Completeness of Attribute Values Representing Partial Information

Fabian Panse
panse@informatik.uni-hamburg.de

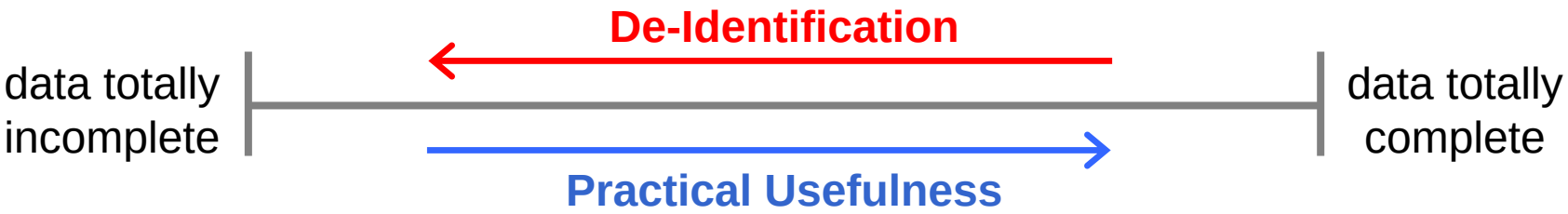
University of Hamburg
Germany

Content

- Motivation
- Completeness in Relational Databases
- Information classification
- Completeness metrics w.r.t. attribute values
  - Countable and finite domains
  - Countable and infinite domains
  - Uncountable and bounded domains
- Conclusion and Future work

