



Der probabilistische Ansatz

Ying Zou

Universität Karlsruhe

Fakultät für Informatik

Institut für Programmstrukturen und Datenorganisation

Lehrstuhl für Systeme der Informationsverwaltung



Wahrscheinlichkeitstheorie

- Das beste verstandene mathematische Paradigma zur Modellierung und Behandlung von unbestimmten Informationen;
- Voraussetzung: Die Wahrscheinlichkeiten der komplexen Fälle können von denen der grundlegenden Fälle berechnet werden;
- Interpretationen : $P(A)$; $P(A | B)$

(weitere siehe 2. Vortrag Folie 13-16)



Ansätze zur probabilistischen Modellierung imperfekter Daten

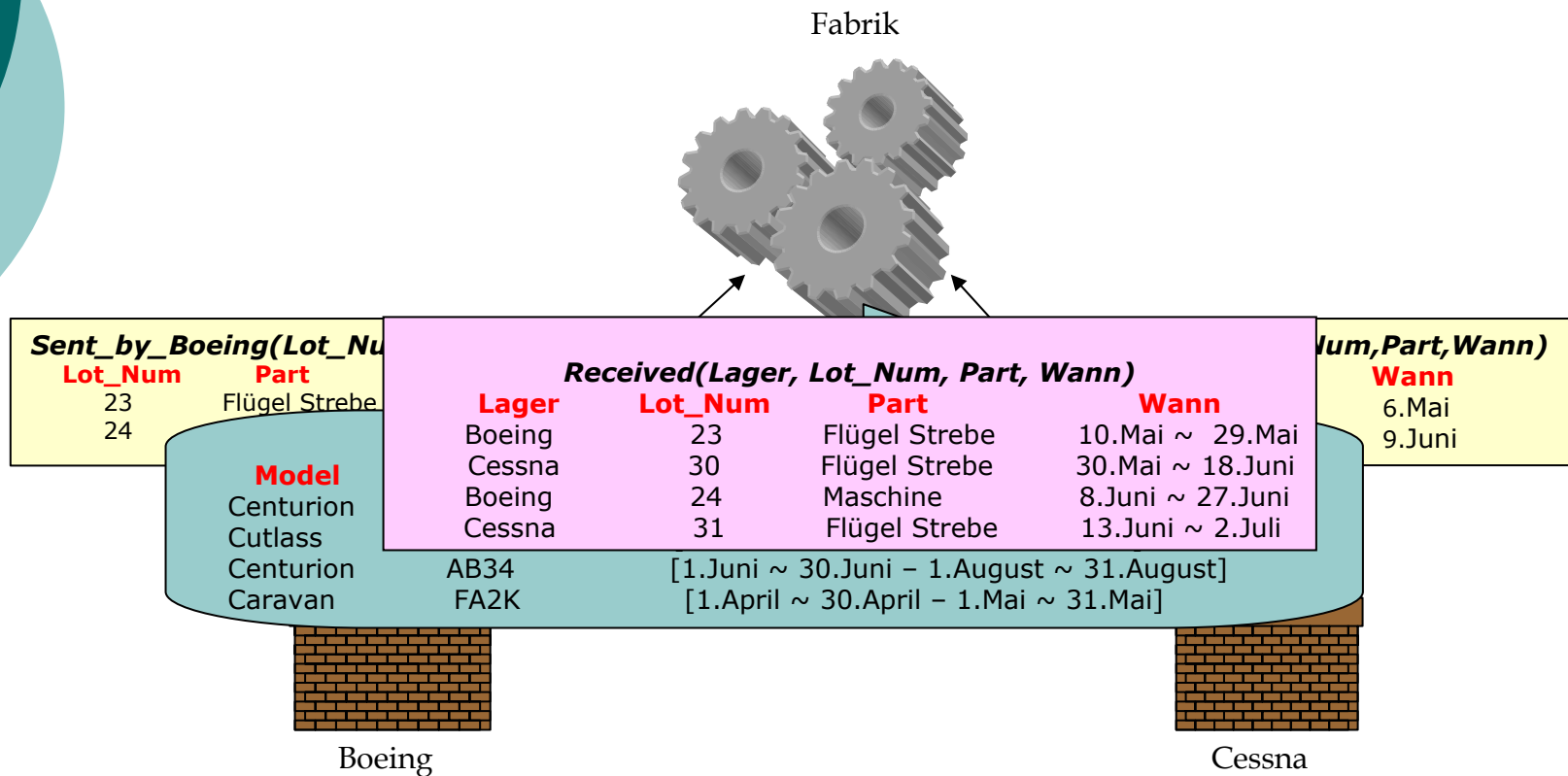
- ProbView

- . Ein Musterdatenbanksystem für probabilistische Daten ;
- . Algebraische Verfahren

