Uncertain Databases - Data Representation

Databases and Information Systems

Fabian Panse

panse@informatik.uni-hamburg.de

University of Hamburg

Representation Systems

- Large number of possible worlds
- Many worlds overlap to a large extent

$\Rightarrow$ Impractical and unnecessary to store all worlds separately

$\Rightarrow$ Compact representation systems required

- Possible worlds representation as naive representation system and reference point
- Each representation system can be transformed into the possible worlds representation
- The possible worlds representation of a probabilistic database *pdb* is defined as $pwr(pdb) = (\mathbf{W}, Pr)$
- Mapping *pws* maps *pdb* to $\mathbf{W}$, i.e. $pws(pdb) = \mathbf{W}$

Representation Systems

Primary key:

- World key (short WK): unique for tuples of the same possible world (graphics: single underline)
- Representation key (short RK): unique for tuples of all possible worlds (graphics: double underline)

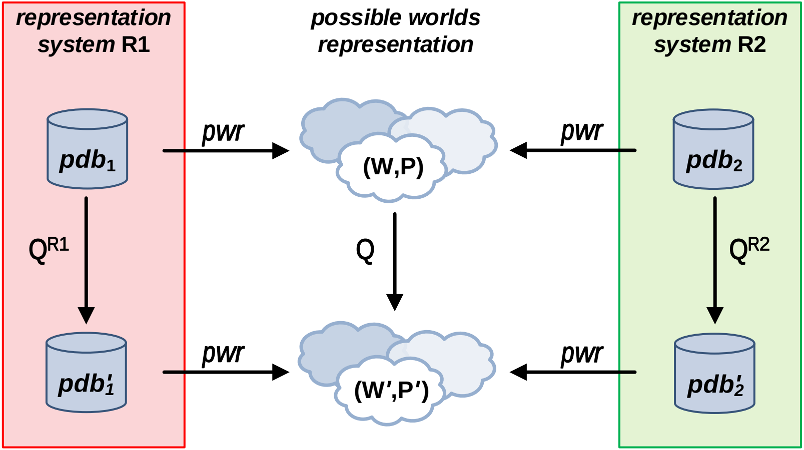
Semantic correctness:

- Representation system R has to be consistent with the possible worlds semantics

$\Rightarrow$ For each query Q, it exists a system-specific query Q^R that computes the compact result of Q, i.e.

$$pwr(Q^R(pdb)) = Q(pwr(pdb))$$

Representation Systems



Representation Systems

Goals:

- Compact representation (i.e. low storage requirements)
- Powerful representation system (i.e. should be able to represent as many possible worlds representations as possible)
- Low modeling/query complexity (i.e. easy to understand and efficient to query)

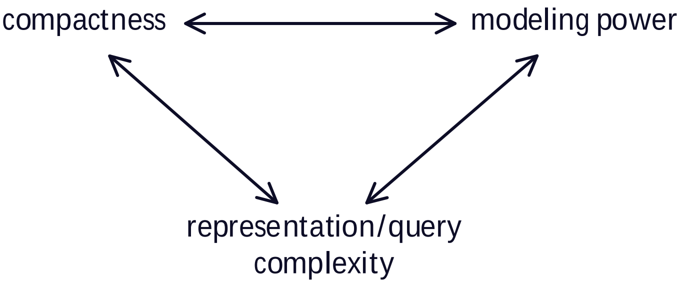
Problem:

- These goals are contradictory
- Increasing two of them comes always at the cost of decreasing the third

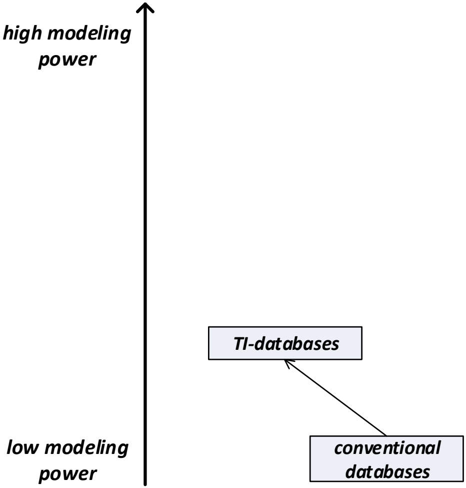
$\Rightarrow$ Choice of a system is always a trade-off between these goals

$\Rightarrow$ Choice of a system depends on use case

Representation Systems



Tuple-Independent Databases (TI-databases)



Tuple-Independent Databases (TI-databases)

- Each tuple is associated with its marginal probability
- Tuples as independent events (tuple-level uncertainty)

Person

	<u>*WK*</u>	*name*	*age*	*p*
t_1	*p1*	*J.Doe*	*27*	0.8
t_2	*p2*	*K.Smith*	*32*	1.0
t_3	*p3*	*J.Ho*	*28*	0.4

- Tuples are mutually independent

$\Rightarrow$ One possible world that contains all tuples

$\Rightarrow$ World key can be used as representation key

Tuple-Independent Databases (TI-databases)

Possible World Generation (formal):

- One possible world per combination of maybe-tuples
- Let pdb be a TI-database
- Let $pdb^!$ the set of all certain-tuples of pdb
- Let $pdb^?$ the set of all maybe-tuples of pdb
- Number of possible worlds: $|\mathbf{W}| = 2^{|pdb^?|}$

Possible world space:

$$\mathbf{W} = pws(pdb) = \{pdb^! \cup S \mid S \subseteq pdb^?\}$$

Probability of a possible world $W \in \mathbf{W}$:

$$Pr(W) = \prod_{t \in W} p(t) \times \prod_{t \in pdb^?, t \notin W} (1 - p(t))$$

Tuple-Independent Databases (TI-databases)

Possible World Generation (example):

Two maybe-tuples $\Rightarrow 2^2 = 4$ possible worlds:

Person

	WK	name	age	p
t_1	p1	J.Doe	27	0.8
t_2	p2	K.Smith	32	1.0
t_3	p3	J.Ho	28	0.4

Tuple-Independent Databases (TI-databases)

Possible World Generation (example):

Possible world $W_1 = \{t_2\}$:

Person

	<u>*WK*</u>	*name*	*age*	*p*
t_1	*p1*	*J.Doe*	*27*	0.8
t_2	*p2*	*K.Smith*	*32*	1.0
t_3	*p3*	*J.Ho*	*28*	0.4

$$\begin{aligned} Pr(W_1) &= (1 - p(t_1)) \times p(t_2) \times (1 - p(t_3)) \\ &= 0.2 \times 1.0 \times 0.6 = \mathbf{0.12} \end{aligned}$$

Tuple-Independent Databases (TI-databases)

Possible World Generation (example):

Possible world $W_2 = \{t_1, t_2\}$:

Person

	<u>*WK*</u>	*name*	*age*	*p*
t_1	*p1*	*J.Doe*	*27*	0.8
t_2	*p2*	*K.Smith*	*32*	1.0
t_3	*p3*	*J.Ho*	*28*	0.4

$$\begin{aligned} Pr(W_2) &= p(t_1) \times p(t_2) \times (1 - p(t_3)) \\ &= 0.8 \times 1.0 \times 0.6 = \mathbf{0.48} \end{aligned}$$

Tuple-Independent Databases (TI-databases)

Possible World Generation (example):

Possible world $W_3 = \{t_2, t_3\}$:

