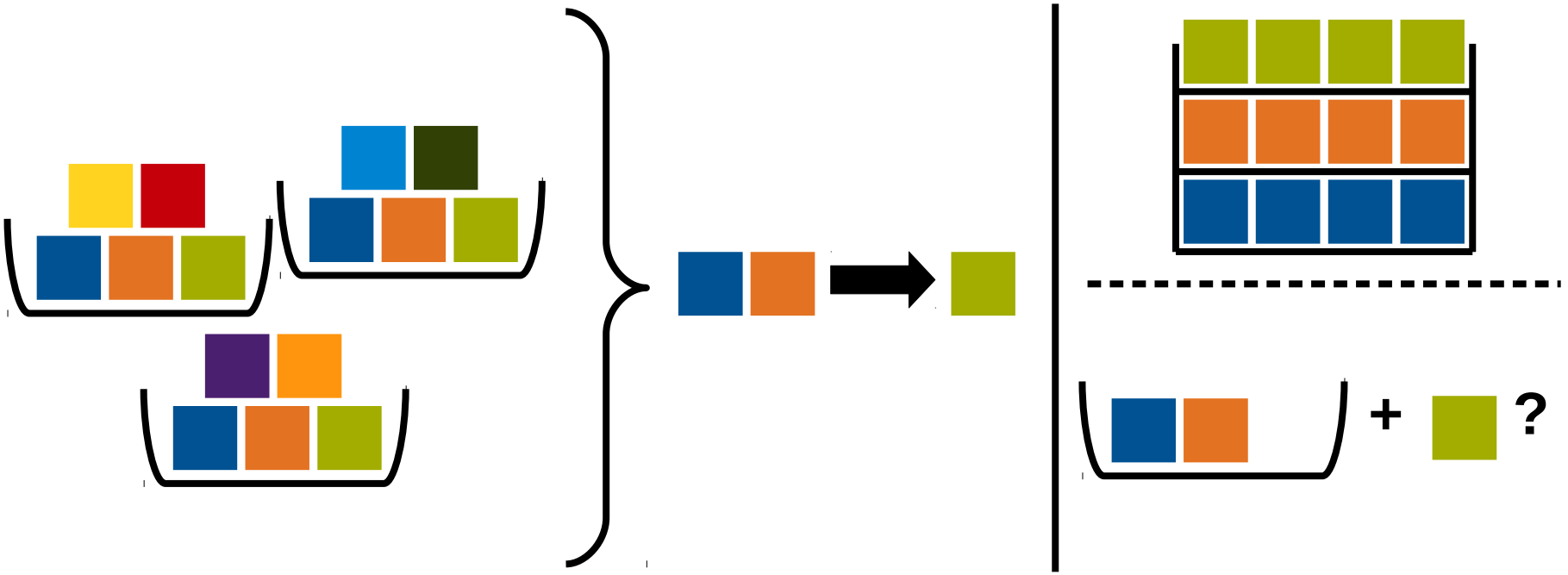
Einsatz und Realisierung von Datenbanksystemen Übungsblatt 08

Maximilian E. Schüle
schuele@in.tum.de
02.11.060

Garching, 26. Juni 2017

Assoziationsregeln

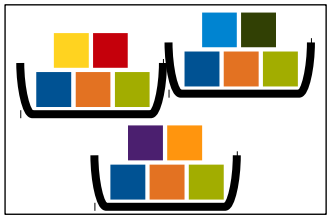


Warenkörbe **Assoziationsregeln** **Produktvorschläge**

Definitionen

Warenkorbdaten $\mathcal{D} = \{T_1, ..., T_n\}$ $T_i \subseteq I = \{i_1, ..., i_m\}$

Daten über miteinander gekaufte Elemente pro Einkauf oder pro Kunde



Frequentitemset $\mathcal{L} := \bigcup_{k=1}^{\infty} \mathcal{L}_k = \{L | L \subseteq I \wedge s(L) \geq s_0\}$

Menge von Elementen, genannt Items, die in einer bestimmten Anzahl von Warenkörben gemeinsam auftreten