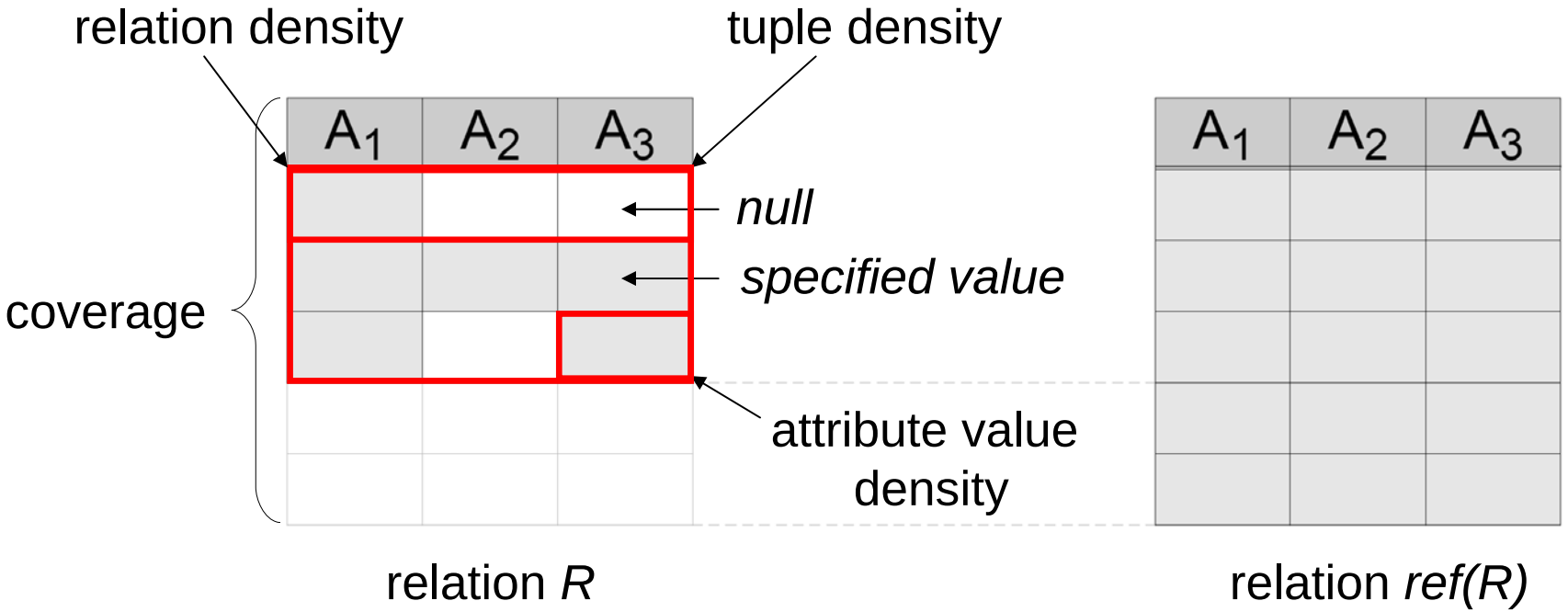
Motivation

- Quality measuring in extended data models
  - Partial information results from:
    - Imperfect information elicitation
    - Information Integration (resolving contradictions)
- Data Anonymization
  - Balancing between de-identification and practical usefulness



Age:	*null*		10-40		20-30	25
ZIP-Code:	*null*	2****	22***	223**	2235*	22351

Completeness in Relational Databases



Decomposition into coverage and density

- **Coverage:** ratio of stored to existing objects
- **Density:** completeness of stored objects
  Different levels of granularity (e.g. relation, tuple, attribute value)

Completeness in Relational Databases (2)

- Density of attribute values

$$d(v) = \begin{cases} 1 & v \text{ is a } specified\ value \\ 0 & v \text{ is } null \end{cases}$$