Person

	WK	name	age	p
t_1	p1	J.Doe	27	0.8
t_2	p2	K.Smith	32	1.0
t_3	p3	J.Ho	28	0.4

$$\begin{aligned} Pr(W_3) &= (1 - p(t_1)) \times p(t_2) \times p(t_3) \\ &= 0.2 \times 1.0 \times 0.4 = \mathbf{0.08} \end{aligned}$$

Tuple-Independent Databases (TI-databases)

Possible World Generation (example):

Possible world $W_4 = \{t_1, t_2, t_3\}$:

Person

	<u>WK</u>	name	age	p
t_1	p1	J.Doe	27	0.8
t_2	p2	K.Smith	32	1.0
t_3	p3	J.Ho	28	0.4

$$\begin{aligned} Pr(W_4) &= p(t_1) \times p(t_2) \times p(t_3) \\ &= 0.8 \times 1.0 \times 0.4 = \mathbf{0.32} \end{aligned}$$

Tuple-Independent Databases (TI-databases)

Possible World Generation (example): Overview

W_1, Pr=0.12

	WK	name	age
t_2	p2	K.Smith	32

W_2, Pr=0.48

	WK	name	age
t_1	p1	J.Doe	27
t_2	p2	K.Smith	32

W_3, Pr=0.08

	WK	name	age
t_2	p2	K.Smith	32
t_3	p3	J.Ho	28

W_4, Pr=0.32

	WK	name	age
t_1	p1	J.Doe	27
t_2	p2	K.Smith	32
t_3	p3	J.Ho	28

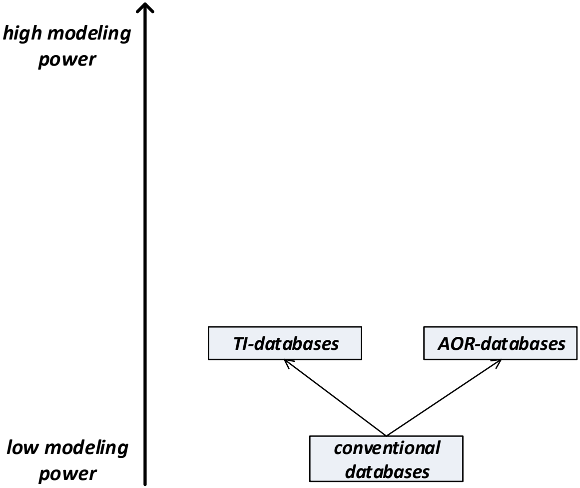
Tuple-Independent Databases (TI-databases)

Computation of the most probable world:

- Removal of all tuples with probability lower than 0.5
- If some tuples have probability 0.5

$\Rightarrow$ More than one most probable world exists

Attribute-OR Databases (AOR-databases)



Attribute-OR Databases (AOR-databases)

- Each tuple is a certain event
- Values in non-key attributes as independent random variables

$\Rightarrow$ Each tuple has several alternative values per attribute (attribute-level uncertainty)

- Tuples are sets of random variables (so-called *A-tuples*)

Person

	WK	name		age	
t_1	*p1*	*J.Doe*	:1.0	*27*	:0.8
				28	:0.2
t_2	*p2*	*K.Smith*	:1.0	32	:1.0
t_3	*p3*	*J.Doe*	:0.7	*28*	:0.5
		J.Ho	:0.3	*29*	:0.5

Attribute-OR Databases (AOR-databases)

- All A-tuples are certain

⇒ Each possible world contains all A-tuples

⇒ World key can be used as representation key

Person

	WK	name		age	
t_1	p1	J.Doe	:1.0	27	:0.8
				28	:0.2
t_2	p2	K.Smith	:1.0	32	:1.0
t_3	p3	J.Doe	:0.7	28	:0.5
		J.Ho	:0.3	29	:0.5

Attribute-OR Databases (AOR-databases)

- Each A-tuple models a set of possible instances
- One instance per combination of one alternative value per attribute
- Let $\{A_1, \ldots, A_k\}$ be the attributes of the considered table
- Let $Prob(t[A] = a)$ be the probability that A-tuple t has the alternative value a in attribute A
- Set of possible instances of an A-tuple t is defined as:

$$
\begin{aligned}
pws(t) = & \{a_1 \in dom(A_1) \mid Prob(t[A_1] = a_1) > 0\} \\
& \times \{a_2 \in dom(A_2) \mid Prob(t[A_2] = a_2) > 0\} \\
& \ldots \\
& \times \{a_k \in dom(A_k) \mid Prob(t[A_k] = a_k) > 0\} \\
= & \{(a_1, \ldots, a_k) \in dom(A_1) \times \ldots \times dom(A_k) \mid \textstyle\prod_{i=1}^{k} Prob(t[A_i] = a_i) > 0\}
\end{aligned}
$$

- Attribute values are mutually independent

$\Rightarrow p(t^{\langle * \rangle}) = \prod_{i=1}^{k} Prob(t[A_i] = t^{\langle * \rangle}[A_i])$ for every $t^{\langle * \rangle} \in pws(t)$.

Attribute-OR Databases (AOR-databases)

Possible Instance Generation (example): A-tuple t_3

2 attributes with each 2 alternative values $\Rightarrow 2^2 = 4$ instances:

Person

	WK	name		age	
t_1	p1	J.Doe	:1.0	27	:0.8
				28	:0.2
t_2	p2	K.Smith	:1.0	32	:1.0
t_3	p3	J.Doe	:0.7	28	:0.5
		J.Ho	:0.3	29	:0.5

Attribute-OR Databases (AOR-databases)

Possible Instance Generation (example): A-tuple t_3

Possible instance $t_3^{\langle 1 \rangle}$=('p3','J.Doe','28'):

Person

	WK	name		age	
t_1	p1	J.Doe	:1.0	27	:0.8
				28	:0.2
t_2	p2	K.Smith	:1.0	32	:1.0
t_3	p3	J.Doe	:0.7	28	:0.5
		J.Ho	:0.3	29	:0.5

$$p(t_3^{\langle 1 \rangle}) = Prob(t_3[\text{'}name\text{'}] = \text{'J.Doe'}) \times Prob(t_3[\text{'}age\text{'}] = \text{'28'})$$

$$= 0.7 \times 0.5 = \mathbf{0.35}$$

Attribute-OR Databases (AOR-databases)

Possible Instance Generation (example): A-tuple t_3

Possible instance $t_3^{\langle 2 \rangle}$=('p3','J.Doe','29'):

Person

	<u>*WK*</u>	*name*		*age*	
t_1	*p1*	*J.Doe*	:1.0	*27*	:0.8
				28	:0.2
t_2	*p2*	*K.Smith*	:1.0	32	:1.0
t_3	*p3*	*J.Doe*	:0.7	*28*	:0.5
		J.Ho	:0.3	*29*	:0.5

$$p(t_3^{\langle 2 \rangle}) = Prob(t_3[\text{'}name\text{'}] = \text{'J.Doe'}) \times Prob(t_3[\text{'}age\text{'}] = \text{'29'})$$
$$= 0.7 \times 0.5 = \mathbf{0.35}$$

Attribute-OR Databases (AOR-databases)

Possible Instance Generation (example): A-tuple t_3

Possible instance $t_3^{\langle 3 \rangle}$=('p3','J.Ho','28'):

