

# **Seminar Impfektion und Datenbanken**

Schemaerweiterung zur Abbildung von  
Imperfekten Daten

Andreas Merkel

# Inhalt

---

## Einführung

- Eigenschaften des relationalen Modells
- Erweiterungsmöglichkeiten

## Zwei unterschiedliche Ansätze:

- **Schemaerweiterung zur unscharfen Klassifikation**
- **Schemaerweiterung zur Einbettung von Subrelationen**

# Eigenschaften des relationalen Modells

---

## Attributwerte sind atomar

=> keine Repräsentation von:

- unscharfen Informationen
- stochastischen Unsicherheiten
- Daten, die nicht in den Wertebereich passen
- unvollständigen Informationen

**Baustellenrelation (**

**Autobahn:    STRING ,**

**Länge:        INT ,**

**Ort:           INT ,**

**Baubeginn:   DATE )**

# Erweiterungsmöglichkeiten

---

## 1. Unscharfes Klassifizieren von scharfen Datenbeständen

-> Schemaerweiterung zur unscharfen Klassifikation

## 2. Fallenlassen der Forderung nach Atomizität der Attributwerte

-> Schemaerweiterung zur Einbettung von Subrelationen

# Schemaerweiterung zur unscharfen Klassifikation

---

## Motivation:

- zahl- und umfangreiche scharfe Datenbestände existieren bereits
- Einteilung in Klassen ist wünschenswert  
-> Komplexitätsreduktion
- Oft ist man an scharfer Einteilung nicht interessiert
- Unscharfe Klassifikation scharfer Daten vermeidet Konversionsaufwand

# Schemaerweiterung durch Kontexte

---

## Definition Kontext:

Einteilung des Wertebereichs eines Attributs in Äquivalenzklassen

## Beispiel:

$D(\text{Staulänge}) = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$

$K(\text{Staulänge}) = \{\{1, 2, 3, 4\}, \{5, 6, 7, 8, 9, 10\}\}$

$D(\text{Stauhäufigkeit}) = \{\text{häufig}, \text{gelegentlich}, \text{selten}, \text{nie}\}$

$K(\text{Stauhäufigkeit}) = \{\{\text{häufig}, \text{gelegentlich}\}, \{\text{selten}, \text{nie}\}\}$

# Schemaerweiterung durch Kontexte

$K(\text{Staulänge}) = \{\{1, 2, 3, 4\}, \{5, 6, 7, 8, 9, 10\}\}$

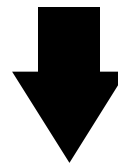
$K(\text{Stauhäufigkeit}) = \{\{\text{häufig, gelegentlich}\}, \{\text{selten, nie}\}\}$

<b>Staulänge</b>	10	<b>C3</b>	<b>C4</b>		
	9	weiträumig	unvorhersehbare		
	8	umfahren	Verzögerungen		
	7				
	6				
	5				
	4	<b>C1</b>	<b>C2</b>		
	3	zusätzlich Zeit	unbedenklich		
	2	einplanen	befahrbar		
	1				
		häufig	gelegentlich	selten	nie

**Stauhäufigkeit**

# Schemaerweiterung durch Kontexte

Autobahn	Abschnitt	Stauhäufigkeit	Staulänge	Klasse
A5	A5-1	gelegentlich	3	C1
A5	A5-2	selten	2	C2
A5	A5-3	nie	4	C2
A8	A8-1	häufig	7	C3
A8	A8-2	gelegentlich	8	C3
A81	A81-1	nie	3	C2
A81	A81-2	selten	5	C4



Autobahn	Abschnitt	Stauhäufigkeit	Staulänge	Klasse
{A5}	{A5-1}	{gelegentlich}	{3}	C1
{A5, A81}	{A5-2, A5-3, A81-1}	{selten, nie}	{2, 3, 4}	C2
{A8}	{A8-1, A8-2}	{häufig, gelegentlich}	{7, 8}	C3
{A81}	{A81-2}	{selten}	{5}	C4