- Erweiterung von SQL

- . Basiert auf SQL
- . Unterstützung der Unbestimmtheiten

Eine Beispieldatenbank



Die Relationen in Datenbank

<i>Sent_by_Boeing(Lot_Num,Part,Wann)</i>		
Lot_Num	Part	Wann
23	Flügel Strebe	6.Mai
24	Maschine	4.June

<i>Sent_by_Cessna(Lot_Num,Part,Wann)</i>		
Lot_Num	Part	Wann
30	Flügel Strebe	6.Mai
31	Flügel Strebe	9.Juni

Select **Lot_Num** from **Received**
Where **Part** = „ Flügel Strebe “
And **Wann** <= 30.Mai

Lot_Num	Während
Centurion	AB33 [1.März ~ 31.März - 1.Juni ~ 30.Juni]
Cutlass	Z19 [1.Juni ~ 30.Juni - 1.Juli ~ 31.Juli]
Centurion	AB34 [1.Juni ~ 30.Juni - 1.August ~ 31.August]
Caravan	FA2K [1.April ~ 30.April - 1.Mai ~ 31.Mai]

Unbestimmten Zeitraum

<i>Received(Lager, Lot_Num, Part, Wann)</i>			
Lager	Lot_Num	Part	Wann
Boeing	23	Flügel Strebe	10.Mai ~ 29.Mai
Cessna	30	Flügel Strebe	30.Mai ~ 18.Juni ?
Boeing	24	Maschine	8.Juni ~ 27.Juni
Cessna	31	Flügel Strebe	13.Juni ~ 2.Juli

Unbestimmten Zeitpunkte

e1
e2
e3
e4



Typen von Antworten

- Bestimmte Antworten : $P(E) = 1$
- Mögliche Antworten : $P(E) > 0$
- Wahrscheinliche Antworten : $P(E) \geq 0.5$

Select * from *Received* where *Wann* < 10.Juni

<i>Received(Lager, Lot_Num, Part, Wann)</i>			
Lager	Lot_Num	Part	Wann
Boeing	23	Flügel Strebe	10.Mai ~ 29.Mai
Cessna	30	Flügel Strebe	30.Mai ~ 18.Juni
Boeing	24	Maschine	8. Juni ~ 27.Juni
Cessna	31	Flügel Strebe	13.Juni ~ 2. Juli

Select * from *Received* where *Wann* < 10.Juni

Bestimmte Antworten

P(E) = 1

Lager	Lot_Num	Part	Wann
Boeing	23	Flügel Strebe	10.Mai ~ 29.Mai

Wahrscheinliche Antworten

P(E) = 0.55 > 0.5

Lager	Lot_Num	Part	Wann
Boeing	23	Flügel Strebe	10.Mai ~ 29.Mai
Cessna	30	Flügel Strebe	30.Mai ~ 18.Juni

Mögliche Antworten

P(E) = 0.10 < 0.5

Lager	Lot_Num	Part	Wann
Boeing	23	Flügel Strebe	10.Mai ~ 29.Mai
Cessna	30	Flügel Strebe	30.Mai ~ 18.Juni
Boeing	24	Maschine	8. Juni ~ 27.Juni