Person

	WK	name		age	
t_1	p1	J.Doe	:1.0	27	:0.8
				28	:0.2
t_2	p2	K.Smith	:1.0	32	:1.0
t_3	p3	J.Doe	:0.7	28	:0.5
		J.Ho	:0.3	29	:0.5

$$p(t_3^{\langle 3 \rangle}) = Prob(t_3[\text{'}name\text{'}] = \text{'J.Ho'}) \times Prob(t_3[\text{'}age\text{'}] = \text{'28'})$$
$$= 0.3 \times 0.5 = \mathbf{0.15}$$

Attribute-OR Databases (AOR-databases)

Possible Instance Generation (example): A-tuple t_3

Possible instance $t_3^{\langle 4 \rangle}$=('p3','J.Ho','29'):

Person

	WK	name		age	
t_1	p1	J.Doe	:1.0	27	:0.8
				28	:0.2
t_2	p2	K.Smith	:1.0	32	:1.0
t_3	p3	J.Doe	:0.7	28	:0.5
		J.Ho	:0.3	29	:0.5

$$p(t_3^{\langle 3 \rangle}) = Prob(t_3[\text{'name'}] = \text{'J.Ho'}) \times Prob(t_3[\text{'age'}] = \text{'29'})$$
$$= 0.3 \times 0.5 = \mathbf{0.15}$$

Minimal Hitting Set

- Let $C = \{S_1, \ldots, S_k\}$ be a collection of sets
- Set H is a hitting set for C if
  - it contains only elements that belong to sets in C, i.e.

    $H \subseteq \bigcup_{i=1}^{k} S_i$
  - it contains at least one element per set in C, i.e.

    $H \cap S_i \neq \emptyset$ for every $i \in \{1, \ldots, k\}$
- Set H is a minimal hitting set for C if
  - no strict subset of H is a hitting set for C

  $\Rightarrow$ H contains exactly one element per set in C, i.e.

    $|H \cap S_i| = 1$ for every $i \in \{1, \ldots, k\}$
- $\mathfrak{H}(C)$ is the set of all minimal hitting sets of C

Minimal Hitting Set - Example

Let $C = \{S_1, S_2, S_3\}$ be a collection of sets with

- $S_1 = \{a, b, c\}$
- $S_2 = \{k, l, m, n\}$
- $S_3 = \{x, y, z\}$

Which of the following sets are (minimal) hitting sets of C?

- $H_1 = \{a, b, k, x\}$ non-minimal hitting set
- $H_2 = \{a, k, q, z\}$ no hitting set
- $H_3 = \{b, l, y\}$ minimal hitting set
- $H_4 = \{m, z\}$ no hitting set

How many minimal hitting sets of C exist? $3 \times 4 \times 3 = 36$

Attribute-OR Databases (AOR-databases)

Possible World Generation (formal):

- Possible world is constructed by selecting for each A-tuple one alternative value per attribute

$\Rightarrow$ One possible world per combination of possible instances (one instance per A-tuple)

- Let *pdb* be an AOR-database
- Number of possible worlds: $|\mathbf{W}| = \prod_{t \in pdb} |pws(t)|$

Possible world space:

$$\mathbf{W} = \mathfrak{H}(C) \text{ where } C = \{pws(t) \mid t \in pdb\}$$

Probability of a possible world $W \in \mathbf{W}$:

$$Pr(W) = \prod_{t^{\langle * \rangle} \in W} p(t^{\langle * \rangle})$$

Attribute-OR Databases (AOR-databases)

Given: A-tuples t_1, t_2 and t_3 with

$$pws(t_1) = \{t_1^{\langle 1 \rangle}, t_1^{\langle 2 \rangle}, t_1^{\langle 3 \rangle}\},\ pws(t_2) = \{t_2^{\langle 1 \rangle}, t_2^{\langle 2 \rangle}\},\ pws(t_3) = \{t_3^{\langle 1 \rangle}, t_3^{\langle 2 \rangle}\}$$

Corresponding Minimal Hitting Sets:

Minimal Hitting Sets	
H_1	$\{t_1^{\langle 1 \rangle}, t_2^{\langle 1 \rangle}, t_3^{\langle 1 \rangle}\}$
H_2	$\{t_1^{\langle 1 \rangle}, t_2^{\langle 1 \rangle}, t_3^{\langle 2 \rangle}\}$
H_3	$\{t_1^{\langle 1 \rangle}, t_2^{\langle 2 \rangle}, t_3^{\langle 1 \rangle}\}$
H_4	$\{t_1^{\langle 1 \rangle}, t_2^{\langle 2 \rangle}, t_3^{\langle 2 \rangle}\}$

Minimal Hitting Sets	
H_5	$\{t_1^{\langle 2 \rangle}, t_2^{\langle 1 \rangle}, t_3^{\langle 1 \rangle}\}$
H_6	$\{t_1^{\langle 2 \rangle}, t_2^{\langle 1 \rangle}, t_3^{\langle 2 \rangle}\}$
H_7	$\{t_1^{\langle 2 \rangle}, t_2^{\langle 2 \rangle}, t_3^{\langle 1 \rangle}\}$
H_8	$\{t_1^{\langle 2 \rangle}, t_2^{\langle 2 \rangle}, t_3^{\langle 2 \rangle}\}$

Minimal Hitting Sets	
H_9	$\{t_1^{\langle 3 \rangle}, t_2^{\langle 1 \rangle}, t_3^{\langle 1 \rangle}\}$
H_{10}	$\{t_1^{\langle 3 \rangle}, t_2^{\langle 1 \rangle}, t_3^{\langle 2 \rangle}\}$
H_{11}	$\{t_1^{\langle 3 \rangle}, t_2^{\langle 2 \rangle}, t_3^{\langle 1 \rangle}\}$
H_{12}	$\{t_1^{\langle 3 \rangle}, t_2^{\langle 2 \rangle}, t_3^{\langle 2 \rangle}\}$

Attribute-OR Databases (AOR-databases)

Possible World Generation (example):

Person

	WK	name		age	
t_1	p1	J.Doe	:1.0	27	:0.8
				28	:0.2
t_2	p2	K.Smith	:1.0	32	:1.0
t_3	p3	J.Doe	:0.7	28	:0.5
		J.Ho	:0.3	29	:0.5

A-tuple t_1: 2 possible instances
A-tuple t_2: 1 possible instances $\Rightarrow$ $2 \times 1 \times 4 = 8$ possible worlds
A-tuple t_3: 4 possible instances

Attribute-OR Databases (AOR-databases)

Possible World Generation (example): Overview

W_1, Pr=0.28

	WK	name	age
t_1	p1	J.Doe	27
t_2	p2	K.Smith	32
t_3	p3	J.Doe	28

W_2, Pr=0.28

	WK	name	age
t_1	p1	J.Doe	27
t_2	p2	K.Smith	32
t_3	p3	J.Doe	29

W_3, Pr=0.12

	WK	name	age
t_1	p1	J.Doe	27
t_2	p2	K.Smith	32
t_3	p3	J.Ho	28

W_4, Pr=0.12

	WK	name	age
t_1	p1	J.Doe	27
t_2	p2	K.Smith	32
t_3	p3	J.Ho	29

W_5, Pr=0.07

	WK	name	age
t_1	p1	J.Doe	28
t_2	p2	K.Smith	32
t_3	p3	J.Doe	28

W_6, Pr=0.07

	WK	name	age
t_1	p1	J.Doe	28
t_2	p2	K.Smith	32
t_3	p3	J.Doe	29

W_7, Pr=0.03

	WK	name	age
t_1	p1	J.Doe	28
t_2	p2	K.Smith	32
t_3	p3	J.Ho	28

W_8, Pr=0.03

	WK	name	age
t_1	p1	J.Doe	28
t_2	p2	K.Smith	32
t_3	p3	J.Ho	29

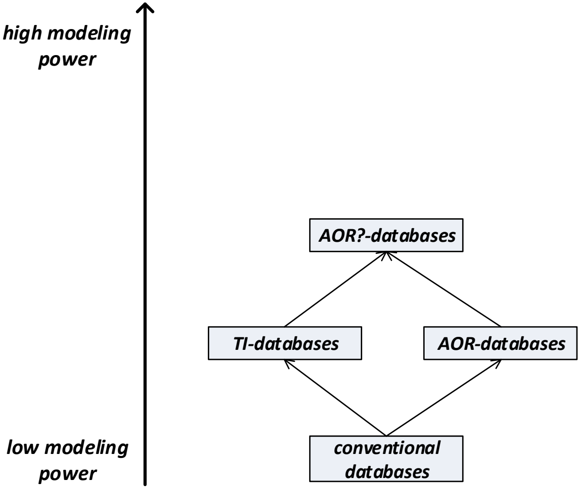
Attribute-OR Databases (AOR-databases)