# unscharfe Zuordnung

bisher: scharfe Zuordnung  
jedes Tupel gehört zu genau einer Klasse

jetzt: unscharfe Zuordnung  
ein Tupel kann zu mehreren Klassen gehören

<b>Autobahn</b>	<b>Stauhäufigkeit</b>	<b>Staulänge</b>	<b>Klasse</b>
{A5}	{gelegentlich}	{3}	C1
{A5, A81}	{selten, nie}	{2, 3, 4}	C2
{A8}	{häufig, gelegentlich}	{7, 8}	C3
{A81}	{selten}	{5}	C4

# linguistische Variable

Ziel: semantische Beschreibung unscharfer Klassen

linguistische Variable

Staulänge

Terme

kurz

lang

Zugehörigkeitsfunktionen

1,00  
0,75  
0,50  
0,25  
0,00

$\mu_{\text{lang}}$

$\mu_{\text{kurz}}$

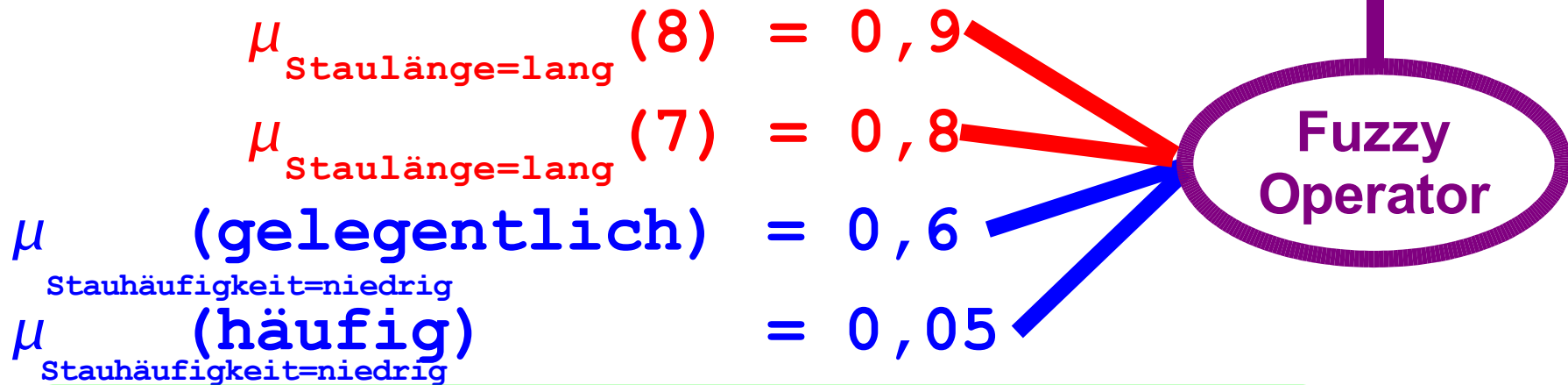
D(Staulänge)

1 2 3 4 5 6 7 8 9 10

# Klassenzuordnung durch Aggregation

Aus den Werten der Zugehörigkeitsfunktionen kann die Zugehörigkeit zu den unscharfen Klassen berechnet werden:

Autobahn	Stauhäufigkeit	Staulänge	Klasse
{A5: 0,3}	hoch	kurz	C1
{A5: 0,7, A8: 0,4}	niedrig	kurz	C2
{A81: 1,0}	hoch	lang	C3
{A8: 0,6}	<u>niedrig</u>	<u>lang</u>	C4



{A8}	{häufig, gelegentlich}	{7, 8}
------	------------------------	--------

# die Anfragesprache fCQL

---

- Anfragesprache zur Definition und Abfrage unscharfer Klassen
- Arbeitet auf der linguistischen Ebene
- Entwickelt an der Universität Fribourg

## Beispiel:

**classify** Autobahn **from** Staurelation **with**  
Staulänge **is** kurz **and** Stauhäufigkeit **is**  
niedrig

# Erweiterungsmöglichkeiten

---

## 1. Unscharfes Klassifizieren von scharfen Datenbeständen

-> Schemaerweiterung zur unscharfen Klassifikation

## 2. Fallenlassen der Forderung nach Atomizität der Attributwerte

-> Schemaerweiterung zur Einbettung von Subrelationen

# Motivation

---

- Relationales Modell legt für jedes Attribut einen Wertebereich fest
- => idealisierte Weltvorstellung
- Es gibt Ausnahmen, die nicht in das Schema passen