Beschreibungen zu unbestimmten Daten

10,Mai ~ 29,Mai

- Untere Beschränkung: α^* $\alpha^* = 10,\text{Mai}$
- Obere Beschränkung: α_* $\alpha_* = 29,\text{Mai}$
- Mass Funktion (p.m.f): P_α

$$P_\alpha(i) = \Pr[\alpha = i] \quad i \in \{0,1, \dots, N\}$$

Wobei $\Pr[\alpha = i]$ die Wahrscheinlichkeit ist, dass die unbestimmte Datum α zum Zeitpunkt i angeordnet ist ;

$$\Pr[i < \alpha_*] = 0 \quad \text{und} \quad \Pr[i > \alpha^*] = 0$$



Syntaktische Erweiterung von SQL(1)

Gebe bei Erzeugen einer Relation an, dass die Eigenschaft der Daten unbestimmt ist

- Füge den Modifikator entweder **INDETERMINATE** oder **INDETERMINATE COMPACT** vor der Eigenschaft der unbestimmten Daten hinzu;
- Füge einen optionalen Satz am Ende der Beschreibung der Eigenschaften hinzu, um die Mass Funktion als **standard** oder **nonstandard** anzugeben.

```
CREATE TABLE Received ( Lager          CHAR(30),  
                        Lot_Num       INTERGER,  
                        Part         CHAR(40),  
                        Wann         INDETERMINATE DATUM);  
CREATE TABLE In_Production ( Model     CHAR(30),  
                              Serial_Num CHAR(10),  
                              Während    INDETERMINATE ZEITRAUM(DATUM));  
ALTER TABLE Received ALTER COLUMN Wann TO NONSTANDARD DISTRIBUTION;
```



Syntaktische Erweiterung von SQL(2)

Bestimme die Stufe der Glaubwürdigkeit in der „*select...from...*“ Klausel

- Bezeichnet die Glaubwürdigkeit durch Ausdruck **WITH CREDIBILITY**;
- Benutze Ersetzung (Replacement) Verfahren
- 4 mögliche Stufe der Glaubwürdigkeit:
 - INDETERMINAT** -- Erhält alle Ungenauigkeiten;
 - EXPECTED** -- Ersetzt die ungenaue Daten mit dem erwarteten Wert;
 - MAX** -- Ersetzt die unbestimmten Daten mit untere Beschränkung
 - MIN** -- Ersetzt die unbestimmten Daten mit obere Beschränkung

```
SELECT Lager FROM Received WITH CREDIBILITY INDETERMINATE
```



Syntaktische Erweiterung von SQL(3)

Bestimme die Reihenfolge der Plausibilität

- Bezeichnet die Plausibilität durch Ausdruck `WITH PLAUSIBILITY` am Ende der `Where` Klausel; Oder durch Anweisung : `SET DEFAULT PLAUSIBILITY`;
- zwischen Integer `1` und `100` zugeordnet:
 - `1` Beliebige mögliche Antwort;
 - `100` Die bestimmte Antwort.

```
SET DEFAULT PLAUSIBILITY 60
```

```
SELECT Lager FROM Received  
WHERE Model = „Centurion“  
AND Wann < 29.Mai  
WITH PLAUSIBILITY 60
```



Syntaktische Erweiterung von SQL(4)

**Zeichen „~“, zeigt die Unbestimmtheit von einem
Zeitdatum**

Datum 5,Mai ~ 29,Mai

Semantik zu Unbestimmtheiten

Die Semantik von der Anweisung -- *Select* in SQL:

$$\begin{aligned} & \llbracket \text{SELECT } \langle \textit{target list} \rangle \text{ FROM } \langle \textit{from list} \rangle \text{ WHERE } \langle \textit{predicate} \rangle \rrbracket_{\text{SQL}}(d) \\ &= \llbracket \langle \textit{target list} \rangle \rrbracket_{\text{SQL}} \left(\llbracket \text{WHERE } \langle \textit{predicate} \rangle \rrbracket_{\text{SQL}} \left(\llbracket \langle \textit{from list} \rangle \rrbracket_{\text{SQL}}(d) \right) \right) \end{aligned}$$

d : Datenbank

Die Semantik zu unbestimmter Datenbank(Erweiterung der Semantik von SQL):

$$\begin{aligned} & \llbracket \text{SELECT } \langle \textit{target list} \rangle \text{ FROM } \langle \textit{from list} \rangle \text{ WHERE } \langle \textit{predicate} \rangle \rrbracket_{\text{ind}}(\delta, \gamma, d) \\ &= \llbracket \langle \textit{target list} \rangle \rrbracket_{\text{ind}}(\gamma, \llbracket \text{WHERE } \langle \textit{predcate} \rangle \rrbracket_{\text{ind}}(\gamma, \llbracket \langle \textit{from list} \rangle \rrbracket_{\text{ind}}(\delta, d))) \end{aligned}$$