Computation of the most probable world:

- Select the most probable value per attribute for each A-tuple
- If an A-tuple has more than one most probable value per attribute

$\Rightarrow$ More than one most probable world exists

Attribute-OR DBs with Maybe-Tuples (AOR?-databases)



Attribute-OR DBs with Maybe-Tuples (AOR?-databases)

- Combines the ideas of AOR-databases and TI-databases

⇒ Values in non-key attributes as independent random variables

⇒ A-tuples as independent events

Person

	<u>*WK*</u>	*name*		*age*		*p*
t_1	*p1*	*J.Doe*	:1.0	*27*	:0.8	0.8
				28	:0.2	
t_2	*p2*	*K.Smith*	:1.0	32	:1.0	1.0
t_3	*p3*	*J.Doe*	:0.7	*28*	:0.5	1.0
		J.Ho	:0.3	*29*	:0.5	

Attribute-OR DBs with Maybe-Tuples (AOR?-databases)

- All A-tuples are mutually independent

⇒ One possible world contains all A-tuples

⇒ World key can be used as representation key

Person

	WK	name		age		p
t_1	p1	J.Doe	:1.0	27 28	:0.8 :0.2	0.8
t_2	p2	K.Smith	:1.0	32	:1.0	1.0
t_3	p3	J.Doe J.Ho	:0.7 :0.3	28 29	:0.5 :0.5	1.0

Attribute-OR DBs with Maybe-Tuples (AOR?-databases)

Possible World Generation:

- Select one possible instance per certain A-tuple
- Select one or none possible instance per maybe A-tuple
- Formal Definition: Similar to BID-databases (see next section)

Attribute-OR DBs with Maybe-Tuples (AOR?-databases)

Possible World Generation (example):

Person

	WK	name		age		p
t_1	p1	J.Doe	:1.0	27	:0.8	0.8
				28	:0.2	
t_2	p2	K.Smith	:1.0	32	:1.0	1.0
t_3	p3	J.Doe	:0.7	28	:0.5	1.0
		J.Ho	:0.3	29	:0.5	

A-tuple t_1 (maybe): 2 poss. instances
A-tuple t_2 (certain): 1 poss. instances $\Rightarrow$ $(2+1) \times 1 \times 4 = 12$ poss. worlds
A-tuple t_3 (certain): 4 poss. instances

Attribute-OR DBs with Maybe-Tuples (AOR?-databases)

Possible World Generation (example): Overview

W_1, Pr=0.224

	WK	name	age
t_1	p1	J.Doe	27
t_2	p2	K.Smith	32
t_3	p3	J.Doe	28

W_2, Pr=0.224

	WK	name	age
t_1	p1	J.Doe	27
t_2	p2	K.Smith	32
t_3	p3	J.Doe	29

W_3, Pr=0.096

	WK	name	age
t_1	p1	J.Doe	27
t_2	p2	K.Smith	32
t_3	p3	J.Ho	28

W_4, Pr=0.096

	WK	name	age
t_1	p1	J.Doe	27
t_2	p2	K.Smith	32
t_3	p3	J.Ho	29

W_5, Pr=0.056

	WK	name	age
t_1	p1	J.Doe	28
t_2	p2	K.Smith	32
t_3	p3	J.Doe	28

W_6, Pr=0.056

	WK	name	age
t_1	p1	J.Doe	28
t_2	p2	K.Smith	32
t_3	p3	J.Doe	29

W_7, Pr=0.024

	WK	name	age
t_1	p1	J.Doe	28
t_2	p2	K.Smith	32
t_3	p3	J.Ho	28

W_8, Pr=0.024

	WK	name	age
t_1	p1	J.Doe	28
t_2	p2	K.Smith	32
t_3	p3	J.Ho	29

W_9, Pr=0.07

	WK	name	age
t_2	p2	K.Smith	32
t_3	p3	J.Doe	28

W_{10}, Pr=0.07

	WK	name	age
t_2	p2	K.Smith	32
t_3	p3	J.Doe	29

W_{11}, Pr=0.03

	WK	name	age
t_2	p2	K.Smith	32
t_3	p3	J.Ho	28

W_{12}, Pr=0.03

	WK	name	age
t_2	p2	K.Smith	32
t_3	p3	J.Ho	29

Attribute-OR DBs with Maybe-Tuples (AOR?-databases)

Transformation from TI-database to AOR?-database:

- One A-tuple per tuple
- One alternative value per attribute

Transformation from AOR-database to AOR?-database:

- Associating every A-tuple with probability 1.0

Attribute-OR DBs with Maybe-Tuples (AOR?-databases)

Computation of the most probable world:

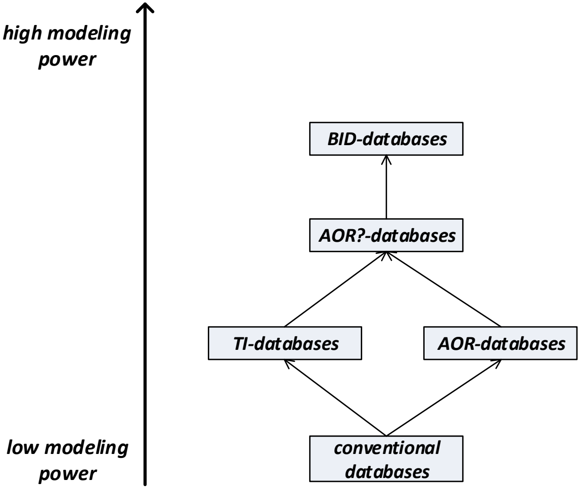
- Select the most probable instance per certain A-tuple

$\Rightarrow$ select the most probable value per attribute for each certain A-tuple

- Select the most probable state (most probable instance or no instance) per maybe A-tuple
- If an A-tuple has more than one most probable instance/state

$\Rightarrow$ More than one most probable world exists

Block-Independent-Disjoint Databases (BID-databases)



Block-Independent-Disjoint Databases (BID-databases)

- Each tuple is associated with its marginal probability
- Tuples are grouped in blocks
  - Tuples of different blocks are mutually independent
  - Tuples of the same block are mutually exclusive
- Block is *maybe* if probabilities of its tuples do not sum up to 1

Person

	RK	BNo.	WK	name	age	p
t_1	1	1	p1	J.Doe	27	0.6
t_2	2	1	p1	J.Doe	28	0.2
t_3	3	2	p2	K.Smith	32	1.0
t_4	4	3	p3	J.Doe	28	0.8
t_5	5	3	p3	J.Ho	29	0.2