## **Beispiel:**

Ausländisches Autokennzeichen passt nicht in das vorgesehene Schema

*Buchstabengruppe-Buchstabengruppe-Zahl*

# Exceptional Conditions

---

**Exceptional Condition: Art, in der Daten von der Schemadefinition abweichen**

## **Beispiele:**

- ausländische Kennzeichen  
-> *passt nicht in vorgesehenen Wertebereich*
- Saisonkennzeichen  
-> *zusätzliche Angaben nötig*
- Moped mit Versicherungsplakette  
-> *semantisch gleichwertige Information*
- Unterschiedliche Kennzeichen auf Vorder- und Rückseite  
-> *Attribut hat mehrere Werte*

# eingebettete Attribute

Lokal werden zusätzliche Attribute eingeführt, um nicht ins Schema passende Informationen zu repräsentieren:

Farzeugtyp	Kennzeichen		Farbe
PKW	Seite	Kennzeichen	rot
	Vorderseite	KA-XY-42	
	Rückseite	KA-XY-23	

**e-Attribut** (red text and line pointing to 'Seite')

**s-Attribut** (blue text and line pointing to 'Kennzeichen')

# Einbettung von Subrelationen

s- und e-Attribute werden als Subrelation eingebettet:

<b>Fahrzeugtyp</b>	<b>Kennzeichen</b>			<b>Farbe</b>	
PKW	KA-AB-1			rot	
PKW	<b>Kennzeichen</b>	<b>Seite</b>		rot	
	KA-XY-42	Vorderseite			
	KA-XY-23	Rückseite			
LKW	KA-YZ-4711			<b>Farbe</b>	<b>Muster</b>
				gelb	grüne Streifen
PKW	<b>Kennzeichen</b>	<b>Saison von</b>	<b>Saison bis</b>	blau	
	KA-XZ-5	März	Oktober		
Moped	<b>Kennzeichen</b>	<b>Versicherungs-Nr</b>			silbergrau
		ABC-123			

# Anfragen an die erweiterte Relation

---

## Zwei Arten von Anfragen sind möglich:

1. Anfrage geht vom idealen Schema aus  
=> nur s-Attribute sind bekannt
2. Anfrage geht vom erweiterten Schema aus  
=> s- und e-Attribute sind bekannt

# Anfragen auf Grundlage des idealen Schemas

---

Anfragender erwartet, dass ideale Relation zurückgegeben wird.

**Statt dessen:**

<u>Kennzeichen</u>	
KA-AB-1	
<u>Kennzeichen</u>	<u>Seite</u>
KA-XY-42	Vorderseite
KA-XY-23	Rückseite

Neuformulierung der Anfrage unter Kenntnis der zusätzlichen Attribute:

<u>Kennzeichen</u>
KA-AB-1
KA-XY-42

# Anfragen unter Kenntnis der erweiterten Attribute

Anfrage nach Attributen auf der

Vorderseite:

**Kennzeichen**

KA-AB-1

KA-XY-42

KA-CD-7

Rückseite:

**Kennzeichen**

KA-AB-1

KA-XY-23

KA-CD-7

<b>Fahrzeugtyp</b>	<b>Kennzeichen</b>	<b>Seite</b>	<b>Farbe</b>
PKW	KA-AB-1	*	rot
PKW	KA-XY-42	Vorderseite	rot
PKW	KA-XY-23	Rückseite	rot
Motorrad	KA-CD-7	*	weiß

# Zusammenfassung

---

- **Unschärfe Anfragen auf scharfen Daten**
  - Klassifizierung zur Komplexitätsreduktion
  - Kontexte
  - unscharfe Klassifizierung
  - linguistische Variable
  - unscharfe Anfragen mit fCQL
- **Schemaerweiterung durch Subrelationen**
  - Exceptional Conditions
  - Eingebettete Attribute
  - Anfragen an die erweiterte Relation
    - auf Grundlage des idealen Schemas
    - auf Grundlage des erweiterten Schemas