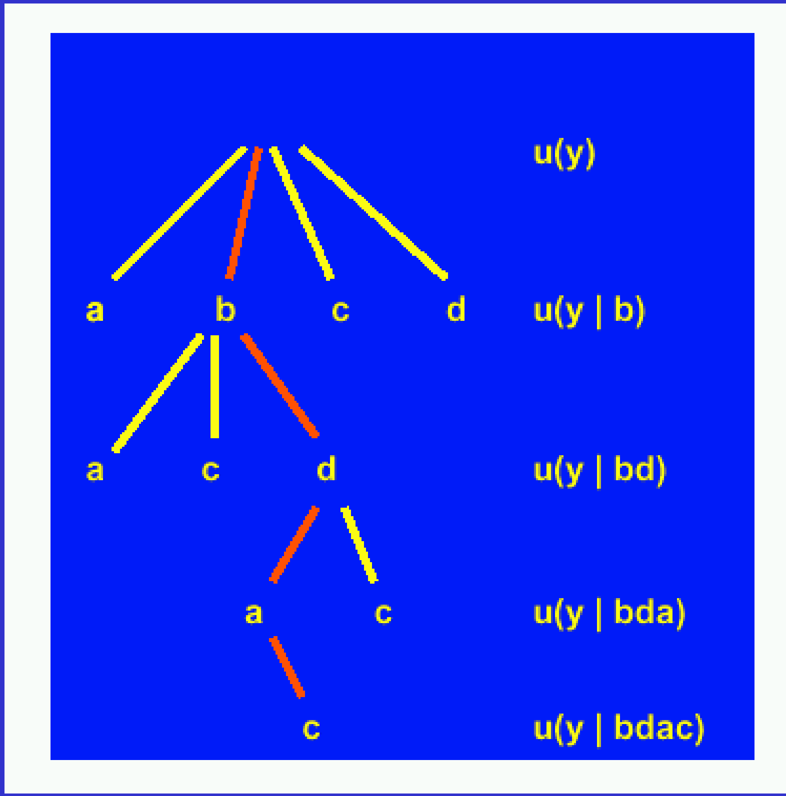
Variable ordering

- **Uncertainty (Shannon):**

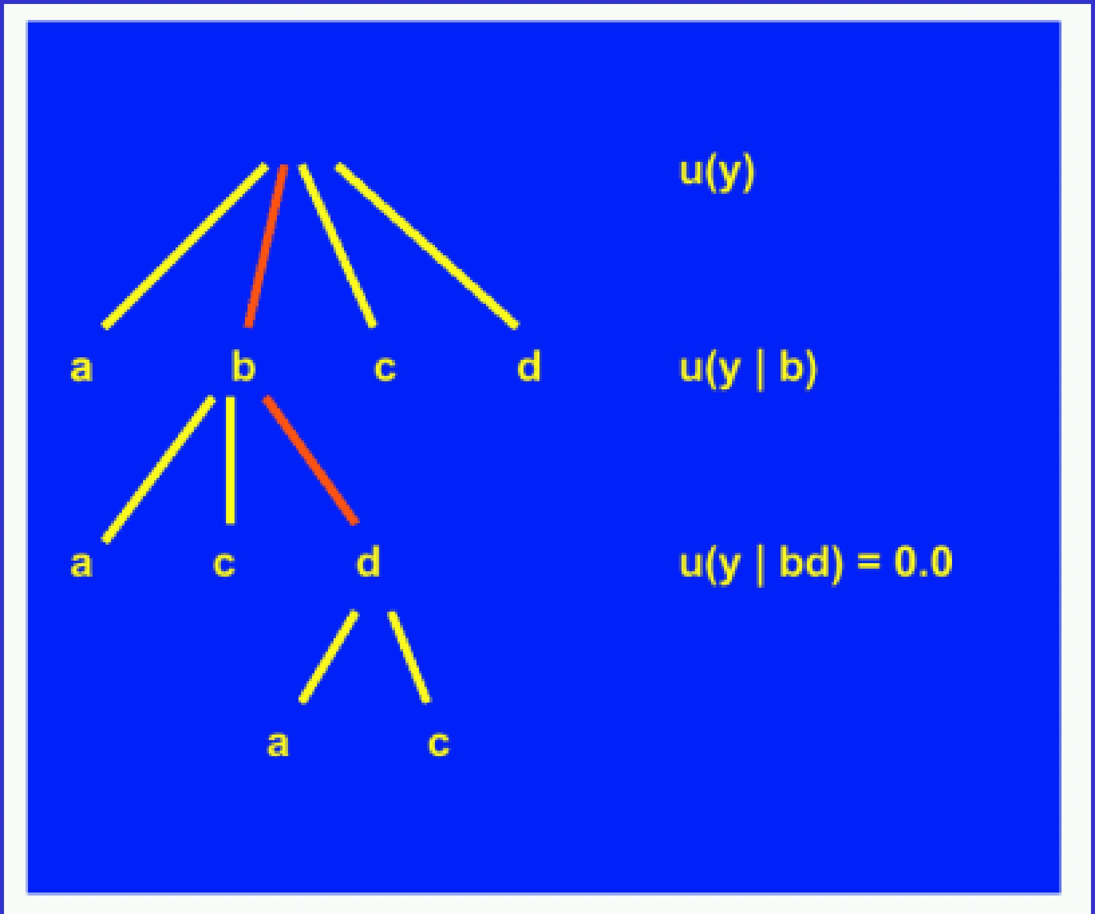
$$u(a) = -\sum_i p(a = a_i)\log_2 p(a = a_i)$$