Block-Independent-Disjoint Databases (BID-databases)

- Each tuple is associated with its marginal probability
- Tuples are grouped in blocks
  - Tuples of different blocks are mutually independent
  - Tuples of the same block are mutually exclusive
- Block is *maybe* if probabilities of its tuples do not sum up to 1

Person

	<u>RK</u>	BNo.	<u>WK</u>	name	age	p	
t_1	1	1	p1	J.Doe	27	0.6	← maybe-tuple
t_2	2	1	p1	J.Doe	28	0.2	← maybe-tuple
t_3	3	2	p2	K.Smith	32	1.0	← certain-tuple
t_4	4	3	p3	J.Doe	28	0.8	← maybe-tuple
t_5	5	3	p3	J.Ho	29	0.2	← maybe-tuple

Block-Independent-Disjoint Databases (BID-databases)

- Each tuple is associated with its marginal probability
- Tuples are grouped in blocks
  - Tuples of different blocks are mutually independent
  - Tuples of the same block are mutually exclusive
- Block is *maybe* if probabilities of its tuples do not sum up to 1

Person

	<u>RK</u>	BNo.	<u>WK</u>	name	age	p
t_1	1	1	p1	J.Doe	27	0.6
t_2	2	1	p1	J.Doe	28	0.2
t_3	3	2	p2	K.Smith	32	1.0
t_4	4	3	p3	J.Doe	28	0.8
t_5	5	3	p3	J.Ho	29	0.2

maybe block

certain block

certain block

Block-Independent-Disjoint Databases (BID-databases)

Person

	<u>*RK*</u>	*BNo.*	<u>*WK*</u>	*name*	*age*	*p*
t_1	1	1	p1	J.Doe	27	0.6
t_2	2	1	p1	J.Doe	28	0.2
t_3	3	2	p2	K.Smith	32	1.0
t_4	4	3	p3	J.Doe	28	0.8
t_5	5	3	p3	J.Ho	29	0.2

- Tuples can be exclusive

$\Rightarrow$ Different tuples can share the same world key value

$\Rightarrow$ World key cannot be used as representation key

Block-Independent-Disjoint Databases (BID-databases)

Possible World Generation (formal):

- Let *pdb* be a BID-database
- Let $\mathcal{B}^!$ the set of all certain blocks of *pdb*
- Let $\mathcal{B}^?$ the set of all maybe blocks of *pdb*
- Number of possible worlds: $|\mathbf{W}| = \prod_{B \in \mathcal{B}^!} |B| \times \prod_{B \in \mathcal{B}^?} (|B| + 1)$

Possible world space:

$$\mathbf{W} = \bigcup_{C \in \{\mathcal{B}^! \cup S | S \subseteq \mathcal{B}^?\}} \mathfrak{H}(C)$$

Probability of a possible world $W \in \mathbf{W}$:

$$Pr(W) = \prod_{t \in W} p(t) \times \prod_{B \in \mathcal{B}^?,\, B \cap W = \emptyset} (1 - p(B))$$

where $p(B) = \sum_{t \in B} p(t)$.

Block-Independent-Disjoint Databases (BID-databases)

Given:

Certain blocks $\mathcal{B}^! = \{B_1, B_2\}$ with $B_1 = \{t_1\}$ and $B_2 = \{t_2, t_3\}$

Maybe blocks $\mathcal{B}^? = \{B_3, B_4\}$ with $B_3 = \{t_4, t_5\}$ and $B_4 = \{t_6\}$

Corresponding Collections and Minimal Hitting Sets:

Collection C_i	**Minimal Hitting Sets $\mathfrak{H}(C_i)$**
$C_1 = \{B_1, B_2\}$	$H_{11} = \{t_1, t_2\}$, $H_{12} = \{t_1, t_3\}$
$C_2 = \{B_1, B_2, B_3\}$	$H_{21} = \{t_1, t_2, t_4\}$, $H_{22} = \{t_1, t_3, t_4\}$ $H_{23} = \{t_1, t_2, t_5\}$, $H_{24} = \{t_1, t_3, t_5\}$
$C_3 = \{B_1, B_2, B_4\}$	$H_{31} = \{t_1, t_2, t_6\}$, $H_{32} = \{t_1, t_3, t_6\}$
$C_4 = \{B_1, B_2, B_3, B_4\}$	$H_{41} = \{t_1, t_2, t_4, t_6\}$, $H_{42} = \{t_1, t_3, t_4, t_6\}$ $H_{43} = \{t_1, t_2, t_5, t_6\}$, $H_{44} = \{t_1, t_3, t_5, t_6\}$

Block-Independent-Disjoint Databases (BID-databases)

Possible World Generation (example):

Two certain blocks (1 & 2 tupels), one maybe block (2 tupels)
$\Rightarrow 1 \times 2 \times 3 = 6$ possible worlds:

Person

	<u>RK</u>	BNo.	<u>WK</u>	name	age	p
t_1	1	1	p1	J.Doe	27	0.6
t_2	2	1	p1	J.Doe	28	0.2
t_3	3	2	p2	K.Smith	32	1.0
t_4	4	3	p3	J.Doe	28	0.8
t_5	5	3	p3	J.Ho	29	0.2

Block-Independent-Disjoint Databases (BID-databases)

Possible World Generation (example):

Possible world $W_1 = \{t_1, t_3, t_4\}$:

Person

	<u>RK</u>	BNo.	<u>WK</u>	name	age	p
t_1	1	1	p1	J.Doe	27	0.6
t_2	2	1	p1	J.Doe	28	0.2
t_3	3	2	p2	K.Smith	32	1.0
t_4	4	3	p3	J.Doe	28	0.8
t_5	5	3	p3	J.Ho	29	0.2

$$\begin{aligned} Pr(W_1) &= p(t_1) \times p(t_3) \times p(t_4) \\ &= 0.6 \times 1.0 \times 0.8 = \mathbf{0.48} \end{aligned}$$

Block-Independent-Disjoint Databases (BID-databases)

Possible World Generation (example):

Possible world $W_2 = \{t_1, t_3, t_5\}$:

Person

	<u>RK</u>	BNo.	<u>WK</u>	name	age	p
t_1	1	1	p1	J.Doe	27	0.6
t_2	2	1	p1	J.Doe	28	0.2
t_3	3	2	p2	K.Smith	32	1.0
t_4	4	3	p3	J.Doe	28	0.8
t_5	5	3	p3	J.Ho	29	0.2

$$
\begin{aligned}
Pr(W_2) &= p(t_1) \times p(t_3) \times p(t_5) \\
&= 0.6 \times 1.0 \times 0.2 = \mathbf{0.12}
\end{aligned}
$$

Block-Independent-Disjoint Databases (BID-databases)

Possible World Generation (example):

Possible world $W_3 = \{t_2, t_3, t_4\}$:

Person

	<u>RK</u>	BNo.	<u>WK</u>	name	age	p
t_1	1	1	p1	J.Doe	27	0.6
t_2	2	1	p1	J.Doe	28	0.2
t_3	3	2	p2	K.Smith	32	1.0
t_4	4	3	p3	J.Doe	28	0.8
t_5	5	3	p3	J.Ho	29	0.2

$$Pr(W_3) = p(t_2) \times p(t_3) \times p(t_4)$$
$$= 0.2 \times 1.0 \times 0.8 = \mathbf{0.16}$$

