Skylines III

Christian Nieke

Übersicht

- **Kardinalitäten von Skylines**
  - Wie groß ist eine Skyline ?
- **Approximate Dominating Representatives**
  - Annäherung von Skylines

Autokauf: Wie Frage ich eine DB an ?

- **SQL-Anfragen:**
  - Genau die Tupel, die ich will.
  - Aber: Welche will ich genau ?
    - „Ein türkiser Ford mit 113 PS für 6.407€“
- **Top-K-Anfragen:**
  - Die besten K Tupel
  - Aber: Welches Kriterium ?
    - „Preis ist mir 3 mal so wichtig wie PS“
- **Skyline:**
  - Alle „guten“ Tupel
  - Probleme ?

Kardinalität

- **Wie viele Tupel liefert die Skyline ?**
  - Das Ergebnis ist **unsortiert** !
- **30 Tupel**
  - Schöne Übersicht
- **3.000 Tupel**
  - Niemand sieht Ergebnis durch
  - Teure Anfrage umsonst

Ich muss <u>vorher</u> wissen, wie viele Tupel zurück kommen

Basismodell - Definitionen

- **Skyline:**
  - Tupel A ist in Skyline ↔ kein Tupel kann A dominieren
  - Kein Tupel ist in mind. einer Dimension besser und sonst gleich gut
- **Min-Sortierung:**
  - Besser = der Wert ist niedriger
- **Sparseness:**
  - Werte in Dimensionen sind einzigartig
- **Unabhängigkeit:**
  - Die Dimensionen sind unkorreliert

Basismodell - Definitionen

- **Gleichverteilung**
  - Die Werte in den Dimensionen sind gleichverteilt
- **$\hat{s}_{d,n}$ = Durchschnittliche Größe der Skyline**
  - n Tupel
  - d Dimensionen
- **Obergrenze:**
  - $O((\ln n)^{d-1})$
- **Tatsächliche Werte ?**

Basismodell

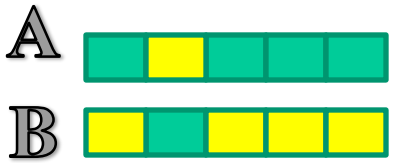
- **Erwartungswert:**
  - $\hat{s}_{d,n} = 1/n * \hat{s}_{d-1,n} + \hat{s}_{d,n-1}$
- **Herleitung:**
  - Ein Tupel hat in d den schlechtesten Wert
  - $1/n * \hat{s}_{d-1,n}$ = Wahrscheinlichkeit, in anderer Dimension Skyline zu sein
  - $\hat{s}_{d,n-1}$ = Anzahl Skyline-Tupel ohne dieses Tupel
- **Abschätzung:**
  - $\Theta\left(\frac{\ln n^{d-1}}{(d-1)!}\right)$

Erweiterung 1 - Dichte

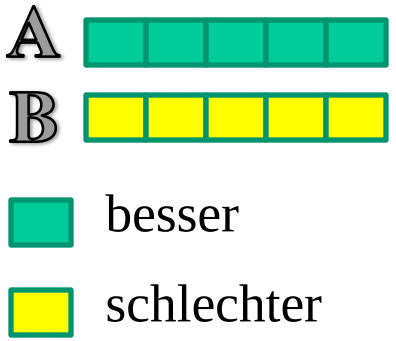
- **Dichte**
  - Tupel dürfen gleiche Werte in den Dimensionen haben
  - Deutlich realistischer
- **Auswirkungen ?**
  - Skyline größer oder kleiner ?
  - Eindeutiger Trend ?
  - Grenzen ?

Erweiterung 1 - Dichte

- **Vergleiche 2 Tupel im <u>Basismodell</u>**
- **Folgende Möglichkeiten:**
  - Tupel sind **unvergleichbar**:
    - jedes hat mind. 1 Attribut besser
    - **=> 2 Skyline-Tupel**



  - Tupel sind **vergleichbar**:
    - Ein Tupel dominiert anderes
    - **=> 1 Skyline-Tupel**

