TECHNISCHE UNIVERSITÄT DRESDEN

Fakultät Informatik, Institut Systemarchitektur, Lehrstuhl für Datenbanken

Sampling: Online oder Offline?

Rainer Gemulla

Gliederung

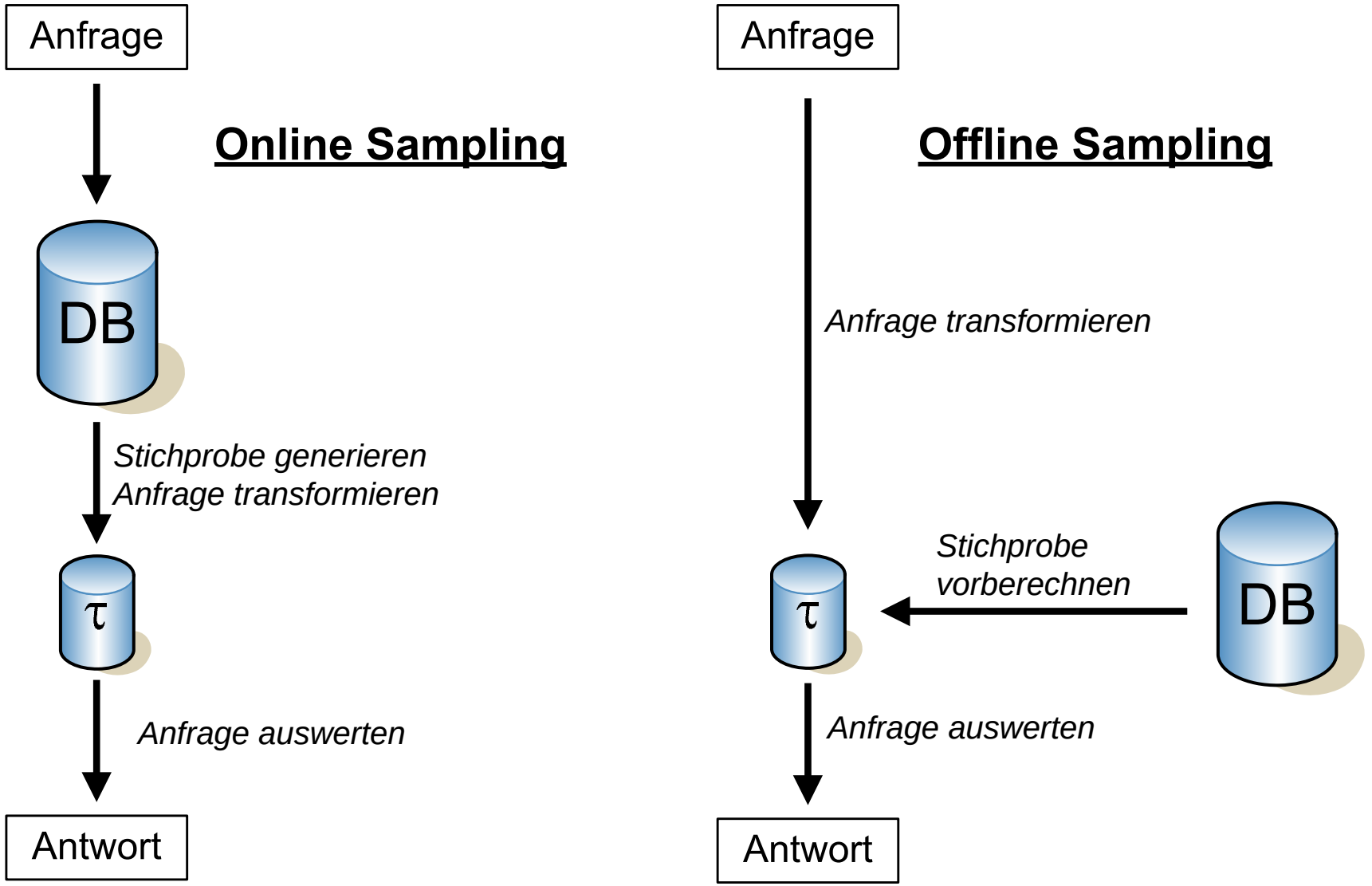
- **Nutzen**
  - Geschwindigkeit
    - große Datenmengen
    - komplexe Algorithmen

Umsatz in Europa (TPCH)		
1%	8.46 Mil. ± 0.15 Mil.	4s
10%	8.51 Mil. ± 0.05 Mil.	52s
100%	8.54 Mil.	200s

- **Einsatzgebiete**
  - näherungsweise Anfragebeantwortung
  - Anfrageoptimierung
  - Lastbalancierung
  - Data Mining
  - interaktives Anfragedesign
  - Antwortzeitschätzung

Woher nehmen?

