Informationssysteme

Sommersemester 2016

Prof. Dr.-Ing. Sebastian Michel
TU Kaiserslautern

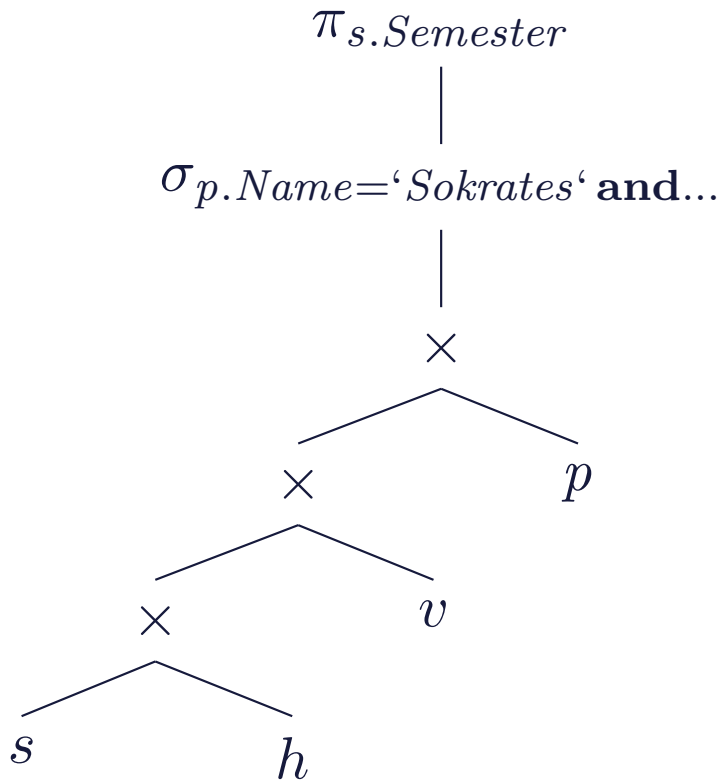
smichel@cs.uni-kl.de

Anfrageoptimierung

Beispiel: SQL Anfrage → Anfrageplan

select distinct s.Semester
from Studenten s, hoeren h
Vorlesungen v, Professoren p
where p.Name=‘Sokrates‘ **and**
v.gelesenVon = p.PersNr **and**
v.VorlNr = h.VorlNr **and**
h.MatrNr = s.MatrNr;

⇒



Regelbasierte Optimierung: Vorgehensweise

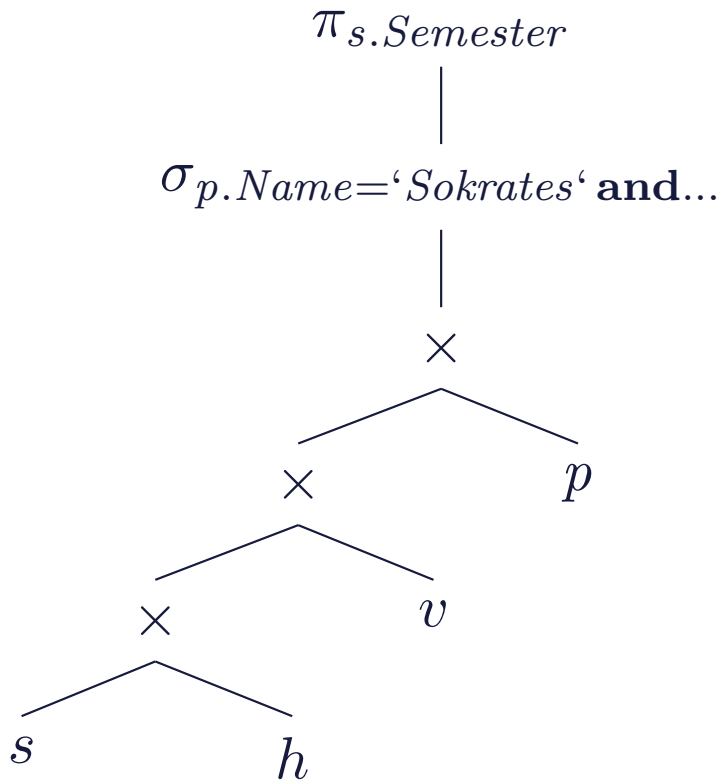
1. Aufbrechen von Selektionen
2. Verschieben der Selektionen soweit wie möglich nach unten im Operatorbaum (englisch: pushing selections)
3. Zusammenfassen von Selektionen und Kreuzprodukten zu Joins
4. Bestimmung der Reihenfolge der Joins in der Form, dass möglichst kleine Zwischenergebnisse entstehen
5. Unter Umständen Einfügen von Projektionen (keine Duplikateleminierung)
6. Verschieben der Projektionen soweit wie möglich nach unten im Operatorbaum

Hier kommen die Äquivalenzregeln der relationalen Algebra zum Einsatz!