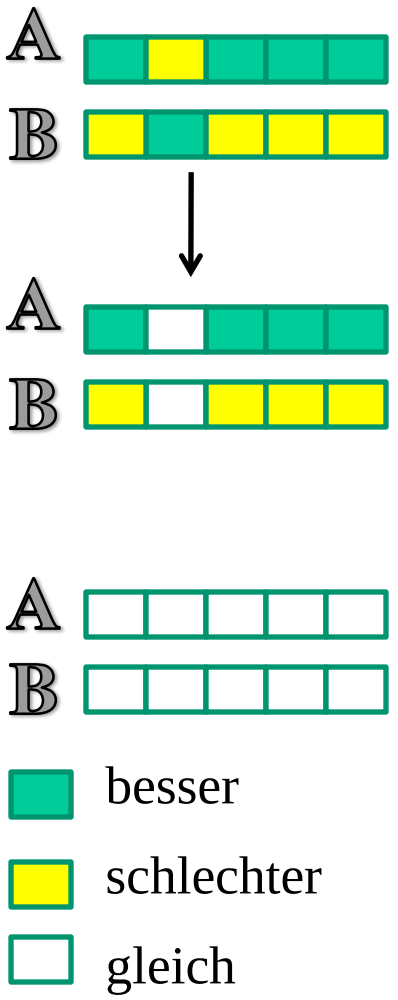


Erweiterung 1 - Dichte

- **Erweiterung auf Dichte**
- **Idee:**
  - Werte →Wertebereiche (Partitionen)
  - Statt: 75 PS, 77 PS
    Nun: 70-80 PS
  - Auswirkungen ?

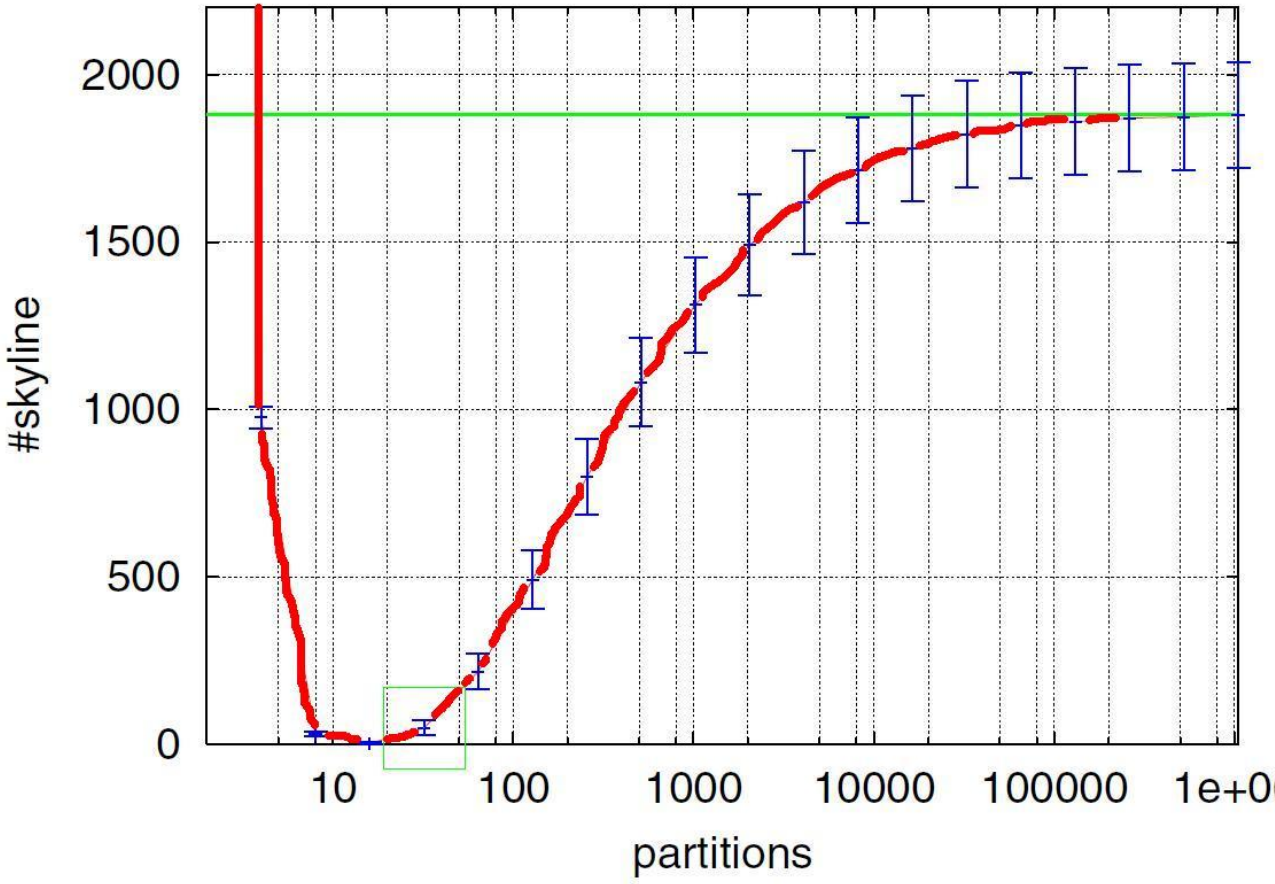
Erweiterung 1 - Dichte

- **Folgende neue Möglichkeiten:**
  - Tupel werden vergleichbar:
    - **=> 1 Skyline-Tupel**
    - **=> Skyline wird kleiner**
  - Tupel sind **Duplikate**:
    - Beide sind Skyline
    - **=> 2 Skyline-Tupel**
    - **=> Skyline wird größer**