- Undefined for values representing partial information

A

AGE
null

B

AGE
20-30

C

AGE
25

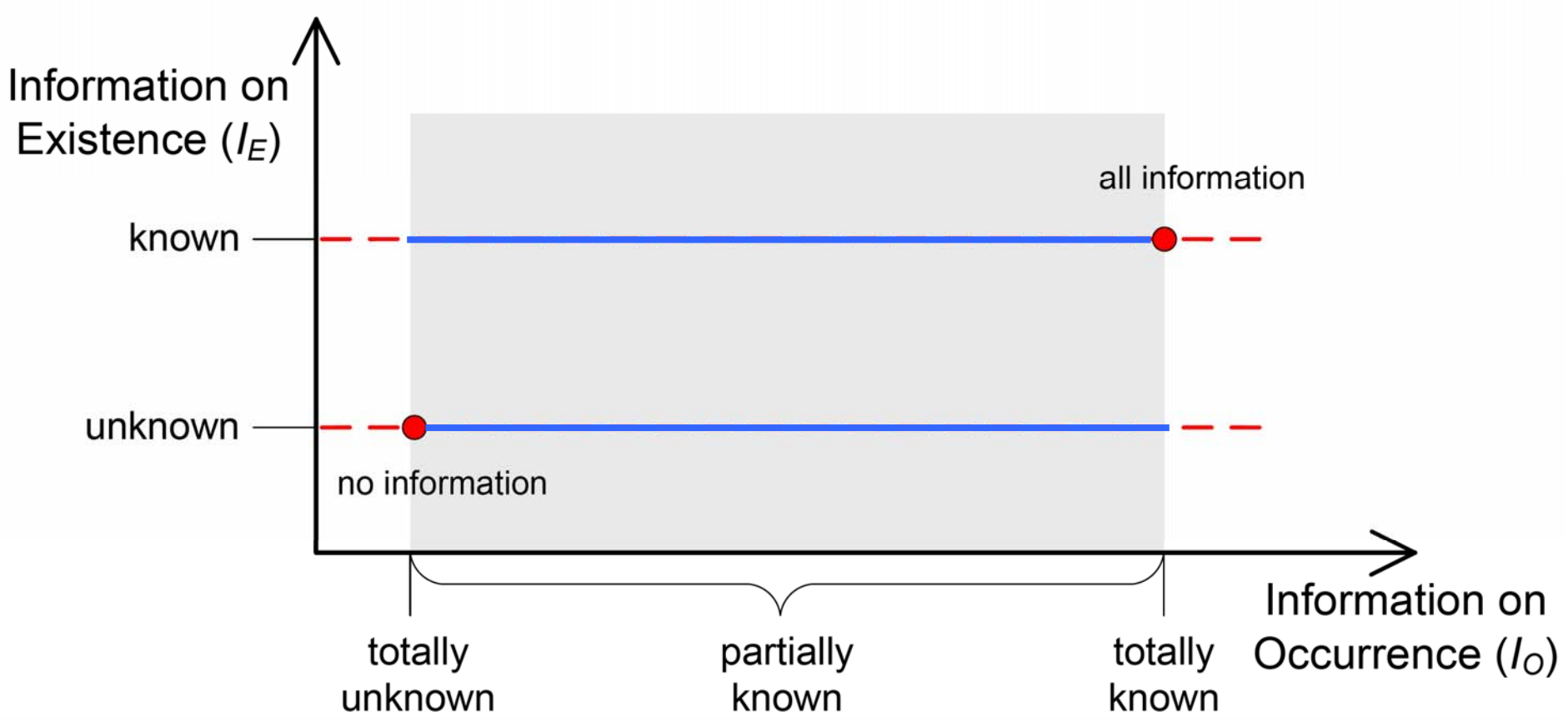
d(A) = 0

d(B) > d(A) $\wedge$ d(B) < d(C)

$\Rightarrow$ 0 < d(B) < 1

d(C) = 1

Information Classification



Density of Attribute Values

Decomposition into

- Density of Existence Information (d_E)

$$d_E(v) = \begin{cases} 1 & \text{existence is known} \\ 0 & \text{else} \end{cases}$$

- Density of Occurrence Information (d_O)

Application domain depending weighting

$$\Rightarrow \quad d(v) = w_E \cdot d_E(v) + w_O \cdot d_O(v)$$

where $w_O = 1 - w_E$

Density of Occurrence Information

- Representation of Occurrence Information by a partial set ($S_V \subseteq D_A$)

- Context-based requirements
  - $S_V = D_A$ $\Rightarrow$ totally unknown $\Rightarrow d_O=0$
  - $|S_V| = 1 \vee |S_V| = 0$ $\Rightarrow$ totally known $\Rightarrow d_O=1$
  - $S_V \subset D_A \wedge |S_V| > 1$ $\Rightarrow$ partially known $\Rightarrow d_O \in (0,1)$

- General metric requirements
  - Normalization
  - Interpretability
  - Interval Scale

Countable and Finite Domains

Number of domain elements
which are already excluded

$$d_O(v) = min(1, \frac{|D_A| - |S_V|}{|D_A| - 1})$$

Number of domain elements
which have to be totally excluded

Example: The color of a car

(D_A={white,green,red,blue,black}, S_V={green,red,blue})

$\Rightarrow$ $d_O(v)$ = min(1,(5-3)/4) = 0.5

Countable and Infinite Domains

- Previous metric is impractical ($|D_A| \Rightarrow$ infinity)
- Scopes of application domains are actually not Infinite (e.g. no person is older than 130 years old)
- Restriction to a finite domain ($D_A^{[l,u]}$) by
  - Knowledge of domain experts
  - Database statistics

Example: The age of a person ($l=0 \wedge u=130$)

($D_A=\mathbb{N}_0$, $S_V=\{v \in \mathbb{N}_0 | v>17\}$)

$\Rightarrow d_O(v) = \min(1,(131-113)/130) = 0.138$

Uncountable and Bounded Domains

Size of domain subset
which is already excluded

$$d_O(v) = \frac{|D_A| - |S_V|}{|D_A|}$$

Size of domain subset
which has to be totally excluded

Example: A test result in percent

(D_A={v∈R| v∈[0,100]}, S_V={v∈R| v≥50})

⇒ d_O(v) = (100-50)/100 = 0.5

Conclusion and Future Look

Conclusion

- Completeness metrics of attribute values
  are undefined for values representing partial information
- Definition of new density metrics of attribute values
  w.r.t. information on existence and information on occurrence
- Consideration of different types of attribute domains

Future Work

- Adaption on possibility/probability distributions of
  - Domains (e.g. more people are 20 than 95 years old)
  - Occurrence Information (e.g. for the age of John D.,
    20 years is more likely than 95 years)
  - Existence Information
    (e.g. it is very likely that John D. has a Phone)

END

Thank you very much for your attention!

Any questions ?