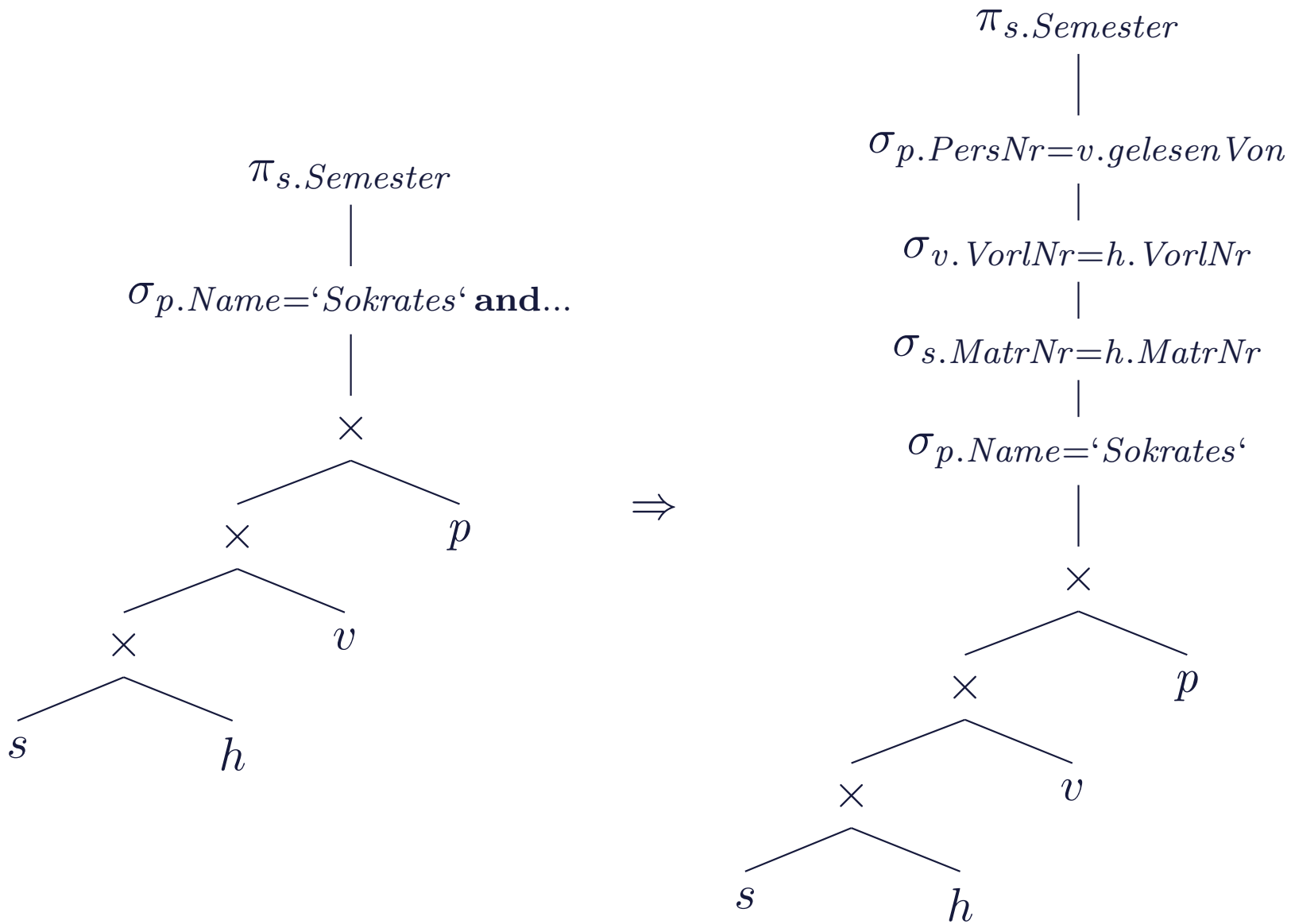
Beispiel

select distinct s.Semester
from Studenten s, hoeren h
Vorlesungen v, Professoren p
where p.Name=‘Sokrates‘ **and**
v.gelesenVon = p.PersNr **and**
v.VorlNr = h.VorlNr **and**
h.MatrNr = s.MatrNr;