Assoziationsregel $X \Rightarrow Y \text{ mit } X \subset T, Y \subset T, X \cap Y = \emptyset$

Korrelation zwischen Elementen
Starke Regel, wenn Konfidenz über Vertrauenswert

Kennzahlen

Support $\sigma(X) := |\{T | T \in \mathcal{D} \wedge X \subseteq T\}|$ $s(X) := \frac{\sigma(X)}{|\mathcal{D}|}$

Relative Häufigkeit einer Item-Menge
Minimale Support: definierter Schwellenwert

Konfidenz $k(X \Rightarrow Y) := P(Y|X) = \frac{\sigma(X \cup Y)}{\sigma(X)} = \frac{s(X \cup Y)}{s(X)}$

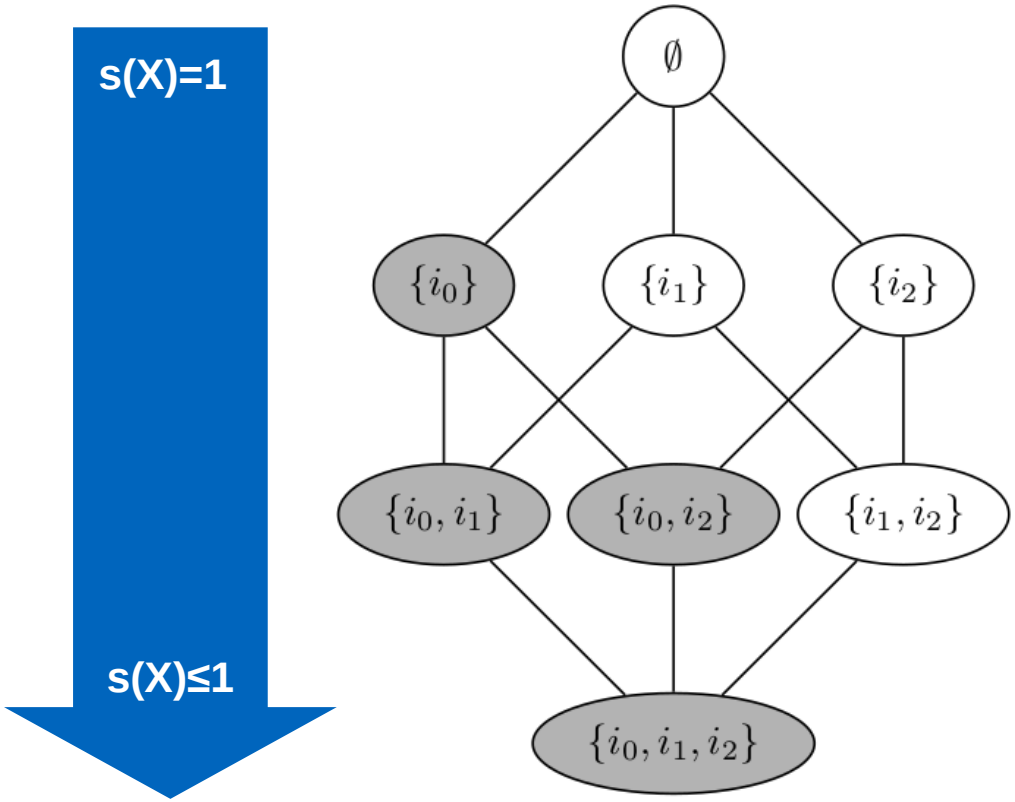
Häufigkeit der Transaktionen die Y enthalten, wenn diese bereits X enthalten

Apriori-Algorithmus – Apriori-Prinzip (1)

Apriori-Prinzip (a priori: lat. *ab*: von; lat. *prior*: vorherig)

Support einer Item-Menge maximal so groß wie der Support ihrer Teilmengen:

$$\forall X, Y \subseteq I.(X \subseteq Y) \Rightarrow s(X) \geq s(Y)$$



Apriori-Algorithmus – Frequentitemsets

Warenkörbe

TID	Items
1	{Stift, Lineal}
2	{Stift, Lineal, Papier}
3	{Stift, Lineal}
4	{Lineal, Papier}