Block-Independent-Disjoint Databases (BID-databases)

Possible World Generation (example):

Possible world $W_4 = \{t_2, t_3, t_5\}$:

Person

	<u>RK</u>	BNo.	<u>WK</u>	name	age	p
t_1	1	1	p1	J.Doe	27	0.6
t_2	2	1	p1	J.Doe	28	0.2
t_3	3	2	p2	K.Smith	32	1.0
t_4	4	3	p3	J.Doe	28	0.8
t_5	5	3	p3	J.Ho	29	0.2

$$
\begin{aligned}
Pr(W_4) &= p(t_2) \times p(t_3) \times p(t_5) \\
&= 0.2 \times 1.0 \times 0.2 = \mathbf{0.04}
\end{aligned}
$$

Block-Independent-Disjoint Databases (BID-databases)

Possible World Generation (example):

Possible world $W_5 = \{t_3, t_4\}$:

Person

	RK	BNo.	WK	name	age	p
t_1	1	1	p1	J.Doe	27	0.6
t_2	2	1	p1	J.Doe	28	0.2
t_3	3	2	p2	K.Smith	32	1.0
t_4	4	3	p3	J.Doe	28	0.8
t_5	5	3	p3	J.Ho	29	0.2

$$
\begin{aligned}
Pr(W_5) &= (1 - p(B_1)) \times p(t_3) \times p(t_4) \\
&= 0.2 \times 1.0 \times 0.8 = \mathbf{0.16}
\end{aligned}
$$

Block-Independent-Disjoint Databases (BID-databases)

Possible World Generation (example):

Possible world $W_6 = \{t_3, t_5\}$:

Person

	<u>RK</u>	BNo.	<u>WK</u>	name	age	p
t_1	1	1	p1	J.Doe	27	0.6
t_2	2	1	p1	J.Doe	28	0.2
t_3	3	2	p2	K.Smith	32	1.0
t_4	4	3	p3	J.Doe	28	0.8
t_5	5	3	p3	J.Ho	29	0.2

$$\begin{aligned} Pr(W_6) &= (1 - p(B_1)) \times p(t_3) \times p(t_5) \\ &= 0.2 \times 1.0 \times 0.2 = \mathbf{0.04} \end{aligned}$$

Block-Independent-Disjoint Databases (BID-databases)

Possible World Generation (example): Overview

W_1, Pr=0.48

	WK	name	age
t_1	p1	J.Doe	27
t_3	p2	K.Smith	32
t_4	p3	J.Doe	28

W_3, Pr=0.16

	WK	name	age
t_2	p1	J.Doe	28
t_3	p2	K.Smith	32
t_4	p3	J.Doe	28

W_5, Pr=0.16

	WK	name	age
t_3	p2	K.Smith	32
t_4	p3	J.Doe	28

W_2, Pr=0.12

	WK	name	age
t_1	p1	J.Doe	27
t_3	p2	K.Smith	32
t_5	p3	J.Ho	29

W_4, Pr=0.04

	WK	name	age
t_2	p1	J.Doe	28
t_3	p2	K.Smith	32
t_5	p3	J.Ho	29

W_6, Pr=0.04

	WK	name	age
t_3	p2	K.Smith	32
t_5	p3	J.Ho	29

Block-Independent-Disjoint Databases (BID-databases)

Transformation from AOR?-database to BID-database:

- One block per A-tuple
- One tuple per possible instance

$\Rightarrow$ Plain presentation of possible instances

$\Rightarrow$ Loss in Compactness

Example:

- A-tuple with 3 alternative values in each of 5 attributes

$\Rightarrow$ $3^5 = 243$ possible instances

$\Rightarrow$ $243 \times 5 = 1215$ attribute values instead of $5 \times 3 = 15$

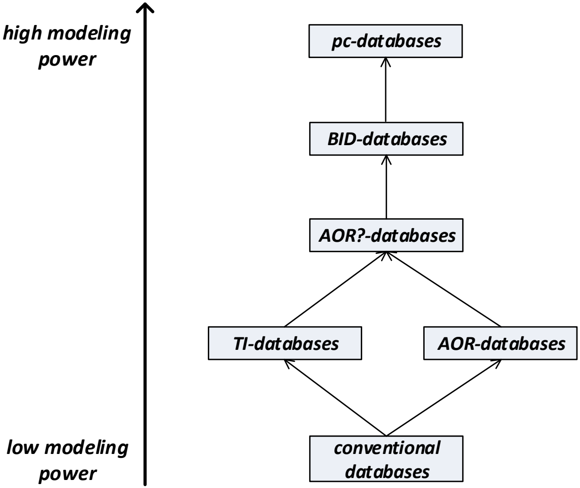
Block-Independent-Disjoint Databases (BID-databases)

Computation of the most probable world:

- Select the most probable tuple per certain block
- Select the most probable state (most probable tuple or no tuple) per maybe block
- If a block has more than one most probable tuple/state

$\Rightarrow$ More than one most probable world exists

Probabilistic Conditional Databases (pc-databases)



Probabilistic Conditional Databases (pc-databases)

- Finite set of mutually independent random variables
- Each random variable has a finite number of possible values
- Each tuple is associated with a condition over these variables (tuple-level uncertainty)

Person

	RK	WK	name	age	condition
t_1	1	p1	J.Doe	27	X=1
t_2	2	p1	J.Doe	28	X=2
t_3	3	p2	K.Smith	32	Y=1
t_4	4	p2	S.Kmith	32	Y=2
t_5	5	p3	J.Doe	28	X=1 v X=3
t_6	6	p3	J.Ho	29	X=2

World-Table

var	value	Prob
X	1	0.6
X	2	0.2
X	3	0.2
Y	1	0.8
Y	2	0.2

Probabilistic Conditional Databases (pc-databases)

- The same variable can appear in conditions of different tuples

⇒ Variables can be used to introduce tuple correlations

Person

	RK	WK	name	age	condition
t_1	1	p1	J.Doe	27	X=1
t_2	2	p1	J.Doe	28	X=2
t_3	3	p2	K.Smith	32	Y=1
t_4	4	p2	S.Kmith	32	Y=2
t_5	5	p3	J.Doe	28	X=1 v X=3
t_6	6	p3	J.Ho	29	X=2

World-Table

var	value	Prob
X	1	0.6
X	2	0.2
X	3	0.2
Y	1	0.8
Y	2	0.2

Probabilistic Conditional Databases (pc-databases)

- The same variable can appear in conditions of different tuples

⇒ Variables can be used to introduce tuple correlations

Exclusion

Person

	<u>RK</u>	<u>WK</u>	name	age	condition
t_1	1	p1	J.Doe	27	X=1
t_2	2	p1	J.Doe	28	X=2
t_3	3	p2	K.Smith	32	Y=1
t_4	4	p2	S.Kmith	32	Y=2
t_5	5	p3	J.Doe	28	X=1 v X=3
t_6	6	p3	J.Ho	29	X=2

World-Table

var	value	Prob
X	1	0.6
X	2	0.2
X	3	0.2
Y	1	0.8
Y	2	0.2

Probabilistic Conditional Databases (pc-databases)