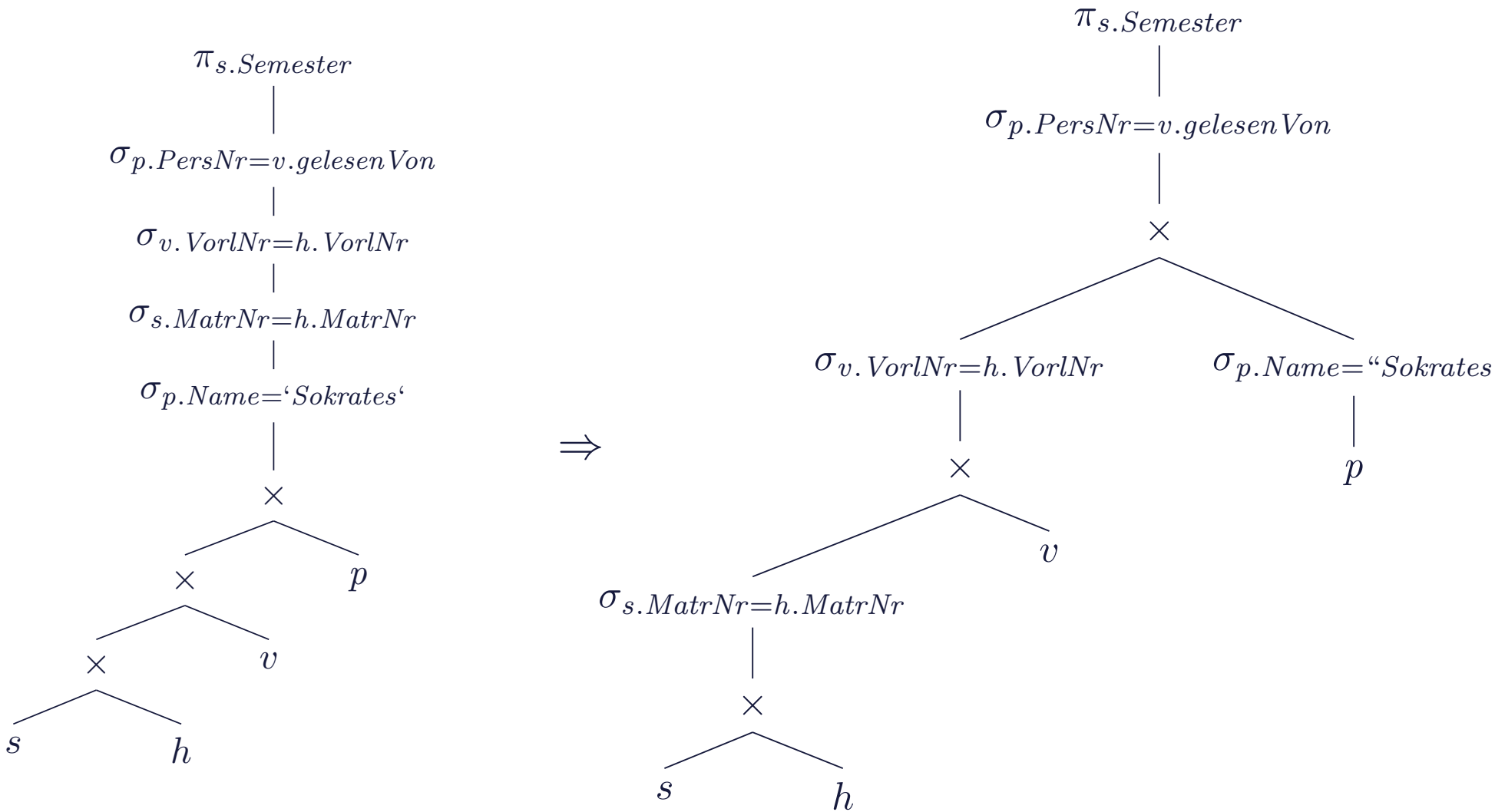
$\Rightarrow$



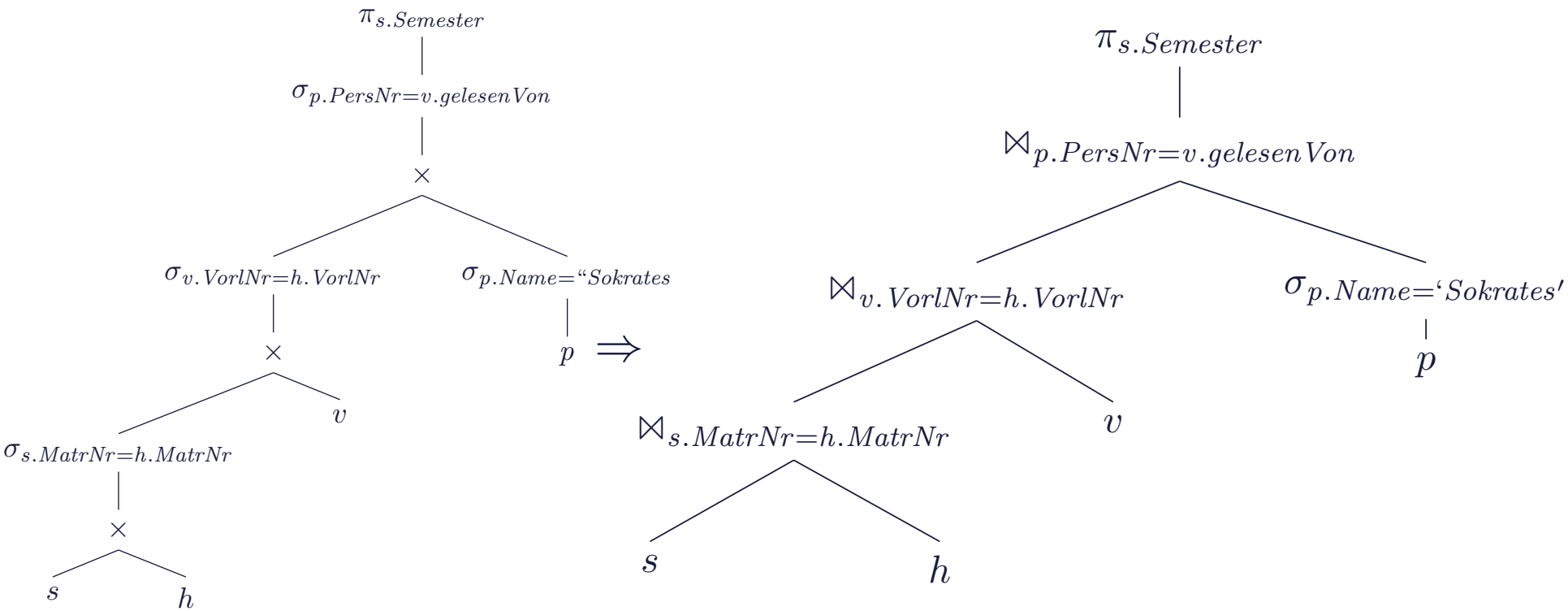