Iteration	Item-Menge	Support
1	{Stift}	3
1	{Lineal}	4
1	{Papier}	2
2	{Stift, Lineal}	3
2	*{Stift, Papier}*	1
2	{Lineal, Papier}	2
3	~~{Stift, Lineal, Papier}~~	

$\mathcal{L}_1 = \{\text{large 1 itemsets}\}$;
for $(k = 2; \mathcal{L}_{k-1} \neq \emptyset; k++)$ **do**
  $\mathcal{C}_k = \text{apriori-gen}(\mathcal{L}_{k-1})$;
  forall $t \in \mathcal{D}$ **do**
    $\mathcal{C}_t = \text{subset}(\mathcal{C}_k, t)$;
    forall $c \in \mathcal{C}_t$ **do**
      $c.\text{count}++$;
    end
  end
  $\mathcal{L}_k = \{c \in \mathcal{C}_k | c.\text{count} \geq \sigma_0\}$;
end
return $\bigcup_k \mathcal{L}_k$;

Hausaufgabe 01 – Apriori

`Siehe Tafel`

VerkaufsTransaktionen	
TransID	Produkt
111	Drucker
111	Papier
111	PC
111	Toner
222	PC
222	Scanner
333	Drucker
333	Papier
333	Toner
444	Drucker
444	PC
555	Drucker
555	Papier
555	PC
555	Scanner
555	Toner

Zwischenergebnisse	
FI-Kandidat	Anzahl
{Drucker}	4
{Papier}	3
{PC}	4
{Scanner}	2
{Toner}	3
{Drucker, Papier}	3
{Drucker, PC}	3
{*Drucker, Scanner*}	
{Drucker, Toner}	3
{Papier, PC}	2
{*Papier, Scanner*}	
{Papier, Toner}	3
{*PC, Scanner*}	
{PC, Toner}	2
{*Scanner, Toner*}	

Hausaufgabe 02 – Threshold und NRA

`Siehe Tafel` Top-1-Berechnung (zur Bestimmung des günstigsten Wohnorts) für eine junge Familie mit zwei Kindern

