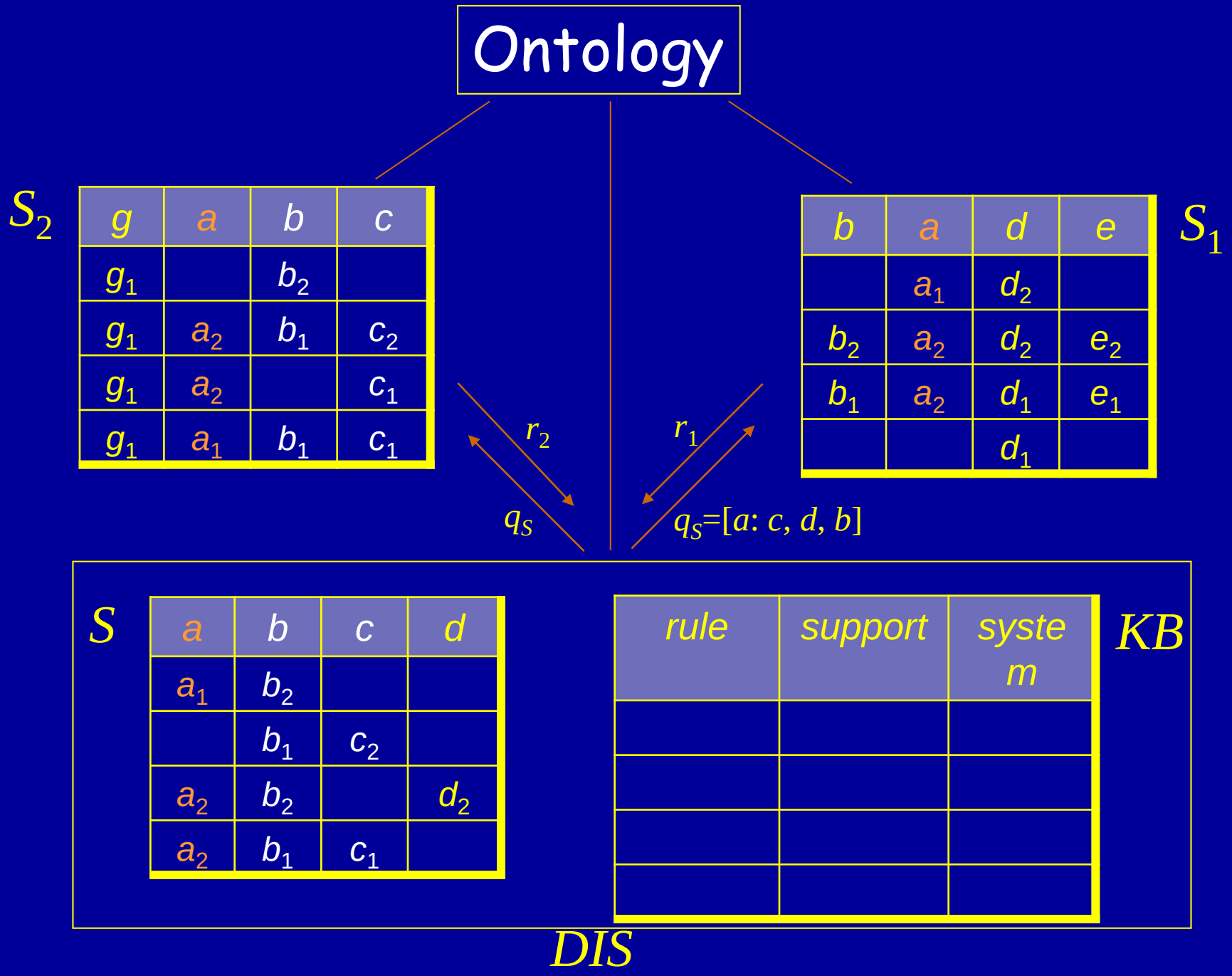
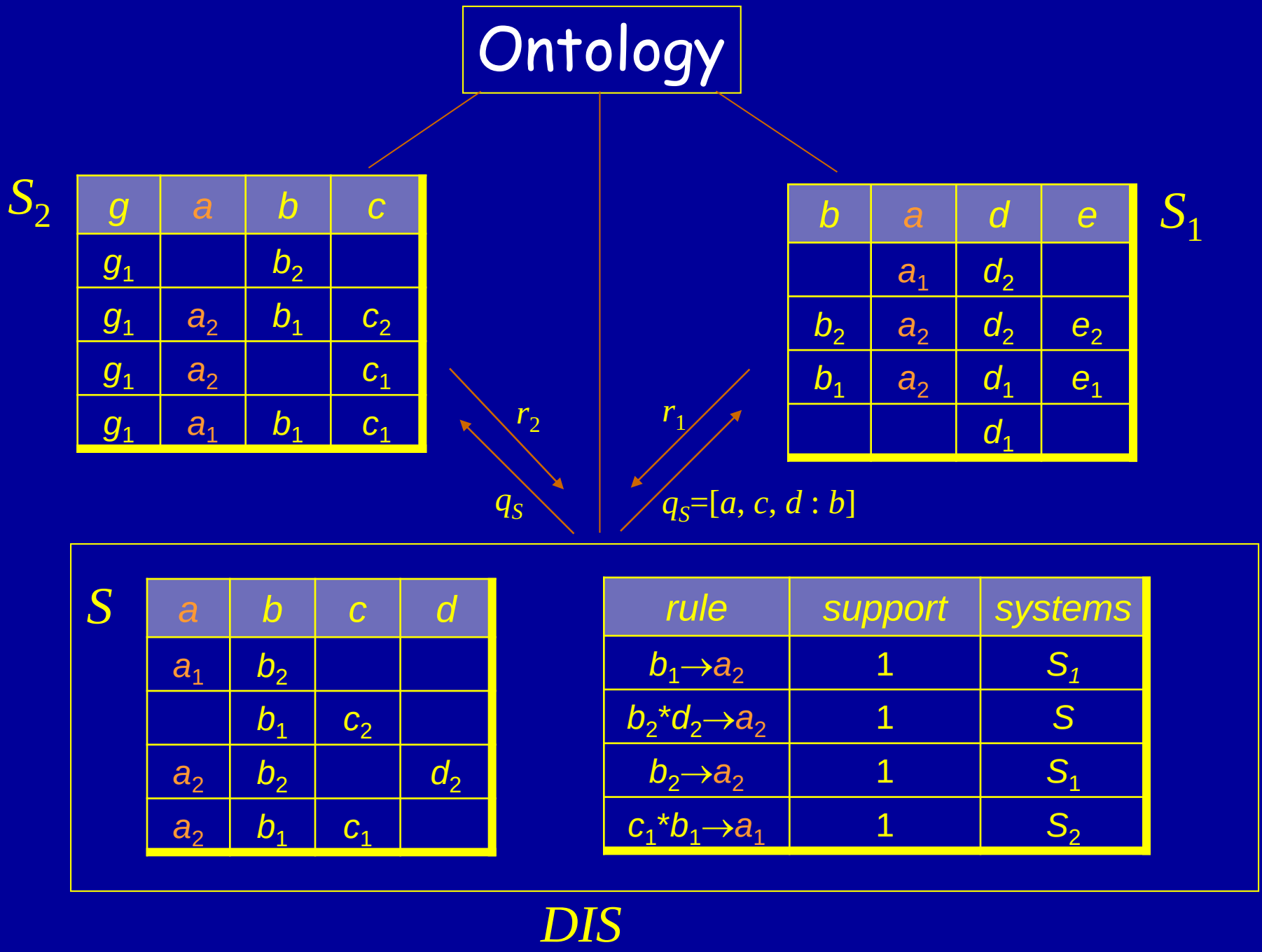
Data Security against Knowledge Loss *)