- The same variable can appear in conditions of different tuples

⇒ Variables can be used to introduce tuple correlations

Person

	RK	WK	name	age	condition
t_1	1	p1	J.Doe	27	X=1
t_2	2	p1	J.Doe	28	X=2
t_3	3	p2	K.Smith	32	Y=1
t_4	4	p2	S.Kmith	32	Y=2
t_5	5	p3	J.Doe	28	X=1 ∨ X=3
t_6	6	p3	J.Ho	29	X=2

positive Implication

World-Table

var	value	Prob
X	1	0.6
X	2	0.2
X	3	0.2
Y	1	0.8
Y	2	0.2

Probabilistic Conditional Databases (pc-databases)

- Tuples can be exclusive

⇒ Different tuples can share the same world key value

⇒ World key cannot be used as representation key

Person

	RK	WK	name	age	condition
t_1	1	p1	J.Doe	27	X=1
t_2	2	p1	J.Doe	28	X=2
t_3	3	p2	K.Smith	32	Y=1
t_4	4	p2	S.Kmith	32	Y=2
t_5	5	p3	J.Doe	28	X=1 v X=3
t_6	6	p3	J.Ho	29	X=2

World-Table

var	value	Prob
X	1	0.6
X	2	0.2
X	3	0.2
Y	1	0.8
Y	2	0.2

Probabilistic Conditional Databases (pc-databases)

Variable Assignment:

- A variable assignment θ maps each random variable to one of its possible values

$\Rightarrow$ $\theta(X) = 1$ means that assignment θ maps variable X to value 1

- All variables are mutually independent

$\Rightarrow$ Probability of assignment θ

$$Prob(\theta) = \prod_{X \in \mathbf{X}} Prob(X = \theta(X))$$

where $Prob(X = \theta(X))$ is the probability that variable X takes value $\theta(X)$.

Probabilistic Conditional Databases (pc-databases)

Variable Assignment (Example):

World-Table

var	*value*	*Prob*
X	1	0.6
X	2	0.2
X	3	0.2
Y	1	0.8
Y	2	0.2

- $3 \times 2 = 6$ possible variable assignments

	$Y=1$	$Y=2$
$X=1$	θ_1	θ_2
$X=2$	θ_3	θ_4
$X=3$	θ_5	θ_6

- Probability of assignment θ_2 is

$$Prob(\theta_2) = Prob(X=1) \times Prob(Y=2)$$
$$= 0.6 \times 0.2 = 0.12$$

Probabilistic Conditional Databases (pc-databases)

Marginal Tuple Probabilities:

- Let Θ be the set of all possible variable assignments
- Let Φ_t be the condition of tuple t
- The marginal probability of a tuple t results from summing up the probabilities of all variable assignments that satisfy condition Φ_t, i.e.

$$p(t) = \sum_{\theta \in \Theta, \Phi_t(\theta) = true} Prob(\theta)$$

Probabilistic Conditional Databases (pc-databases)

Marginal Tuple Probabilities (Example):

Person

	<u>RK</u>	<u>WK</u>	name	age	condition
t_1	1	p1	J.Doe	27	X=1
t_2	2	p1	J.Doe	28	X=2
t_3	3	p2	K.Smith	32	Y=1
t_4	4	p2	S.Kmith	32	Y=2
t_5	5	p3	J.Doe	28	X=1 v X=3
t_6	6	p3	J.Ho	29	X=2

World-Table

var	value	Prob
X	1	0.6
X	2	0.2
X	3	0.2
Y	1	0.8
Y	2	0.2

- Condition of tuple t_5 is satisfied if $\theta(X) = 1$ or $\theta(X) = 3$

$$p(t_5) = Prob(\theta_1) + Prob(\theta_2) + Prob(\theta_5) + Prob(\theta_6)$$

$$= Prob(X = 1) + Prob(X = 3) = 0.6 + 0.2 = 0.8$$

Probabilistic Conditional Databases (pc-databases)

Possible World Generation (formal):

- One possible world per variable assignment
- Let *pdb* be a pc-database
- Let $W^{\theta}_{pdb} = \{t \mid t \in pdb, \Phi_t(\theta) = true\}$ be the world that results from assignment θ

Possible world space:

$$\mathbf{W} = pws(pdb) = \{W^{\theta}_{pdb} \mid \theta \in \Theta\}$$

Probability of a possible world $W \in \mathbf{W}$:

$$Pr(W) = \sum_{\theta \in \Theta, W^{\theta}_{pdb} = W} Prob(\theta)$$

Probabilistic Conditional Databases (pc-databases)

Possible World Generation (example):

Person

	RK	*WK*	*name*	*age*	*condition*
t_1	*1*	*p1*	*J.Doe*	*27*	X=1
t_2	*2*	*p1*	*J.Doe*	*28*	X=2
t_3	*3*	*p2*	*K.Smith*	*32*	Y=1
t_4	*4*	*p2*	*S.Kmith*	*32*	Y=2
t_5	*5*	*p3*	*J.Doe*	*28*	X=1 v X=3
t_6	*6*	*p3*	*J.Ho*	*29*	X=2

World-Table

var	*value*	*Prob*
X	1	0.6
X	2	0.2
X	3	0.2
Y	1	0.8
Y	2	0.2

Probabilistic Conditional Databases (pc-databases)

Possible World Generation (example):

Possible world $W_1 = \{t_1, t_3, t_5\}$:

Person

	RK	WK	name	age	condition
t_1	1	p1	J.Doe	27	X=1
t_2	2	p1	J.Doe	28	X=2
t_3	3	p2	K.Smith	32	Y=1
t_4	4	p2	S.Kmith	32	Y=2
t_5	5	p3	J.Doe	28	X=1 v X=3
t_6	6	p3	J.Ho	29	X=2

World-Table

var	value	Prob
X	1	0.6
X	2	0.2
X	3	0.2
Y	1	0.8
Y	2	0.2

$$Pr(W_1) = Prob(X = 1) \times Prob(Y = 1)$$
$$= 0.6 \times 0.8 = \mathbf{0.48}$$

Probabilistic Conditional Databases (pc-databases)

Possible World Generation (example):

Possible world $W_2 = \{t_2, t_3, t_6\}$:

Person

	<u>*RK*</u>	<u>*WK*</u>	*name*	*age*	*condition*
t_1	*1*	*p1*	*J.Doe*	*27*	X=1
t_2	*2*	*p1*	*J.Doe*	*28*	X=2
t_3	*3*	*p2*	*K.Smith*	*32*	Y=1
t_4	*4*	*p2*	*S.Kmith*	*32*	Y=2
t_5	*5*	*p3*	*J.Doe*	*28*	X=1 v X=3
t_6	*6*	*p3*	*J.Ho*	*29*	X=2

World-Table

var	*value*	*Prob*
X	1	0.6
X	2	0.2
X	3	0.2
Y	1	0.8
Y	2	0.2

$$Pr(W_2) = Prob(X = 2) \times Prob(Y = 1)$$

$$= 0.2 \times 0.8 = \mathbf{0.16}$$

Probabilistic Conditional Databases (pc-databases)

Possible World Generation (example):

Possible world $W_3 = \{t_3, t_5\}$:

Person

	<u>RK</u>	<u>WK</u>	name	age	condition
t_1	1	p1	J.Doe	27	X=1
t_2	2	p1	J.Doe	28	X=2
t_3	3	p2	K.Smith	32	Y=1
t_4	4	p2	S.Kmith	32	Y=2
t_5	5	p3	J.Doe	28	X=1 v X=3
t_6	6	p3	J.Ho	29	X=2

World-Table

var	value	Prob
X	1	0.6
X	2	0.2
X	3	0.2
Y	1	0.8
Y	2	0.2

$$Pr(W_3) = Prob(X = 3) \times Prob(Y = 1)$$

$$= 0.2 \times 0.8 = \mathbf{0.16}$$

Probabilistic Conditional Databases (pc-databases)