Erweiterung 1 - Dichte

- **Welcher Effekt überwiegt ?**



Grün = $\hat{s}_{d,n}$
Rot = Skyline
Blau = Standardabweichung

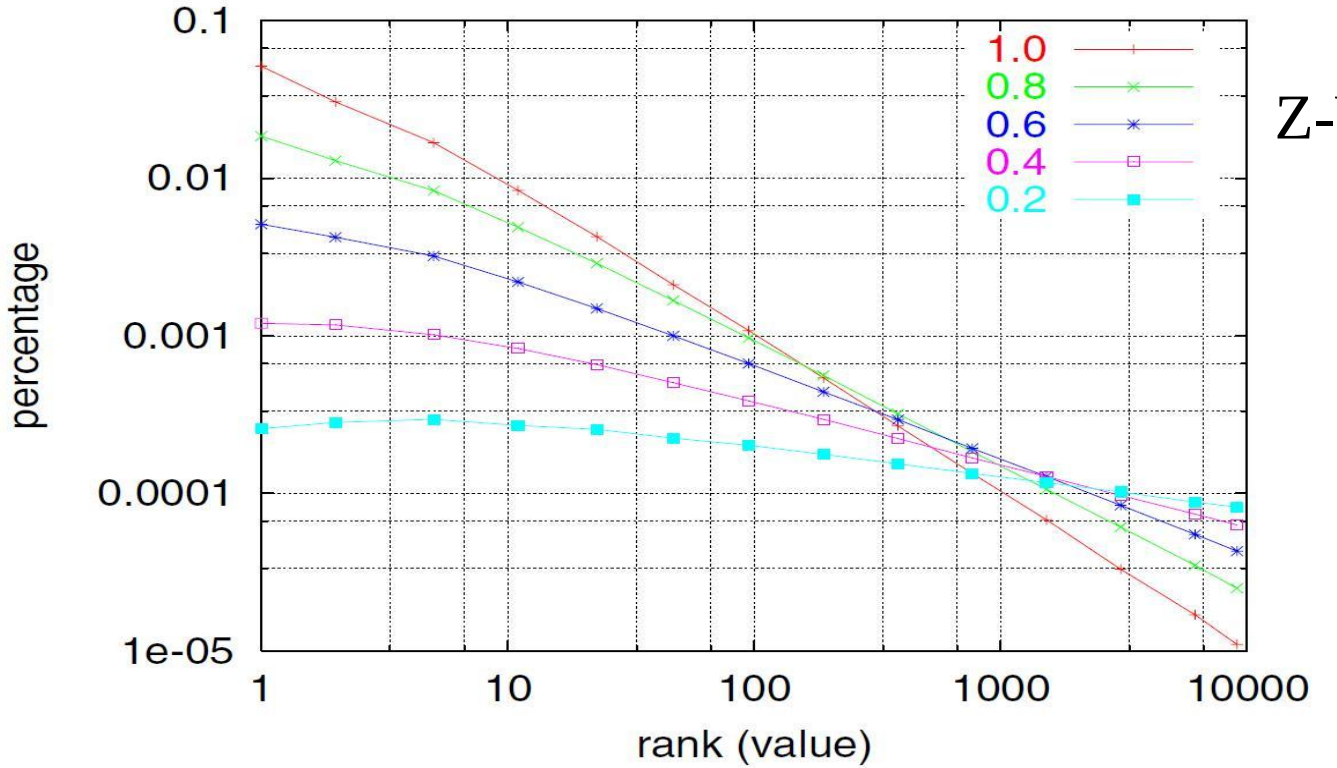
d = 5
n = 10^6

Erweiterung 2 – Zipf Verteilung

- **Werte nicht sparse→ Verteilung ist wichtig**
- **Welche Werte kommen häufig vor ?**
  - Viele Gute ?
  - Viele Schlechte ?
  - Viele Mittlere ?
- **Annahme:**
  - Werte haben Zipf-Verteilung