Aufspalten der Selektionsprädikate



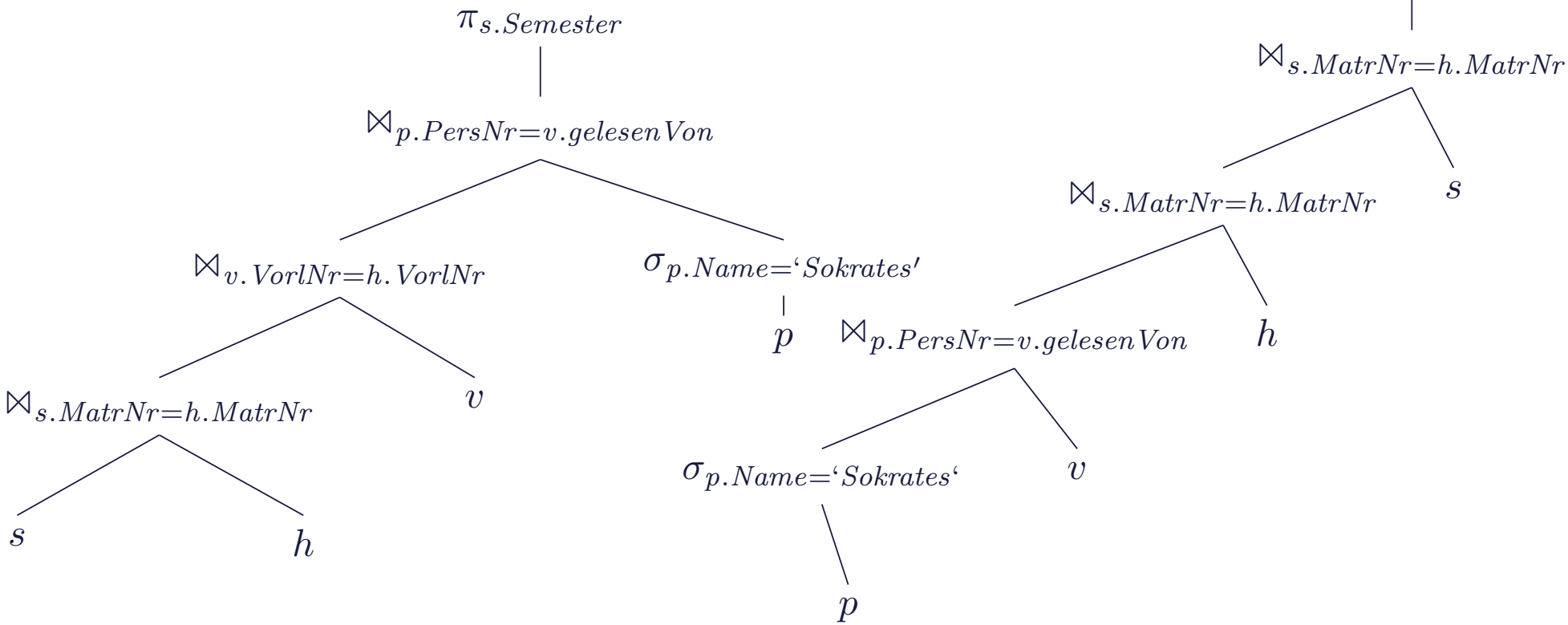
Verschieben der Selektionsprädikate