- **Conditional Uncertainty (Shannon):**

$$u(a|b) = u(ab) - u(b)$$

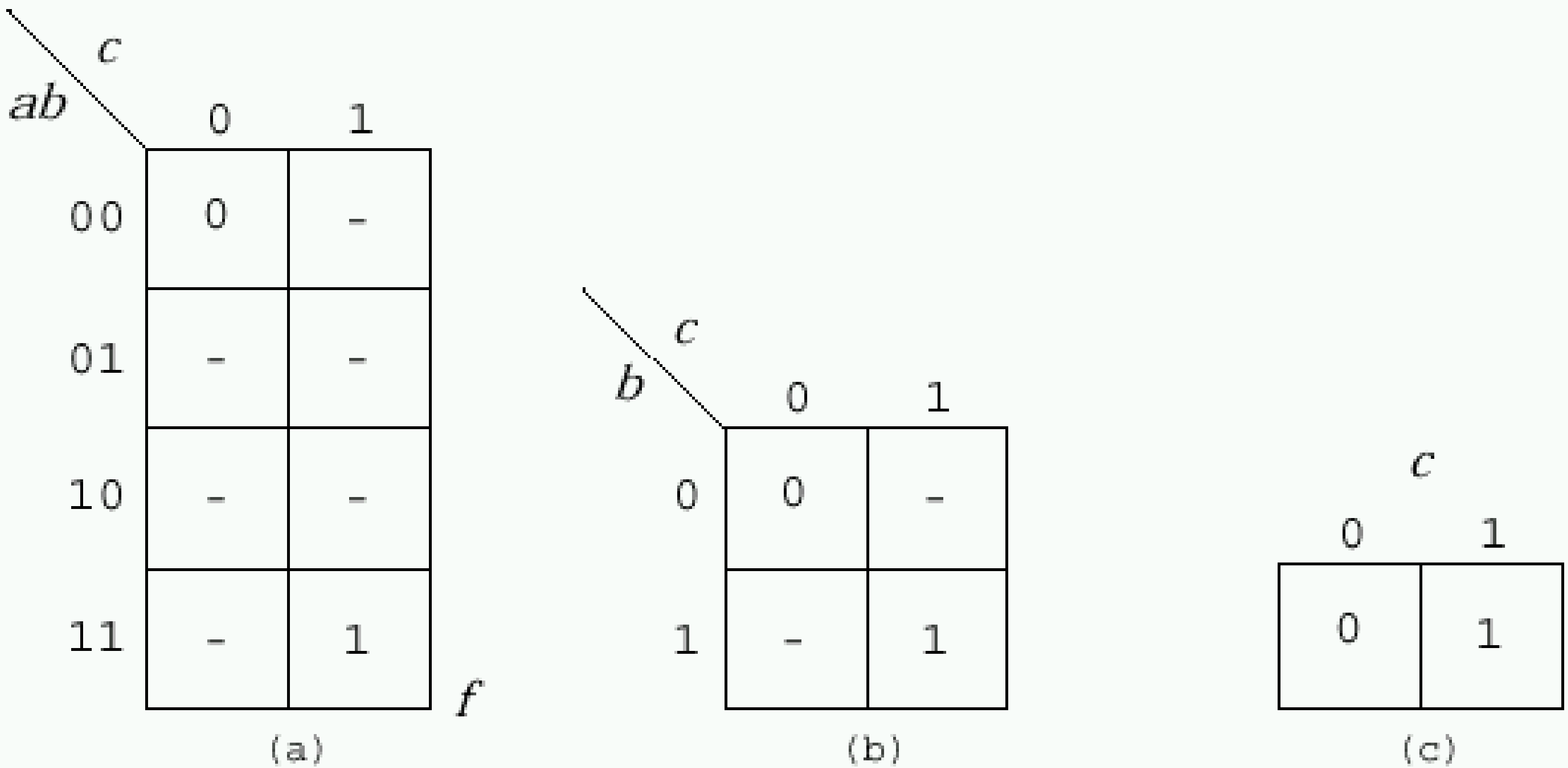


Vacuous variables removing



- Variables b and d reduce uncertainty of y to 0 which means they provide all the information necessary for determination of the output **y**
- Variables a and c are vacuous

Example of removing inessential variables (a) original function (b) variable a removed (c) variable b removed, variable c is no longer inessential.