Erweiterung 2 – Zipf Verteilung

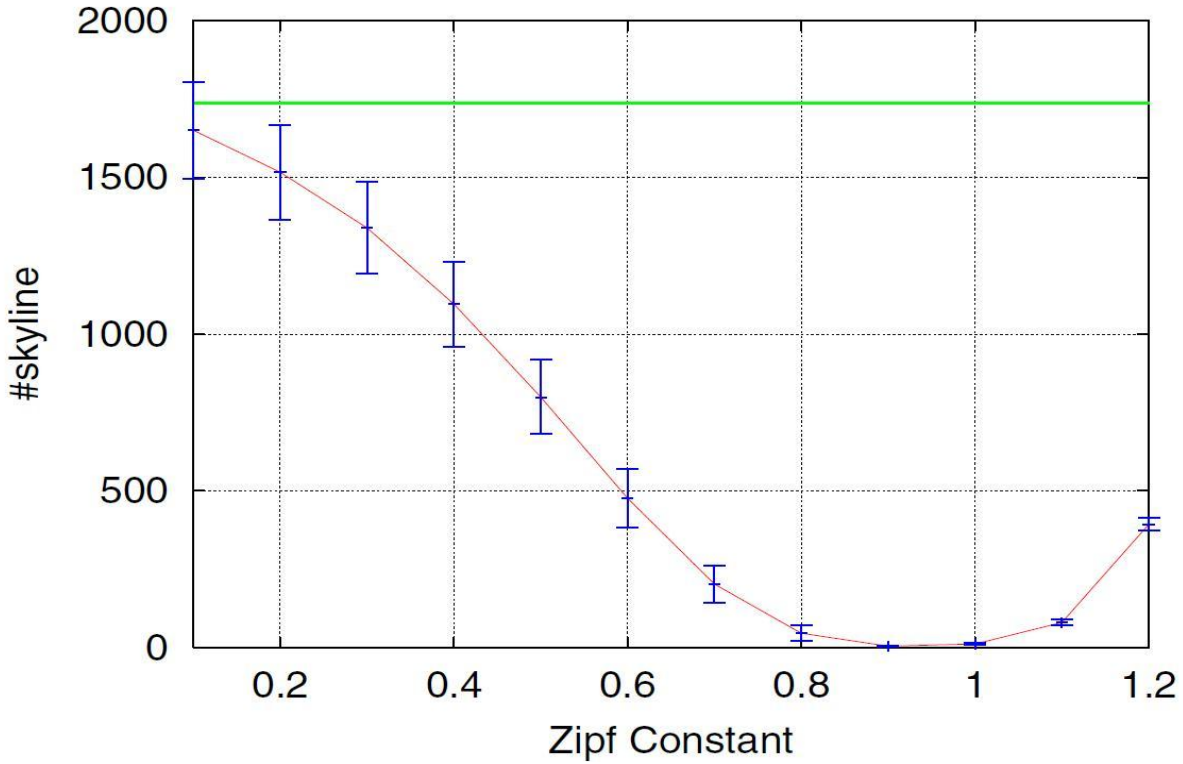
- **Zipf Verteilung:**
  - Wahrscheinlichkeit von Wert i ~ $1/i^z$
  - Je höher z, desto mehr „Zipf-ähnlich“