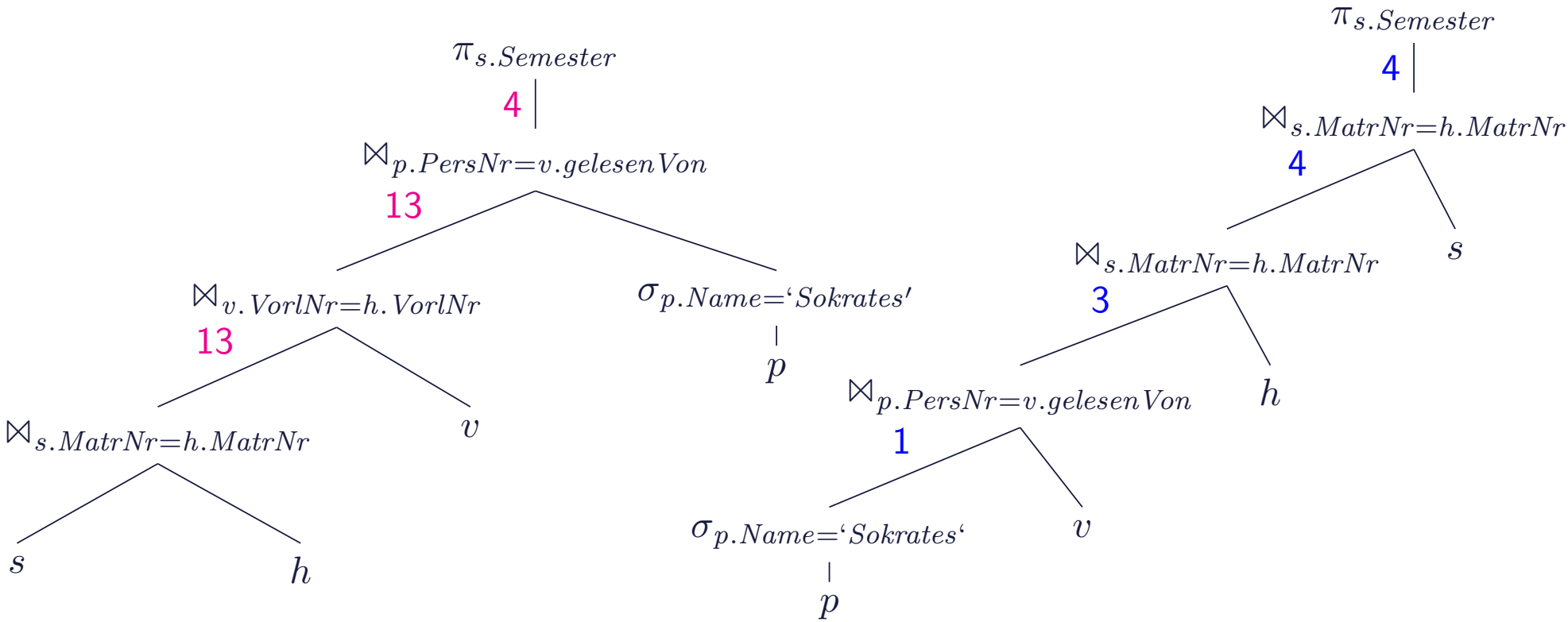
Zusammenfassung von Selektionen und Kreuzprodukten zu Joins