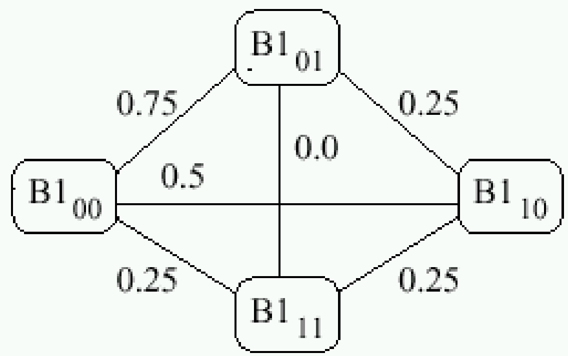
Generalization of the Ashenhurst-Curtis decomposition model

Compatibility graph construction for data with noise

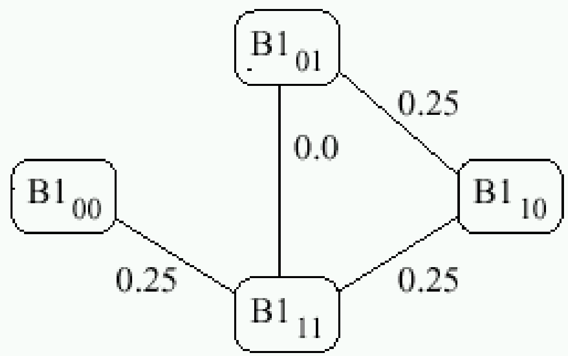
ab \ cd	$B1_{00}$ 00	$B1_{01}$ 01	$B1_{11}$ 11	$B1_{10}$ 10
00	0 [0]	3 [4]	1,3 [7]	2 [9]
01	1 [1]	-	0,1 [8]	1 [10]
11	0 [2]	3 [5]	-	-
10	0 [3]	4 [6]	-	4 [11]

f

Kmap



Compatibility Graph for Threshold 0.75



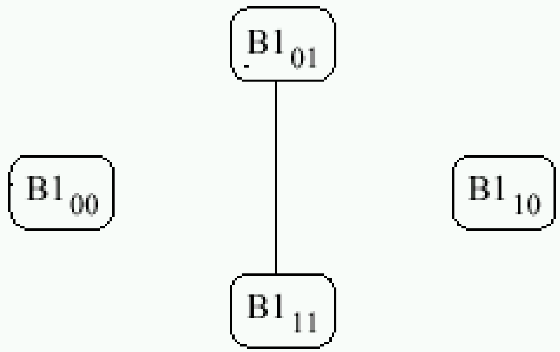
Compatibility Graph for Threshold 0.25

Compatibility graph for metric data

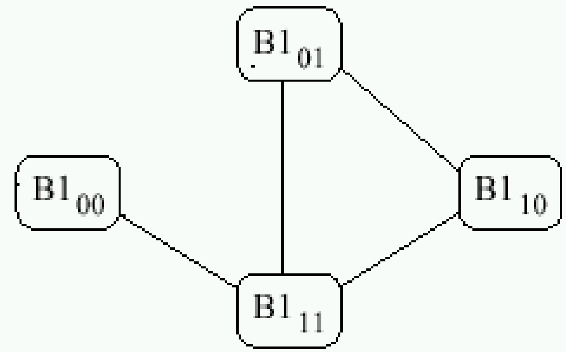
ab \ cd	00 ($B1_{00}$)	01 ($B1_{01}$)	11 ($B1_{11}$)	10 ($B1_{10}$)
00	0 [0]	3 [4]	1,3 [7]	2 [9]
01	1 [1]	-	0,1 [8]	1 [10]
11	0 [2]	3 [5]	-	-
10	0 [3]	4 [6]	-	4 [11]

f

Kmap



Compatibility Graph for nominal data



Compatibility Graph for metric data

Difference of 1

MV relations can be created from contingency tables

ab \ cd	00	01	11	10
00	77	57	3	2
01	1	110	12	1
11	12	28	200	1
10	0	423	21	52

a)

ab \ cd	00	01	11	10
00	1	0	0	0
01	0	1	0	0
11	0	0	1	0
10	0	1	0	0

ab \ cd	00	01	11	10
00	1	1	0	0
01	0	1	0	0
11	0	0	1	0
10	0	1	0	1

THRESHOLD 70

THRESHOLD 50

cd

a \ b	0	1
0	00	01
1	01	11

d)

cd

a \ b	0	1
0	00,01	01
1	01,10	11

e)

Figure 1: Contingency tables

Example of decomposing a Curtis non-decomposable function.