by

Zbigniew W. Ras

University of North Carolina, Charlotte, USA


Ontology
S_2
g a b c
g_1 b_2
g_1 a_2 b_1 c_2
g_1 a_2 c_1
g_1 a_1 b_1 c_1
S_1
b a d e
a_1 d_2
b_2 a_2 d_2 e_2
b_1 a_2 d_1 e_1
d_1
r_2
q_S
r_1
q_S=[a: c, d, b]
S
a b c d
a_1 b_2
b_1 c_2
a_2 b_2 d_2
a_2 b_1 c_1
rule support syste m
KB
DIS


Ontology
S_2
g a b c
g_1 b_2
g_1 a_2 b_1 c_2
g_1 a_2 c_1
g_1 a_1 b_1 c_1
b a d e
a_1 d_2
b_2 a_2 d_2 e_2
b_1 a_2 d_1 e_1
d_1
S_1
r_2
q_S
r_1
q_S=[a, c, d : b]
S
a b c d
a_1 b_2
b_1 c_2
a_2 b_2 d_2
a_2 b_1 c_1
rule support systems
$b_1 \rightarrow a_2$ 1 S_1
$b_2 * d_2 \rightarrow a_2$ 1 S
$b_2 \rightarrow a_2$ 1 S_1
$c_1 * b_1 \rightarrow a_1$ 1 S_2
DIS

Data Security against Knowledge Discovery

Problem…

Give a strategy for identifying minimal number of cells which additionally have to be hidden at site S (part of DIS) in order to guarantee that hidden attribute in S cannot be reconstructed by Distributed Knowledge Discovery.

S

a	b	c	d
a_1	b_2	c_2	d_1
a_2	b_1	c_2	
a_2	b_2	c_1	d_2
a_2	b_1	c_1	d_2

rule	*confidence*	*system*

KB

Data Security against Knowledge Discovery

Problem…

Give a strategy for identifying the minimal number of cells in S which additionally have to be hidden in order to guarantee that attribute a cannot be reconstructed by Distributed Knowledge Discovery.

S

a	b	c	d
	b_2	c_2	d_1
	b_1	c_2	
	b_2	c_1	d_2
	b_1	c_1	d_2

rule	*confidence*	*system*

KB

Data Security against Knowledge Discovery

Problem…

Give a strategy for identifying the minimal number of cells in S which additionally have to be hidden in order to guarantee that attribute a cannot be reconstructed by Distributed Knowledge Discovery.

S

a	b	c	d
	b_2	c_2	d_1
	b_1	c_2	
	b_2	c_1	d_2
	b_1	c_1	d_2

KB

rule	*confidence*	*system*
$b_1 \rightarrow a_2$	3/4	S_2
$b_2 * d_1 \rightarrow a_1$	1	S_2
$b_2 * d_2 \rightarrow a_2$	1	S_1
$c_1 * b_1 \rightarrow a_1$	1	S_2

a	b	c	d
a_1	b_2	c_2	d_1
a_2	b_1	c_2	
a_2	b_2	c_1	d_2
a_2	b_1	c_1	d_2

rule	confidence	system

Original site S

a	b	c	d
a_1	b_2	c_2	d_1
a_2	b_1	c_2	
a_2	b_2	c_1	d_2
a_1	b_1	c_1	d_2

rule	confidence	system
$b_1 \rightarrow a_2$	3/4	S_2
$b_2 * d_1 \rightarrow a_1$	1	S_2
$b_2 * d_2 \rightarrow a_2$	1	S_1
$c_1 * b_1 \rightarrow a_1$	1	S_2

Reconstructed site S

Disclosure Risk of Confidential Data

Problem

Give a strategy that identifies minimum number of attribute values that need to be additionally hidden in Information System S to guarantee that a hidden attribute cannot be reconstructed by Local & Distributed Chase

Chain of predictions by global and local rules

Object x6

x6	a2	b2	c3	sal=$50,000

Confidential data sal=$500 is hidden

x6	a2	b2	c3	

If we have a rule r1 = a2•b2 $\rightarrow$ sal=$50,000, then we hide b2 additionally

X6	a2		c3	

Now if we have a rule r2 = c3 $\rightarrow$ b2, confidential data is restored

x6	a2	b2	c3	sal=$50,000

Algorithm SCIKD (bottom-up strategy)