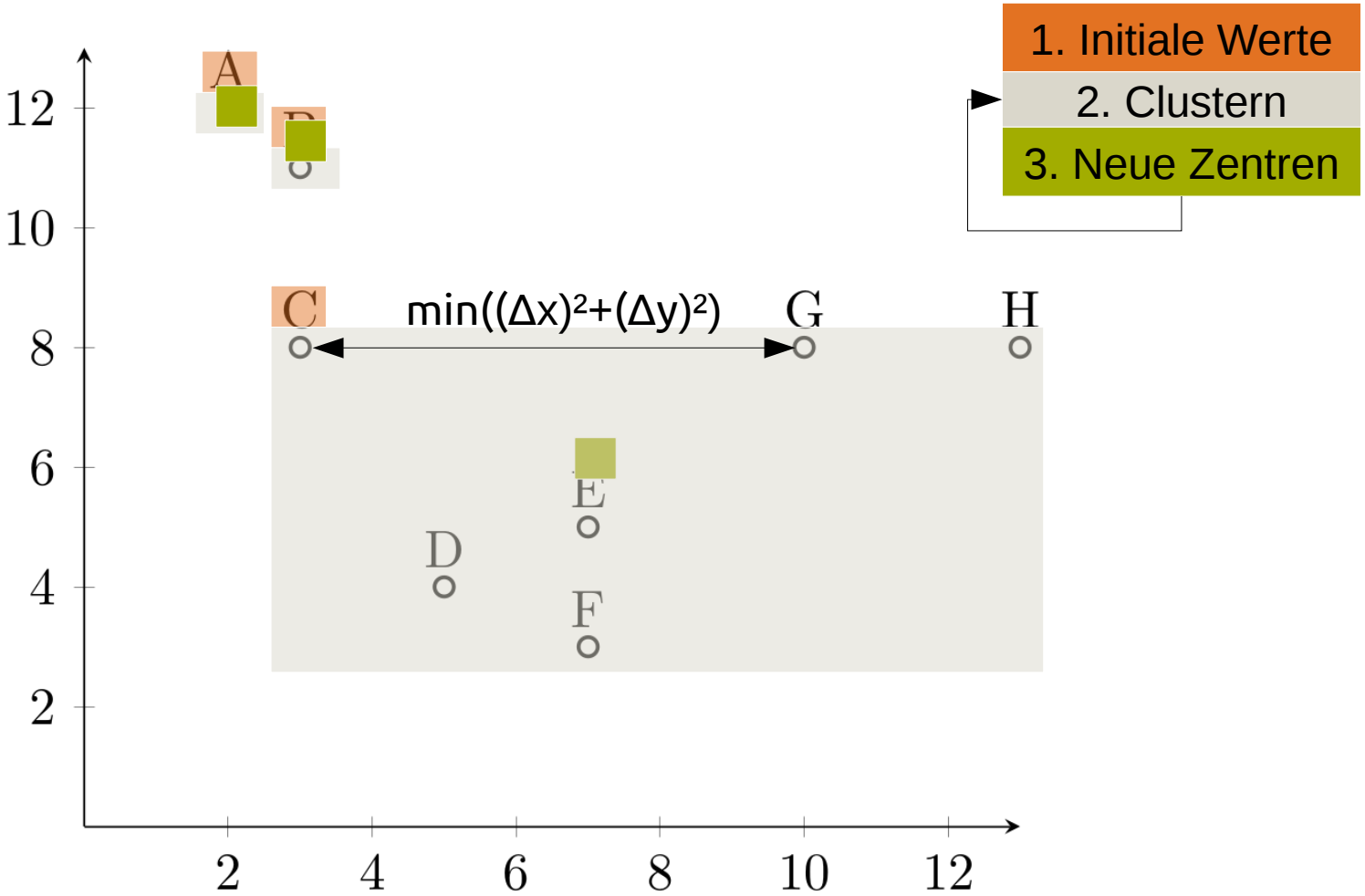
Mietspiegel	
Ort	Miete
Garching	800
Ismaning	900
Unterföhring	1000
Nymphenburg	1500
Bogenhausen	1600
Grünwald	1700

Kindergarten	
Ort	Beitrag
Grünwald	-100
Unterföhring	0
Bogenhausen	100
Ismaning	200
Garching	250
Nymphenburg	300