Z-Werte

Erweiterung 2 – Zipf Verteilung

- **Auswirkungen:**



Grün = $\hat{s}_{d,n}$
Rot = Skyline
Blau = Standardabweichung

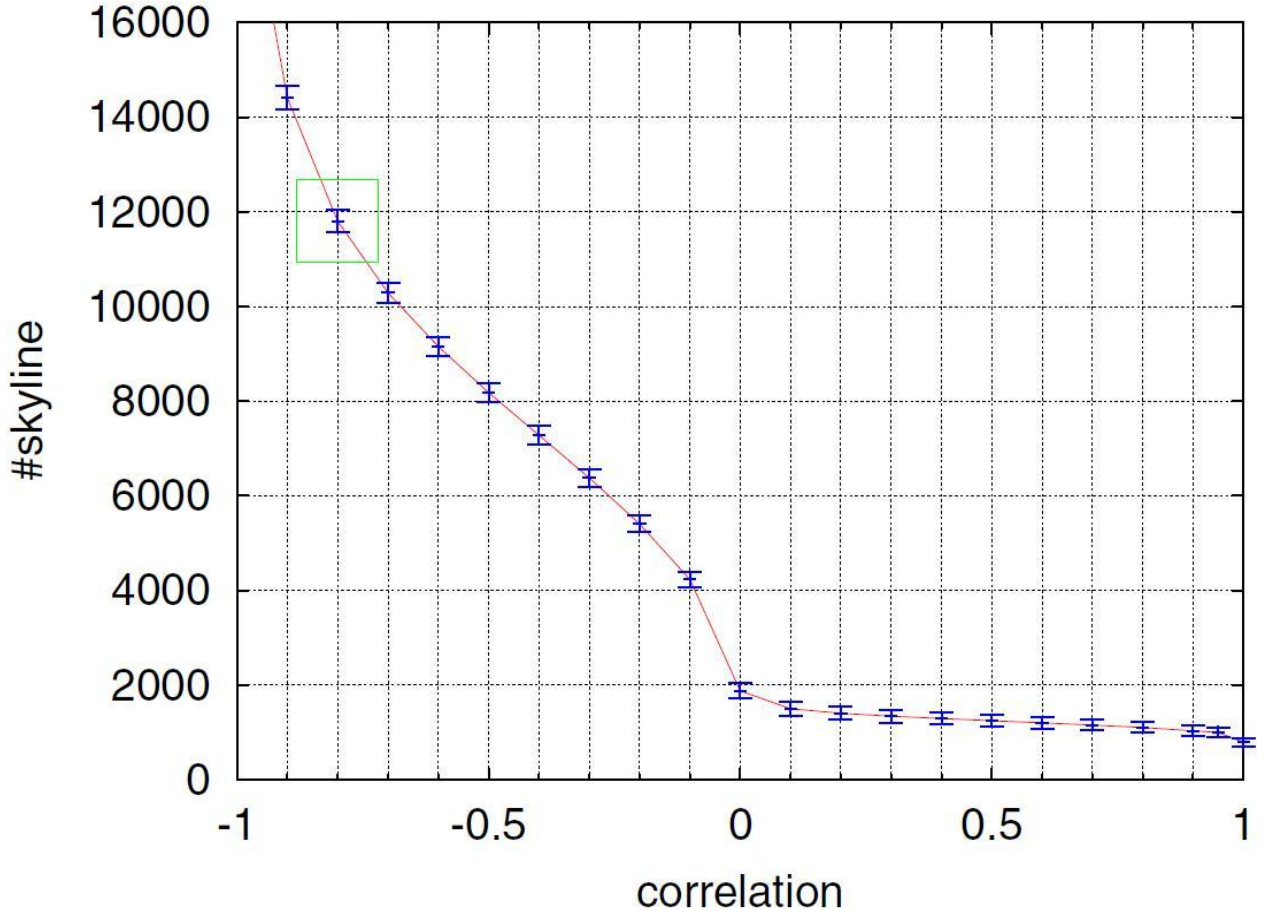
$d = 5$
$n = 10^6$
$p = 10^4$

Erweiterung 2 – Zipf Verteilung

- **Was ist bei vielen schlechten Werten ?**
  - Tupel mit guten Werten werden sparse
  - Annäherung an $\hat{s}_{d,n}$
- **Kritik**
  - Normalverteilung erscheint mir wahrscheinlicher
  - Vermutlich ähnliche Effekte wie Zipf mit vielen schlechten Werten

Erweiterung 3 – (Anti)-Korrelation

- **Auswirkungen:**



Rot = Skyline
Blau = Standardabweichung

d = 5
$n = 10^6$

Kardinalitäten: Zusammenfassung

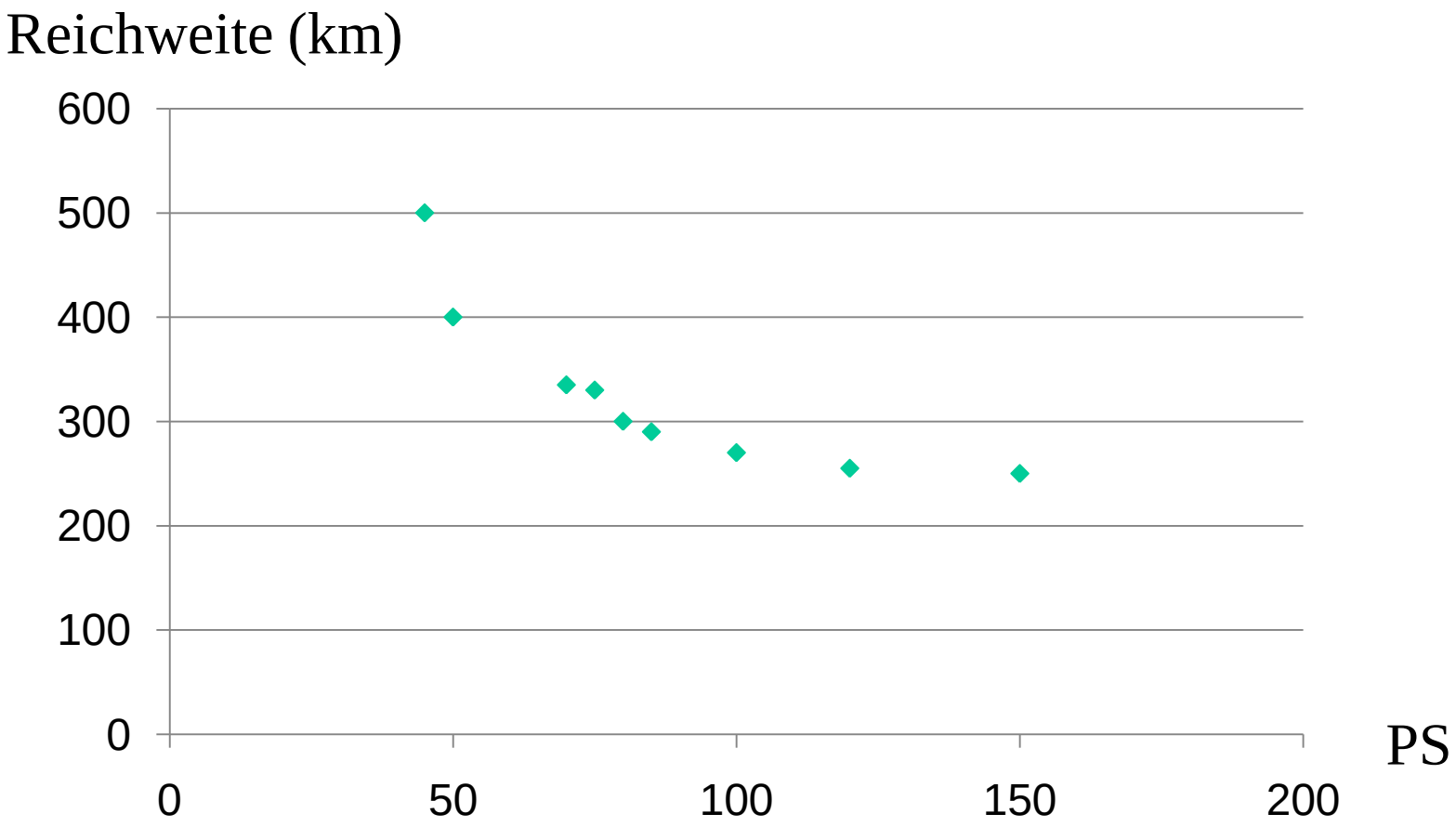
- **Erwartungswert:**
  - $\hat{s}_{d,n} = \Theta\left(\frac{\ln n^{d-1}}{(d-1)!}\right)$
- **Effekte:**
  - Skyline wird kleiner:
    - Dichte
    - Zipf-Verteilung (viele Gute)
    - Korrelation
  - Skyline wird größer:
    - Anti-Korrelation
    - Duplikate