δ : Glaubwürdigkeit

γ : Plausibilität

d : Datenbank



Unterschiede von der Erweiterung

1. Mit zwei zusätzlichen Parametern :

δ -- Glaubwürdigkeit γ - Plausibilität

2. Mit mehreren Interpretationen:

- Die **SELECT** Anweisung wendet nur die vollständige Informationen an und hat unter SQL Semantik eine einzelne Interpretation;
- Für die unvollständigen Informationen braucht die Semantik mindestens zwei Interpretationen:

Eine Abfrage wählt die Informationen aus durch **mögliche Zuordnung**

→ **Mögliche Interpretation**

Eine Abfrage wählt die Informationen aus durch **bestimmte Zuordnung**

→ **Bestimmte Interpretation**



Ein Konzept zu Beschreibung der Interpretation

Es ist wichtig zu garantieren, dass eine Abfrage keine unmögliche Antworten erzeugen wird, d.h.: die Antworten der Abfrage soll eine Submenge zu möglicher Interpretation, bzw. eine Obermenge zu bestimmte Interpretation sein.

Wir stellen uns ein unbestimmte Zeitpunkt als eine Menge von möglichen Zeitpunkte vor, einer ist „echt“, aber welcher ist unbekannt. Jedes möglicher Zeitpunkt stellt eine verschiedene, vollständige Realität dar. Jede Möglichkeit wird eine Vervollständigung eines Zeitpunktes benannt.

Vervollständigung (Completion) eines unbestimmten Zeitpunktes

Sei $\alpha = (\alpha_* \sim \alpha^*, \mathbf{P}_\alpha)$. Eine Vervollständigung eines unbestimmten Zeitpunktes α ist α_i ,

Wobei α_i ein bestimmte Zeitpunkt ist, sodass $\alpha_* \leq \alpha_i \leq \alpha^*$.

Die Menge allen Vervollständigungen für ein Zeitpunkt α bezeichnet man als $\mathbf{C}(\alpha)$.

* Statt Zeitpunkt kann auch Zeitraum, Intervall verwendet werden. Wenn die unbestimmte Zeitelement in einem Tupel aufgetaucht wird, nennen wir die Menge des Tupels als $\mathbf{C}(\mathbf{t})$.

Interpretationen zu Unbestimmtheiten

- **Bestimmte Interpretation**, Bsp.: in *WHERE* Klausel, $\gamma=100$

$$\llbracket \text{WHERE} \langle \text{predicate} \rangle \rrbracket_{\text{ind}}(100, r) = \{ t \mid t \in r \wedge \forall t' \in C(t) (\llbracket \langle \text{predicate} \rangle \rrbracket_{\text{SQL}}(t')) \}$$

- **Mögliche Interpretation**, mit Plausibilität 1,

$$\llbracket \text{WHERE} \langle \text{predicate} \rangle \rrbracket_{\text{ind}}(1, r) = \{ t \mid t \in r \wedge \exists t' \in C(t) (\llbracket \langle \text{predicate} \rangle \rrbracket_{\text{SQL}}(t')) \}$$

- Andere Interpretation, mit andere Wert von Glaubwürdigkeiten und Plausibilität

Eigenschaften der möglichen Interpretation :

- **Zuverlässig**

$$\forall r' \in C(r) [\llbracket W \rrbracket_{\text{SQL}}(r') \in C(\llbracket W \rrbracket_{\text{ind}}(1, r'))]$$

- **Maximal**

$$\forall c \in C(\llbracket W \rrbracket_{\text{ind}}(1, r)) [\exists r' \in C(r) (c = \llbracket W \rrbracket_{\text{SQL}}(r'))]$$