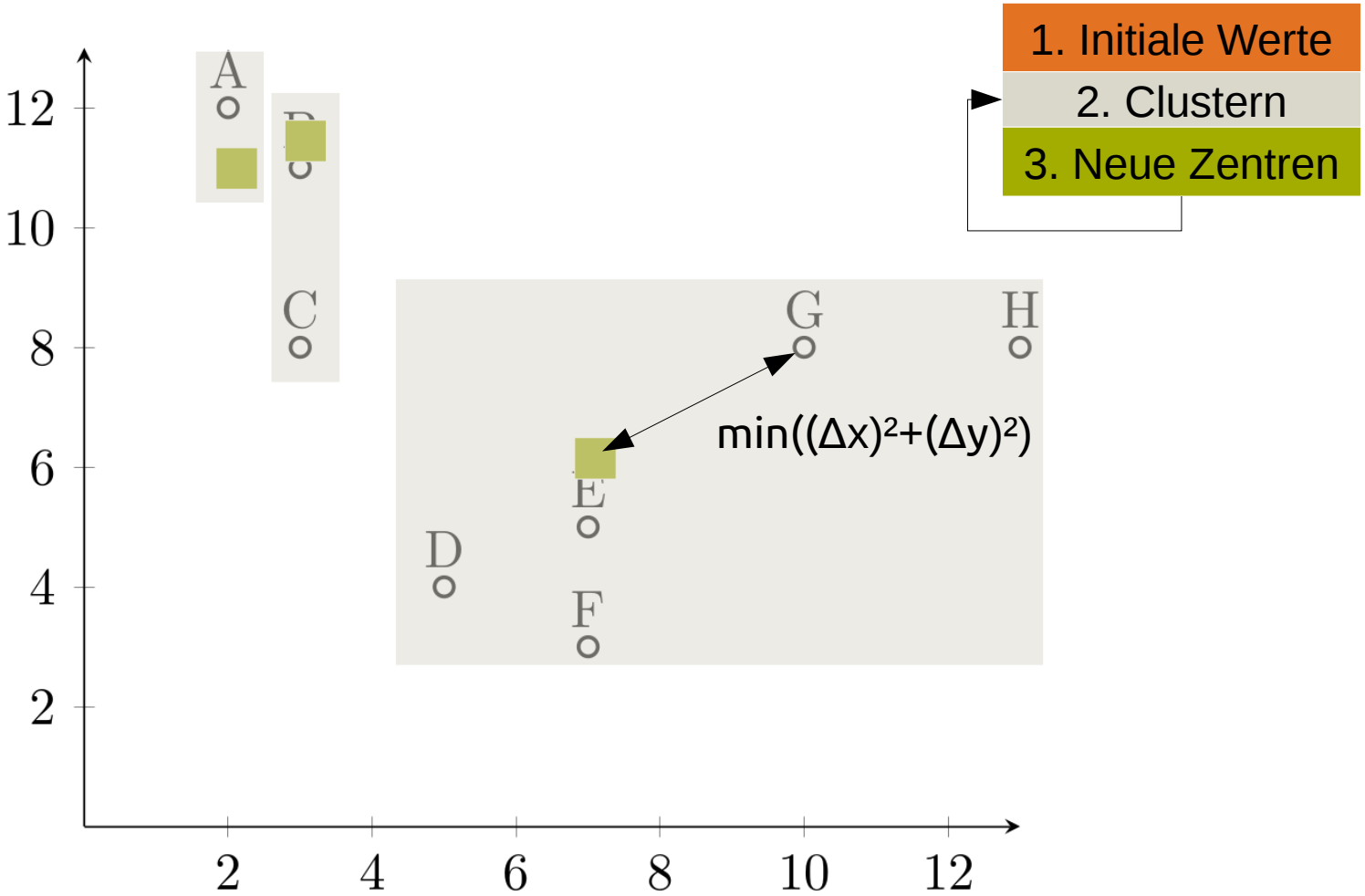
Hausaufgabe 03 – Skyline

```
SQL mit Skyline:

select MatrNr
from Klausur k
skyline of k.Vorbereitungszeit min, k.Note min


SQL ohne Skyline:

select MatrNr
from Klausur k
where not exists (
    Select *
    from klausur dom
    Where -- wird nicht dominiert
        dom.Vorbereitungszeit <= k.Vorbereitungszeit and
        dom.Note <= k.Note and ( -- und sonst mind. genauso
        dom.Vorbereitungszeit < k.Vorbereitungszeit or
        dom.Note < k.Note) -- von einem eines besser
)
```