Und nun ?

- **Skyline ist zu groß !**
  - Was tue ich ?
- **Tupel entfernen**
  - Eine Annäherung der Skyline

- **Aber welche Tupel nehme ich ?**
  - Random-Sample ?
  - Die ersten k ? (keine Sortierung)

Auswahl:



Eine gute Auswahl

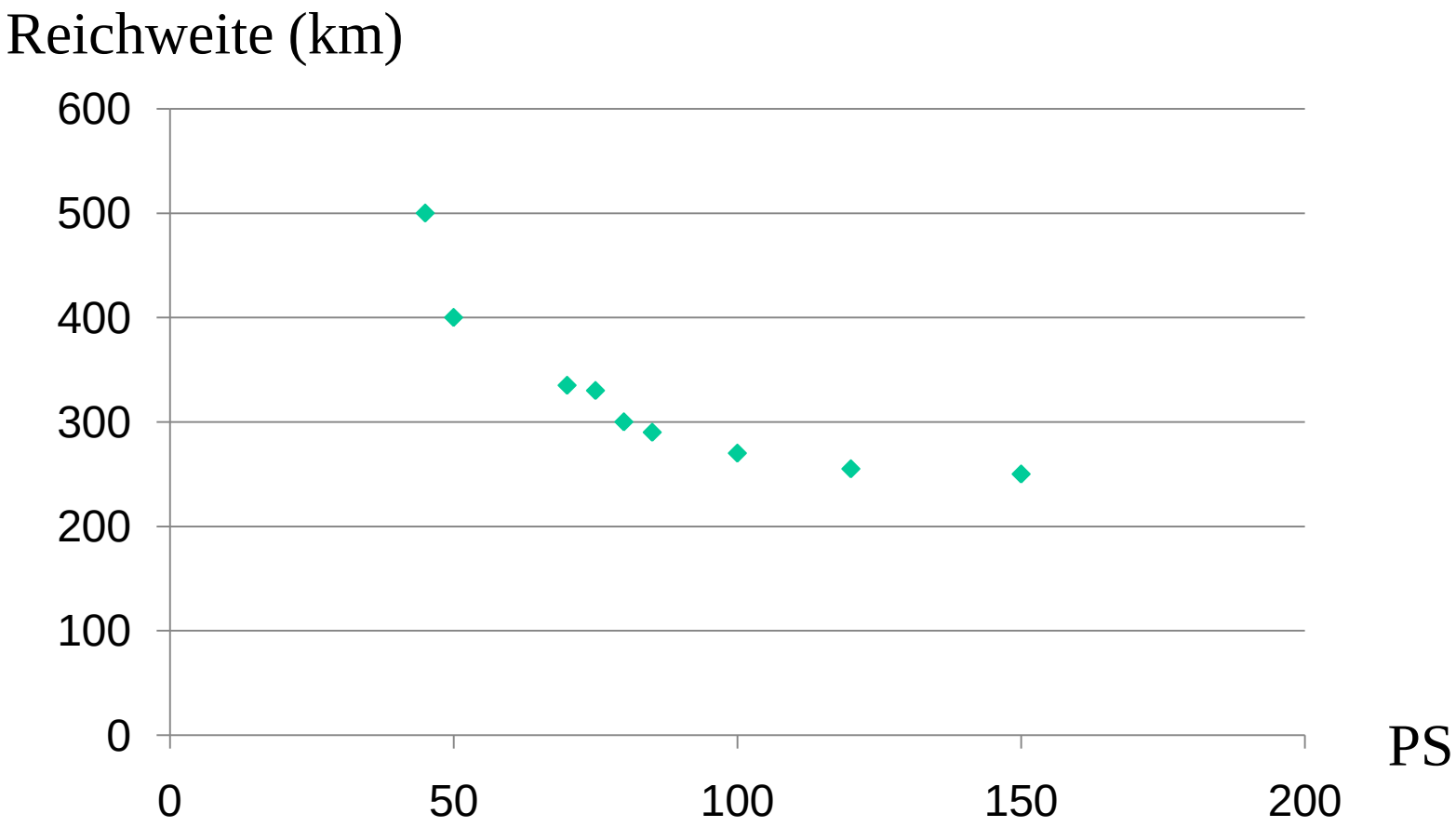
- **Optionen gut repräsentieren**
  - Übersicht !