Operationale Semantik zu Unbestimmtheit

- Ziel: erhöhe die Effizienz
- 3 Veränderungen zur SQL – Semantik
 1. Predikate *Before*
 - Wahrscheinliche Anordnung für unbestimmte Zeitpunkt
 - Benutzt der Wert der Plausibilität in *WHERE*-Klausel
 2. 4-sortiger Wertbereich
 - Die Ausgabewerte von *Before*
 3. *Replace* Verfahren
 - Unterstütze die Glaubwürdigkeit
 - in *from*-Klausel



Wahrscheinliche Anordnung--*Before*

- Unter SQL-Semantik ordnet die Reihenfolge drei Zeitpunkte: E1,E2,E3 durch nachfolgende Form an; Der Wahrheitswert ist von den Ausgabewert von *Before* abhängig

$$\llbracket \langle E1 \rangle \text{ OVERLAPS Zeitraum}(\langle E2 \rangle, \langle E3 \rangle) \rrbracket = \\ \textit{Before}(\llbracket \langle E2 \rangle \rrbracket, \llbracket \langle E1 \rangle \rrbracket) \wedge \textit{Before}(\llbracket \langle E1 \rangle \rrbracket, \llbracket \langle E3 \rangle \rrbracket)$$

Beispiel: Unbestimmte Daten

$$E1 = [01,05,2003 \sim 31,05,2003] \\ E2 = [15,04,2003 \sim 15,05,2003] \quad E3 = [16,05,2003 \sim 31,06,2003]$$

- Erweiterung der Semantik zu Unbestimmtheit, füge in der Prädikate *Before* die Plausibilität hinzu; *Before* wird durch $\alpha \leq \beta$ dargestellt.

$$\textit{Before}(\alpha, \beta, \gamma) = \{ \textit{True} \mid \text{Pr}[\alpha \leq \beta] \times 100 \geq \gamma \} \\ \cup \{ \textit{False} \mid \text{Pr}[\beta < \alpha] \times 100 \geq \gamma \}$$

γ : Plausibilität α, β Zeitpunkt

4 – sortiger Wertbereich

Prädikate *Before*

$$\begin{aligned} \textit{Before}(\alpha, \beta, \gamma) = & \{ \textit{True} \mid \text{Pr}[\alpha \leq \beta] \times 100 \geq \gamma \} \\ & \cup \{ \textit{False} \mid \text{Pr}[\beta < \alpha] \times 100 \geq \gamma \} \end{aligned}$$

4 Werte für die Ausgabe der Prädikate *Before*

$\{\}$: gilt weder $\alpha \leq \beta$, noch die Negation;
$\{\textit{True}\}$: gilt nur $\alpha \leq \beta$, aber die Negation nicht;
$\{\textit{False}\}$: gilt nicht $\alpha \leq \beta$, aber die Negation gilt;
$\{\textit{True}, \textit{False}\}$: $\alpha \leq \beta$ und die Negation haben gleiches Gewicht.

Prädikate und Logische Formel

$$\llbracket \langle \textit{pred1} \rangle \textit{AND} \langle \textit{pred2} \rangle \rrbracket (\gamma, r) = \llbracket \langle \textit{pred1} \rangle \rrbracket (\gamma, r) \cap \llbracket \langle \textit{pred2} \rangle \rrbracket (\gamma, r)$$

$$\llbracket \langle \textit{pred1} \rangle \textit{OR} \langle \textit{pred2} \rangle \rrbracket (\gamma, r) = \llbracket \langle \textit{pred1} \rangle \rrbracket (\gamma, r) \cup \llbracket \langle \textit{pred2} \rangle \rrbracket (\gamma, r)$$

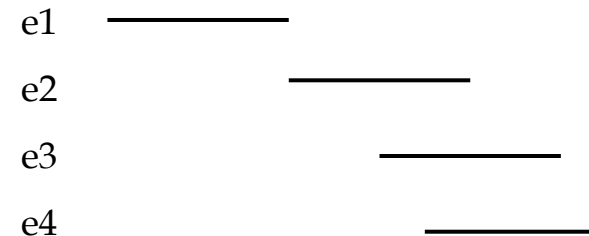
$$\llbracket \textit{NOT} \langle \textit{predicate} \rangle \rrbracket (\gamma, r) = \{x \mid \neg x \in \llbracket \langle \textit{predicate} \rangle \rrbracket (\gamma, r)\}$$

r: Relation

γ : Plausibilität

Beispiel zu *Before* und Wertbereich

	e1	e2	e3	e4
e1	1.00	1.00	1.00	1.00
e2	0	1.00	0.86	0.96
e3	0	0.16	1.00	0.73
e4	0	0.05	0.27	1.00