Possible World Generation (example):

Possible world $W_4 = \{t_1, t_4, t_5\}$:

Person

	RK	WK	name	age	condition
t_1	1	p1	J.Doe	27	X=1
t_2	2	p1	J.Doe	28	X=2
t_3	3	p2	K.Smith	32	Y=1
t_4	4	p2	S.Kmith	32	Y=2
t_5	5	p3	J.Doe	28	X=1 v X=3
t_6	6	p3	J.Ho	29	X=2

World-Table

var	value	Prob
X	1	0.6
X	2	0.2
X	3	0.2
Y	1	0.8
Y	2	0.2

$$Pr(W_4) = Prob(X = 1) \times Prob(Y = 2)$$

$$= 0.6 \times 0.2 = \mathbf{0.12}$$

Probabilistic Conditional Databases (pc-databases)

Possible World Generation (example):

Possible world $W_5 = \{t_2, t_4, t_6\}$:

Person

	RK	WK	name	age	condition
t_1	1	p1	J.Doe	27	X=1
t_2	2	p1	J.Doe	28	X=2
t_3	3	p2	K.Smith	32	Y=1
t_4	4	p2	S.Kmith	32	Y=2
t_5	5	p3	J.Doe	28	X=1 v X=3
t_6	6	p3	J.Ho	29	X=2

World-Table

var	value	Prob
X	1	0.6
X	2	0.2
X	3	0.2
Y	1	0.8
Y	2	0.2

$$Pr(W_5) = Prob(X = 2) \times Prob(Y = 2)$$

$$= 0.2 \times 0.2 = \mathbf{0.04}$$

Probabilistic Conditional Databases (pc-databases)

Possible World Generation (example):

Possible world $W_6 = \{t_4, t_5\}$:

Person

	RK	WK	name	age	condition
t_1	1	p1	J.Doe	27	X=1
t_2	2	p1	J.Doe	28	X=2
t_3	3	p2	K.Smith	32	Y=1
t_4	4	p2	S.Kmith	32	Y=2
t_5	5	p3	J.Doe	28	X=1 v X=3
t_6	6	p3	J.Ho	29	X=2

World-Table

var	value	Prob
X	1	0.6
X	2	0.2
X	3	0.2
Y	1	0.8
Y	2	0.2

$$
\begin{aligned}
Pr(W_6) &= Prob(X = 3) \times Prob(Y = 2) \\
&= 0.2 \times 0.2 = \mathbf{0.04}
\end{aligned}
$$

Probabilistic Conditional Databases (pc-databases)

Transformation from BID-database to pc-database:

- One variable per block
- Certain block B

$\Rightarrow$ Variable has $|B|$ possible values

- Maybe block B

$\Rightarrow$ Variable has $|B| + 1$ possible values

Consequences:

- No loss in compactness
- Increase in modeling/query complexity

Probabilistic Conditional Databases (pc-databases)

Computation of the most probable world:

Case 1: Every assignment leads to another possible world:

- Select the most probable value per variable
- Compute all tuples whose conditions are satisfied by the selected assignment

Case 2: Different assignments lead to the same possible world:

- Compute all assignments
- Compute all possible worlds

$\Rightarrow$ Infeasible in practice

Representation Systems - Overview

TI-database

- uncertain tuples (mutually independent)

AOR-database

- alternative values per attribute (mutually independent)

AOR?-database

- uncertain tuples with alternative values per attribute

BID-database

- mutually independent blocks of exclusive tuples

pc-database

- tuple conditions defined on independent random variables

$\Rightarrow$ any correlation possible

Properties: Completeness

Definition: A representation system is called *complete* if it can be used to represent any discrete probability distribution over a set of possible worlds.

- pc-databases are complete
- BID-databases are not complete

$\Rightarrow$ TI-databases, AOR-databases and AOR?-databases are not complete

Properties: Closeness

Definition: A representation system is called *closed* under a query language if the result of each query of this language can be represented with this system.

- pc-databases are complete

$\Rightarrow$ pc-databases are closed under every query language

- BID-databases are not closed under the join-operator

$\Rightarrow$ TI-databases, AOR-databases and AOR?-databases are not closed under the join-operator

- AOR-databases are not closed even under the selection-operator

Properties: Closeness - Example

Person

	RK	WK	name	age	p
t_1	1	p1	J.Doe	27	0.6
t_2	2	p1	J.Doe	28	0.2
t_3	3	p2	K.Smith	32	1.0
t_4	4	p3	J.Doe	28	0.8
t_5	5	p3	J.Ho	29	0.2

SELECT t.name **AS** nameA, u.name **AS** nameB
FROM Person t, Person u
WHERE t.WK <> u.WK
AND t.age < u.age

W_A, Pr=0.48

	nameA	nameB
t_6	J.Doe	K.Smith
t_7	J.Doe	J.Doe

W_B, Pr=0.32

	nameA	nameB
t_6	J.Doe	K.Smith

W_C, Pr=0.16

	nameA	nameB
t_6	J.Doe	K.Smith
t_8	J.Ho	K.Smith
t_9	J.Doe	J.Ho

W_D, Pr=0.04

	nameA	nameB
t_8	J.Ho	K.Smith

Tuple t_9 pos. implicates tuple t_8 $\Rightarrow$ cannot be represented with a BID-database

Coupling Representation Systems with Views

Observations:

- Queries can introduce tuple dependencies
- BID- and TI-databases are not complete by themselves, but are complete if they are combined with views

Benefits:

- Simple representation system is used and dependencies are introduced on demand
- We can control which dependencies are allowed to exist in the database

⇒ Specific dependency assumptions can be made for such views

⇒ Often more efficient querying than in pc-databases

- Useful if many tuples are correlated in the same way
- Less useful if many tuples are correlated in different ways (one view per individual dependency?)

Choice of Representation System - Examples

Uncertain existence (or relevance) of individual persons

$\Rightarrow$ TI, AOR?, BID, pc

Uncertain attribute values of individual persons

$\Rightarrow$ AOR, AOR?, BID, pc

Correlations between different attribute values of the same person

$\Rightarrow$ BID, pc

Correlations between attribute values of different persons

$\Rightarrow$ pc

Exclusive existences of different persons

$\Rightarrow$ BID, pc

Correlations between existences of different persons

$\Rightarrow$ pc

Choice of Representation System - Use Cases

Use Case 1: Duplicate Merging

- Given: Set of duplicate tuples with conflicting values
- Problem: Uncertainty on correct values for some attributes

PNo.	firstname	lastname	DoB	city
P23	William	Schulz	12.10.1987	HH
P14	Bill	Schultz	10.12.1987	St.Pauli
P31	William	Schultz	$\perp$	Berlin

Solutions:

- Requires exclusion between alternative values or tuples
- AOR-database (loss of correlations between values)
- BID-database (no loss of value correlations)

Choice of Representation System - Use Cases

Use Case 2: Duplicate Detection

- Given: Set of potential duplicates
- Problem: Uncertainty whether or not these tuples are duplicates

PNo.	firstname	lastname	DoB	city
P23	William	Schulz	12.10.1987	HH
P14	Bill	Schultz	10.12.1987	St.Pauli

Solutions:

- Two cases: Different persons (2 tuple), same person (1 tuple)

⇒ Requires modeling of complex relationships

- BID-database with view
- pc-database