- **Werte „nah“ am echten Optimum**
  - Kein Tupel für meine Bewertungsfunktion deutlich besser

Approximate Dominating Representatives - ADR

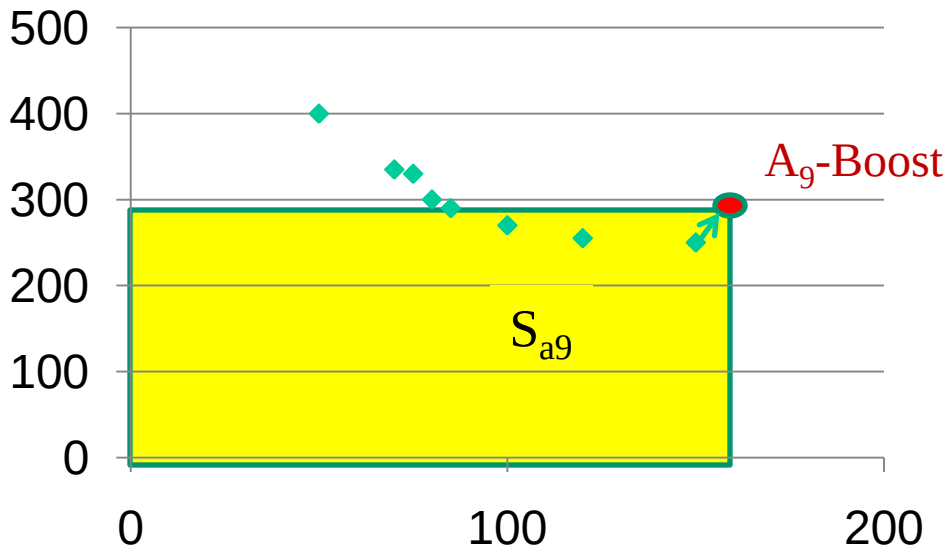
- **Max-Sortierung**
  - Große Werte sind gut
- **Grundidee:**
  - **ε-ADR:**
    Menge von Tupeln, die alle anderen Tupel dominieren würden, wären sie in allen Dimensionen um Faktor ε „geboostet“
  - A * (1+ ε) ≥ B (Pseudo-dominanz)
  - Jedes Tupel in ADR maximal ε von Optimum entfernt
  - Teilmenge der Skyline-Tupel

ADR-Beispiel:



Optimale ADR

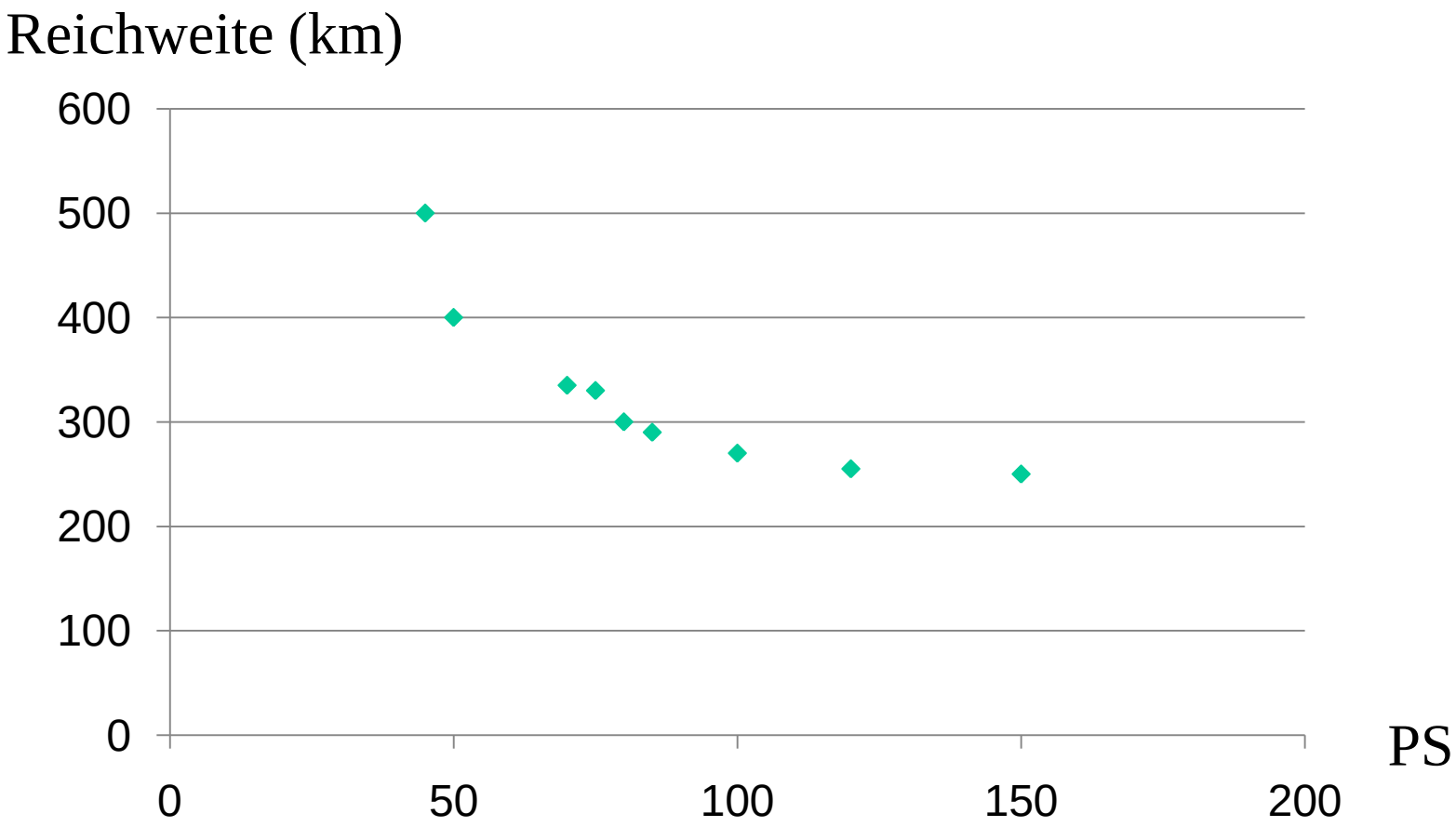
- **Wie finde ich möglichst kleine ADR ?**
- **Set-Cover Problem:**
  - S_{a1} = Menge aller von a_1 Pseudo-dominierten Tupel
  - Welche a_i reichen, um alle Tupel ab zu decken ?



Optimale ADR

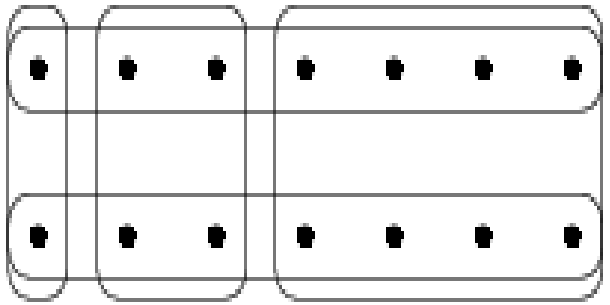
- **2-Dimensional:**
  - Greedy-Algorithmus
  - Linear (bei sortierter Menge)
  - Optimal (Ergebnisgröße)
- **Vorgehen:**
  - Dimensionen D1 und D2
  - Finde das aktuell beste Tupel in D1 A
  - Finde alle Tupel, die A Pseudo-dominieren könnten B_i
  - Wähle das B_i mit bestem Wert in D2 B_i^*
  - B_i^* ist in ADR

ADR-Greedy-2D:



Optimale ADR

- **Mehr als 2 Dimensionen:**
  - Problem ist NP-hart
- **Genau 3 Dimensionen:**
  - Approximation
  - Polynomieller Greedy Algorithmus
  - Fehler im Bereich ln n



- **Verfahren:**
  - Wähle größte verbleibende Teilmenge

Fazit

- **Pro:**
  - Reduziert Ergebnismenge innerhalb vorgegebener Grenzen
  - Gutes Verfahren in 2 Dimensionen
- **Kontra:**
  - Ab 3 Dimensionen schlechte Laufzeit
  - Keine (direkte) Kontrolle über Ergebnisgröße
  - Praktische Anwendung ?


Fragen ?