Optimierung der Joinreihenfolge



Effekt: Reduzierung der Zwischenergebnisse



Diese Zwischenkosten müssen natürlich geschätzt werden. Der Optimierer kann dann den günstigsten Plan bzgl. dieser geschätzten Kosten auswählen .

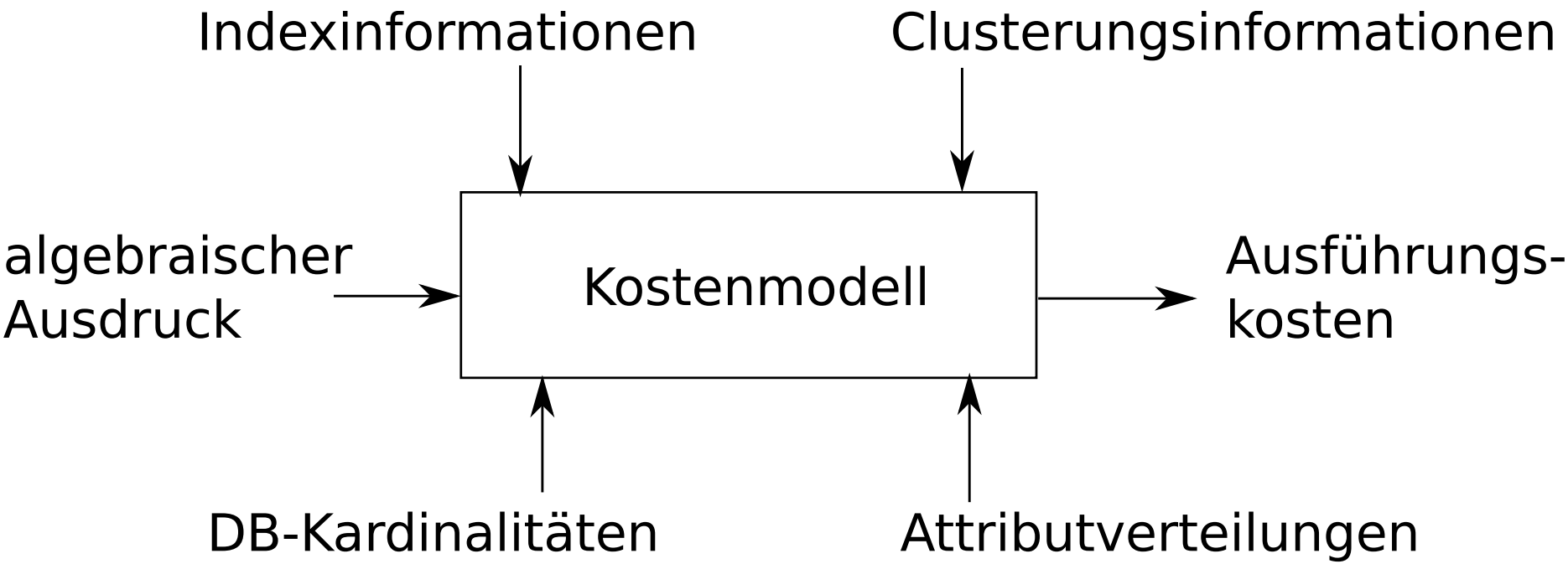
Kostenschätzung

(c) xkcd.com

Übersicht: Kostenbasierte Optimierung

- Generiere **alle** denkbaren Anfrageauswertungspläne:
  = Enumeration/Aufzählung möglicher Pläne
- Schätze deren Kosten ab:
  - Kostenmodell
  - Statistiken
  - Histogramme
  - Kalibrierung gemäß verwendetem Rechner
  - Abhängig vom verfügbaren Speicher
  - Aufwands-Kostenmodell
    - Durchsatz-maximierend
    - versus Antwortzeit-minimierend
- Behalte den Plan mit den geringsten geschätzten Kosten

Übersicht: Kostenmodelle



In dieser Vorlesung betrachten wir nur einfache Kosten in Form von Anzahl Zwischenergebnisse. Genauere Kosten hängen auch von Wahl der Implementierung der Operatoren, von verfügbaren Indexen, Performance der Hardware, etc. ab. $\rightarrow$ VL Datenbanksysteme.

Selektivitätsschätzung

Selektivitätsschätzung: Idee

- Gegeben: Anfrage und Relationen
- Wie viele Tupel sind als Ergebnis zu erwarten?
- Wie viele Tupel fallen als Zwischenergebnisse an?

Selektivitätsschätzung: Werkzeuge

- Kenntnisse/Statistiken über zugrunde liegende Daten
- Generische Annahmen über Selektivität benutzter Prädikate
- Schätzung der Selektivität von o.g. Operatoren (z.B., Join)

Selektivitätsschätzung: Grundlagen

- $T(R)$ ist die **Anzahl der Tupel** in Relation R
- $V(R,A)$ ist die **Anzahl der verschiedenen Attributsausprägungen** für Attribut A.