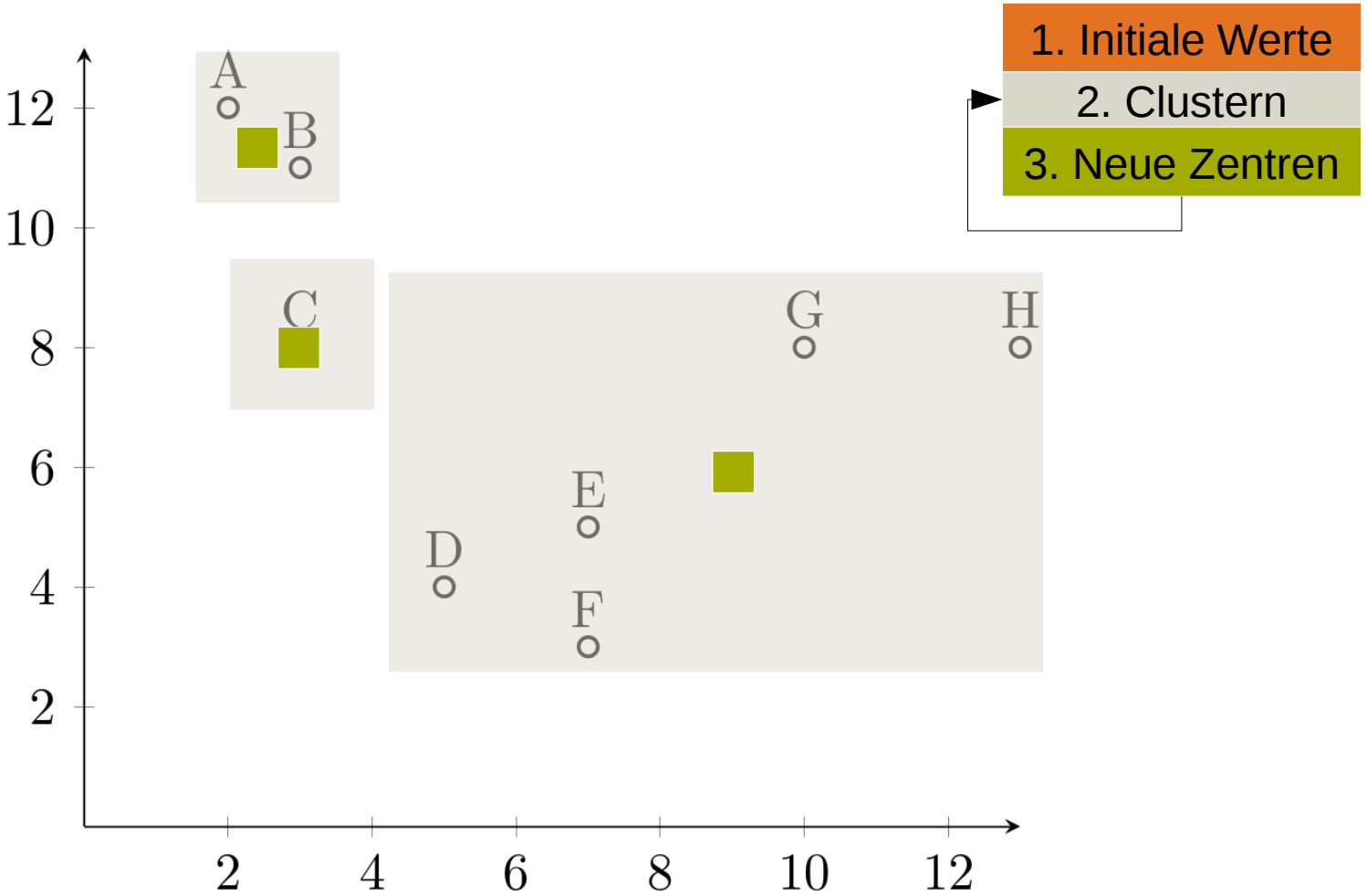
Hausaufgabe 04 – K-Means



Hausaufgabe 04 – K-Means



Hausaufgabe 04 – K-Means



Hausaufgabe 04 – K-Means

```
with points(id,x,y) as (
        VALUES ('A', 2, 12), ('B', 3, 11), ('C', 3,8), ('D', 5,4),
        ('E',7,5),('F',7,3),('G',10,8),('H',13,8)
),
clusters_0(cid,x,y) as (
  VALUES ('1', 2, 12), ('2', 3, 11), ('3', 3,8)
),
clusters_1(cid, x,y, count) as (
  select cid, avg(px), avg(py), count(*) from (
    select cid, p.x as px, p.y as py, rank() OVER (
      partition by p.id
      order by (p.x-c.x)*(p.x-c.x)+(p.y-c.y)*(p.y-c.y) asc,
        (c.x*c.x+c.y*c.y) asc)
    from points p, clusters_0 c
  ) x
  where x.rank=1
  group by cid
),
clusters_2 (cid, x,y, count) as (
  select cid, avg(px), avg(py), count(*) from (
    select cid, p.x as px, p.y as py, rank() OVER (
      partition by p.id
      order by (p.x-c.x)*(p.x-c.x)+(p.y-c.y)*(p.y-c.y) asc,
        (c.x*c.x+c.y*c.y) asc)
    from points p, clusters_1 c
  ) x
  where x.rank=1
  group by cid
),
clusters_3 (cid, x,y, count) as (
  select cid, avg(px), avg(py), count(*) from (
    select cid, p.x as px, p.y as py, rank() OVER (
      partition by p.id
      order by (p.x-c.x)*(p.x-c.x)+(p.y-c.y)*(p.y-c.y) asc,
        (c.x*c.x+c.y*c.y) asc)
    from points p, clusters_2 c
  ) x
  where x.rank=1
  group by cid
)
select * from clusters_3
```