Rule	A	B	C	D	E	F	G
r_1	a_1	b_1	c_1				
r_2	a_1		c_1			f_1	
r_3		b_1	c_1				
r_4		b_1			e_1		
r_5	a_1		c_1			f_1	
r_6	a_1		c_1		e_1		
r_7			c_1		e_1		g_1
r_8	a_1		c_1	d_1			
r_9		b_1	c_1	d_1			
r_{10}				d_1		f_1	

KB - knowledge base

$r_1 = [\, b_1 * c_1 \rightarrow a_1 \,]$, $r_2 = [\, c_1 * f_1 \rightarrow a_1 \,]$,

D – decision attribute

Algorithm SCIKD (bottom-up strategy)

a1	b1	c1	d1	e1	f1	g1

$\{a_1\}^* = \{a_1\}$ unmarked
$\{b_1\}^* = \{b_1\}$ unmarked
$\{c_1\}^* = \{a_1,b_1,c_1,d_1,e_1\} \supseteq \{d_1\}$ marked
$\{e_1\}^* = \{b_1, e_1\}$ unmarked
$\{f_1\}^* = \{d_1,f_1\} \supseteq \{d_1\}$ marked
$\{g_1\}^* = \{g_1\}$ unmarked

$\{a_1, b_1\}^* = \{a_1, b_1\}$ unmarked
$\{a_1, e_1\}^* = \{a_1, b_1, e_1\}$ unmarked
$\{a_1, g_1\}^* = \{a_1, g_1\}$ unmarked
$\{b_1, e_1\}^* = \{b_1, e_1\}$ unmarked
$\{b_1, g_1\}^* = \{b_1, g_1\}$ unmarked
$\{e_1, g_1\}^* = \{a_1,b_1,c_1,d_1,e_1, g_1\} \supseteq \{d_1\}$ marked

---

$\{a_1, b_1, e_1\}^* = \{a_1, b_1, e_1\}$ unmarked /maximal subset/
$\{a_1, b_1, g_1\}^* = \{a_1, b_1, g_1\}$ unmarked /maximal subset/
$\{b_1, e_1, g_1\}^* = \{e_1, g_1\}^*$ marked $\{a1,e1,g1\}^* = \{e1,g1\}^*$ marked

---

$\{a1, b1, e1, g1\}^* = \{e1, g1\}^*$ marked

Data security versus knowledge loss

X	A	B	C	D	E	F	G
x_1	a_1	b_1	c_1	d_1	e_1	f_1	g_1
x_2							

Database
d_1 - has to be hidden

X	A	B	C	D	E	F	G
x_1	a_1	b_1					g_1
x_2							

X	A	B	C	D	E	F	G
x_1	a_1	b_1			e_1		
x_2							

{a1, b1, e1}* = {a1, b1, e1} unmarked /maximal subset/
{a1, b1, g1}* = {a1, b1, g1} unmarked /maximal subset/

Data security versus knowledge loss

X	A	B	C	D	E	F	G
x_1	a_1	b_1					g_1
x_2							

Database D_1

X	A	B	C	D	E	F	G
x_1	a_1	b_1			e_1		
x_2							

Database D_2

$R_k = \{r_{k,i}: i \in I\}$ set of rules extracted from D_k and $[c_{k,i}, s_{k,i}]$ denotes the confidence and support of rule r_i for all $i \in I_k$, k=1,2
With R_k we associate the number $K(R_k) = \Sigma\{ c_{k,i} \cdot s_{k,i} : i \in I_k\}$.

D_2 can be trusted more than D_1, if $K(R_1) > K(R_2)$.

Data security versus knowledge loss

X	A	B	C	D	E	F	G
x_1	a_1	b_1					g_1
x_2							

Database D_1

X	A	B	C	D	E	F	G
x_1	a_1	b_1			e_1		
x_2							

Database D_2

$R_k = \{r_{k,i}: i \in I\}$ set of rules extracted from D_k and $[c_{k,i}, s_{k,i}]$ denotes the coverage and support of rule r_i for all $i \in I_k$, k=1,2
With R_k we associate the number $K(R_k) = \Sigma\{ c_{k,i} \cdot s_{k,i} : i \in I_k\}$.

D_1 is preferable than D_2, if $K(R_1) > K(R_2)$.

Questions?

Thank You