- Dementsprechend für mehrere Attribute: $V(R,[A_1, A_2, ..., A_n])$

Aufgabe der Kostenschätzung

- Gegeben eine Anfrage, wie groß ist das Ergebnis und wie viele Tupel fallen als Zwischenergebnisse an?
- z.B. wie viele Tupel sind in $\sigma_{A=13434}(R)$

Selektivität: Anteil der Tupel der Eingaberelation, die nicht herausgefiltert werden. Also: Größe der Ausgabe geteilt durch Größe der Eingabe.

Schätzungen für Selektion

Gegeben eine Selektion $S = \sigma_{A=c}(R)$.
Wie viele Tupel sind in S?
Schätzung:

$$T(S) = \frac{T(R)}{V(R,A)}$$

Gilt falls die Werte für A zufällig aus allen möglichen Werten gezogen wurden.

Ungleichheit

Was ist bei $S = \sigma_{A<c}(R)$?
Im Allgemeinen, ohne weitere Annahmen: $T(S) = T(R)/2$

Aber, Intuition: man wählt mit solchen Bedingungen meist weniger Tupel aus.
Besser: $T(S) = T(R)/3$

Schätzung für "not-equals"

Gegeben eine Selektion $S = \sigma_{A \neq c}(R)$.
Wie viele Tupel sind in S?

Einfache Schätzung

- "Mehr oder weniger" alle Tupel erfüllen die Bedingung (naja, bis auf ein paar, aber egal)
- Also: $T(S) = T(R)$

Leicht verbessert

- Die Tupel mit $A = c$ erfüllen die Bedingung nicht.
- Also: $T(S) = T(R)\dfrac{V(R,A) - 1}{V(R,A)}$

Was passiert mit Kaskaden von Selektionen?

Selektivität ist Produkt der einzelnen Selektivitäten!

Selektion mit ODER-Bedingungen

Gegeben:

$$S = \sigma_{C_1 \vee C_2}(R)$$

Annahme: die Bedingungen werden nie gemeinsam erfüllt

- Also: **entweder** gilt C_1 **oder** C_2
- Schätzung: Summe der beiden einzelnen Selektivitäten.
- Beobachtung: Überschätzt oft. Was kann dann passieren?