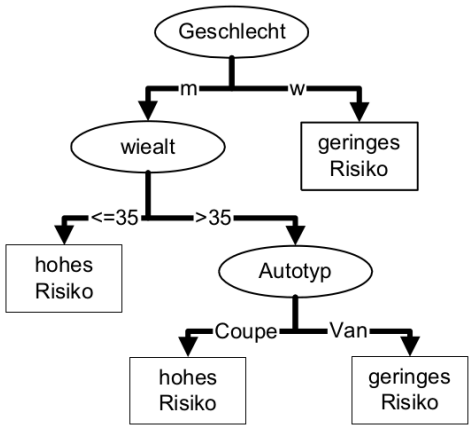
```
with points(id,x,y) as (
        VALUES ('A',2,12), ('B',3,11),
        ('C',3,8), ('D',5,4),('E',7,5),
        ('F',7,3),('G',10,8),('H',13,8)
)

select *
from kmeans(
    (table points),
    λ(a,b) sqrt((a.x-b.x)^2+(a.y-b.y)^2),
    3
)
```

Hausaufgabe 05 – Klassifikationsbaum

1) Vorhersageattribut
2) Partitionieren

```
Select wiealt, Schäden, count(*)
From Schadenshöhe
Group by wiealt, Schäden

Select Autotyp, Schäden, count(*)
From Schadenshöhe
Group by Autotyp, Schäden

Select Geschlecht, Schäden, count(*)
From Schadenshöhe
Group by Geschlecht, Schäden
```

Schadenshöhe			
wiealt	Geschlecht	Autotyp	Schäden
45	w	Van	gering
18	w	Coupé	gering
22	w	Van	gering
38	w	Coupé	gering
19	m	Coupé	hoch
24	m	Van	hoch
40	m	Coupé	hoch
40	m	Van	gering
⋮	⋮	⋮	⋮



Hausaufgabe 06 – Threshold

`Preis − (100 ∗ P S) + 24 ∗ Unnterhalt`

Auto	Preis	PS
Seat Leon	25000€	200
Audi A1	17000€	96
Citroen DS 4	20679€	100
Mini One	16500€	75
Mercedes C-Klasse	35000€	160
Porsche Cayenne	80100€	420

Auto	Unterhalt p. Monat
Seat Leon	215€
Audi A1	220€
Citroen DS 4	225€
Mini One	262€
Mercedes C-Klasse	290€
Porsche Cayenne	430€

Zw. Ergebnis: Phase 4

Auto	Score
Seat Leon	10160
Audi A1	12680
Mini One	15288
Citroen DS 4	16079
Threshold	16967

Hausaufgabe 06 – NRA

`Preis - (100 * PS) + 24 * Unterhalt`

Auto	Preis	PS
Seat Leon	25000€	200
Audi A1	17000€	96
Citroen DS 4	20679€	100
Mini One	16500€	75
Mercedes C-Klasse	35000€	160
Porsche Cayenne	80100€	420

Auto	Unterhalt p. Monat
Seat Leon	215€
Audi A1	220€
Citroen DS 4	225€
Mini One	262€
Mercedes C-Klasse	290€
Porsche Cayenne	430€

NRA: Phase 4

Auto	Score
Seat Leon	10160
Audi A1	12680
Mini One	15288
Citroen DS 4	16079

http://db.in.tum.de/teaching/ss17/impldb/

Maximilian E. Schüle
schuele@in.tum.de
02.11.060

Viel Spaß!