$[[e2 \leq e3]] (100, r) = \text{before}(e2, e3, 100) = \{\}$	$[[\text{NOT } \langle e2 \leq e3 \rangle]] (100, r) = \{\}$
$[[e2 \leq e3]] (50, r) = \{\text{True}\}$	$[[\text{NOT } \langle e2 \leq e3 \rangle]] (50, r) = \{\text{False}\}$
$[[e2 \leq e3]] (1, r) = \{\text{True}, \text{False}\}$	$[[\text{NOT } \langle e2 \leq e3 \rangle]] (1, r) = \{\text{True}, \text{False}\}$
$[[e2 \leq e3 \text{ AND } e1 \leq e4]] (100, r) = \{\}$	$[[\text{NOT } (\langle e2 \leq e3 \rangle \text{ AND } \langle e1 \leq e4 \rangle)]] (100, r) = \{\}$
$[[e2 \leq e3 \text{ OR } e1 \leq e4]] (100, r) = \{\text{True}\}$	$[[\text{NOT } (\langle e2 \leq e3 \rangle \text{ OR } \langle e1 \leq e4 \rangle)]] (100, r) = \{\text{True}\}$

r: relation

e: Zeitelement

Replace Verfahren

- Unterstütze die Glaubwürdigkeit
- in *from*-Klausel

	Zeitpunkt	Zeitraum		Abstand
		Start	Ende	
INDETERMINATE	α	α	α	α
EXPECTED	$E[\alpha]$	$E[\alpha]$	$E[\alpha]$	$E[\alpha]$
MIN	α_*	α_*	α^*	α_*
MAX	α^*	α^*	α_*	α^*

Beispiel: $t = (\text{Centurion}, \text{AB33}, [1, \text{März} \sim 31, \text{März} - 1, \text{Juni} \sim 30, \text{Juni}])$

$\text{Replace}(\text{INDETERMINATE}, t) = t$

$\text{Replace}(\text{EXPECTED}, t) = (\text{Centurion}, \text{AB33}, [15, \text{März} - 15, \text{Juni}])$

$\text{Replace}(\text{MIN}, t) = (\text{Centurion}, \text{AB33}, [31, \text{März} - 1, \text{Juni}])$

$\text{Replace}(\text{MAX}, t) = (\text{Centurion}, \text{AB33}, [1, \text{März} - 30, \text{Juni}])$

Beispiel und Zusammenfassung

- **Syntax**

```
SET DEFAULT PLAUSIBILITY 60
SELECT r.Lager, r.Lot_Num, p.Serial_Num, r.Wann
      FROM Received AS r WITH CREDIBILITY INDETERMINATE,
           In_Production AS p WITH CREDIBILITY INDETERMINATE
      WHERE p.Model = „Centurion“ AND r.Part = „Flügel Strebe“
           AND r.Wann OVERLAPS p.Während
```

- **Semantik**

$$\llbracket Q \rrbracket (d) = \{(r.Lager, r.Lot_Num, p.SerialNum, r.Wann) \mid$$
$$r \in \text{Replace}(\text{INDETERMINATE}, \text{Received})$$
$$\wedge p \in \text{Replace}(\text{INDETERMINATE}, \text{In_production})$$
$$\wedge p.Model = 'Centurion' \wedge r.Part = 'Flügel Strebe'$$
$$\wedge True \in \{\text{Before}(p.Während_{start}, r.Wann, 60) \cap$$
$$\text{Before}(r.Wann, p.Während_{ende}, 60)\}\}$$

- **Antworten**

Lager	Lot_Num	Serial_Num	Wann
Boeing	23	AB33	10,Mai ~ 29,Mai
Cessna	30	AB33	30,Mai ~ 18,Juni
Cessna	31	AB34	13,Juni ~ 02,Juli