Besser: Annahme die Bedingungen sind unabhängig

- Annahme: m_1 Tupel erfüllen C_1 und m_2 Tupel erfüllen C_2
- Dann $T(S) = T(R) * (1 - (1 - \frac{m_1}{T(R)})(1 - \frac{m_2}{T(R)}))$

Selektion mit ODER-Bedingungen: Erläuterung

Wieso $T(S) = T(R) * (1 - (1 - \frac{m_1}{T(R)})(1 - \frac{m_2}{T(R)}))$?

$1 - \frac{m_1}{T(R)}$ ist der Anteil der Tupel die C_1 **nicht** erfüllen.

analog für C_2. Dann ist

$(1 - \frac{m_1}{T(R)})(1 - \frac{m_2}{T(R)})$ ist der Anteil der Tupel die C_1 **und** C_2 **nicht erfüllen**.

$(1 - ***)$ der Anteil der Tupel die C_1 **oder** C_2 erfüllen.

Klar: Multipliziert mit $T(R)$ ergibt $T(S)$

Andere Schätzer

Projektion

- Ändert die Kardinalität nicht (unter Bag Semantik)
- Sehr wohl aber die Größe der Tupel!
- Bei Mengensemantik: $T(S) = V(R,A)$ mit $S = \pi_A(R)$

Kartesisches Produkt

- Einfach: Produkt der beteiligten Kardinalitäten

Jetzt wird es **unklar was zu tun ist**:

Union

- Obere Schranke: Summe der beiden Kardinalitäten
- Untere Schranke: Größere der beiden Kardinalitäten
- Empfehlung aus Literatur: Irgendwas dazwischen, z.B. größere plus die halbe kleinere Kardinalität

Andere Schätzer

Schnitt

- Obere Schranke: Kleinste der beiden Kardinalitäten
- Untere Schranke: 0
- **Empfehlung aus Literatur:** Durchschnitt dieser beiden Werte.

Differenz $R - S$

- Obere Schranke: $T(R)$
- Untere Schranke: $T(R) - T(S)$
- **Empfehlung aus Literatur:** $T(R) - \frac{T(S)}{2}$

Schätzung für Joins: Natürlicher Join

Zwei Relationen: $R = (X,Y)$ und $S = (Y,Z)$
Annahme: Y ist ein einfaches Attribut, keine Menge von Attributen. X und Z dürfen Mengen sein.

Was können wir dann sagen? Leider nur sehr wenig, da z.B. folgende Fälle auftreten können:

- **Die beiden Relationen haben disjunkte Mengen für Y-Werte.** Also ist der Join leer, d.h. $T(R \bowtie S) = 0$.
- **Y ist Schlüssel von S und Fremdschlüssel in R.** Dann findet jedes Tupel in R einen Joinpartner in S, also $T(R \bowtie S) = T(R)$
- **Fast alle Tupel aus R und S haben den gleichen Wert für Y,** also $T(R \bowtie S) = T(R) * T(S)$

Schätzung für Joins: Natürlicher Join: Annahmen

Häufig auftretende Fälle. Weitere Annahmen:

- Es gibt Y-Werte $y_1, y_2, y_3, ...$
- Relationen benutzen diese Werte in dieser Reihenfolge.
- Dann: Falls $V(R,Y) \leq V(S,Y)$ so gilt auch, dass jeder Y-Wert in R auch ein Y-Wert in S ist.

Erhaltung der Wertemengen

- Falls A kein Join-Attribut ist, gilt

$$V(R \bowtie S, A) = V(R,A)$$

- D.h. Attribut A verliert durch den Join keine möglichen Werte.

Schätzung für Joins: Natürlicher Join

Gesucht: Größe des Joins $R(X,Y) \bowtie S(Y,Z)$

- Gegeben, zwei Tupel: $r \in R$ und $s \in S$
- Was ist die Wahrscheinlichkeit, dass $r.Y = s.Y$?
- Annahme: $V(R,Y) \geq V(S,Y)$, also gibt es den Y-Wert von s in R.
- Also: Wahrscheinlichkeit, dass $r.Y = s.Y$ ist $1/V(R,Y)$
- Umgekehrt: falls $V(R,Y) < V(S,Y)$ analog.
- **Insgesamt:** Übereinstimmung in Y mit Wahrscheinlichkeit $1/max(V(R,Y), V(S,Y))$

$$T(R \bowtie S) = \frac{T(R) * T(S)}{max(V(R,Y), V(S,Y))}$$

"Essentially, all models are wrong, but some are useful."
(George E. P. Box)