Online versus Offline Sampling



ICICLES

Congressional Sampling

Join Synopses

Small Group Sampling

Histogram construction

Outlier Indexing

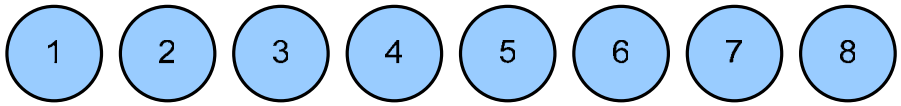
Online Aggregation

Sampling with Preaggregated Data

usw.

Simple Random Sample (SRS)

- **Gleichwahrscheinlichkeit**
  - aller möglichen Stichproben (einer Größe)
  - allgemeinste Form
- **Beispiel:** 2 aus 8

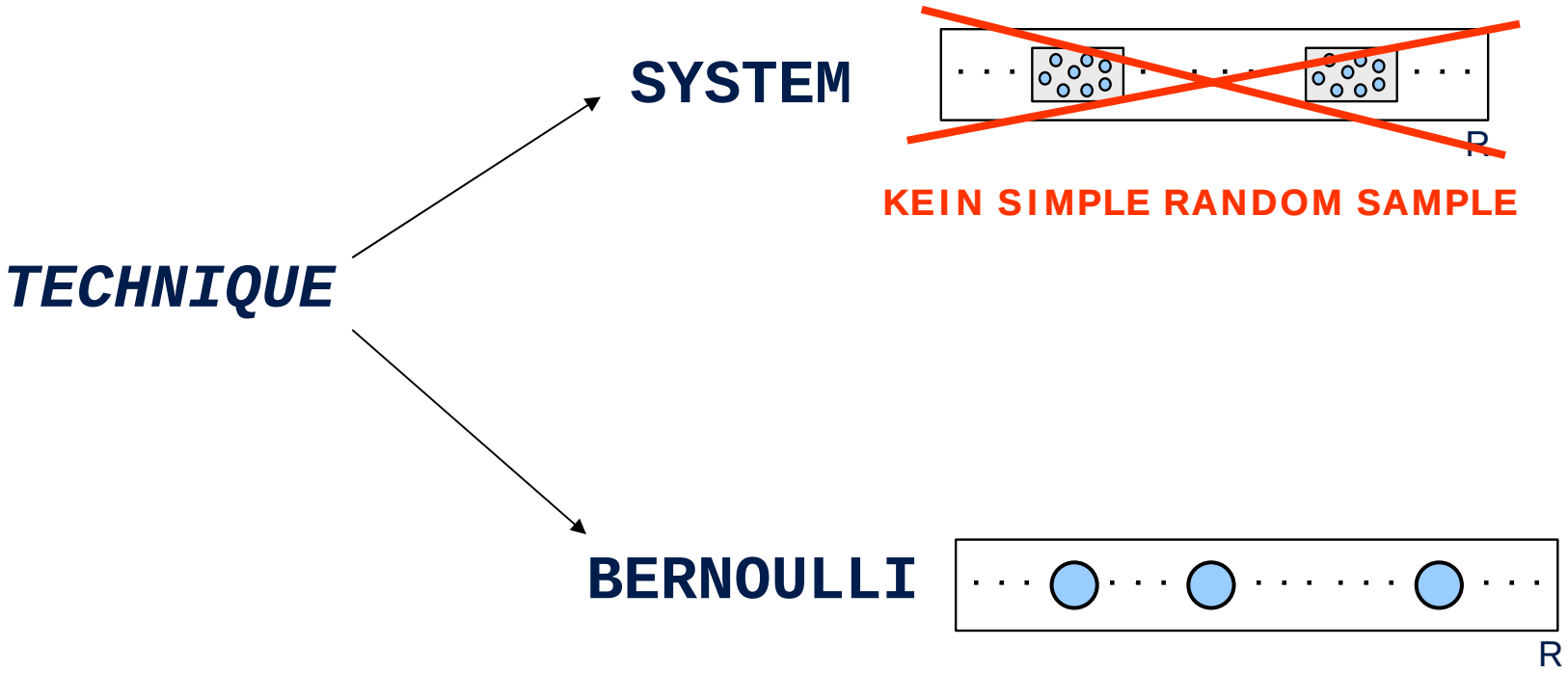


- **Möglichkeiten**

11	12	13	14	15	16	17	18
21	**22**	23	24	25	26	27	28
31	32	**33**	34	35	36	37	38
41	42	43	**44**	45	46	47	48
51	52	53	54	**55**	56	57	58
61	62	63	64	65	**66**	67	68
71	72	73	74	75	76	**77**	78
81	82	83	84	85	86	87	**88**

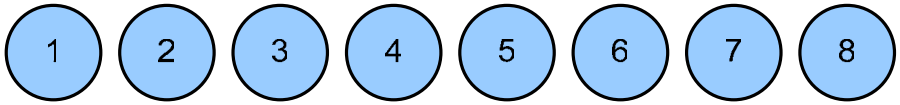
- **Anzahl verschiedener Stichproben**
  - ohne Zurücklegen: (N n) = 28 (13983816)
  - mit Zurücklegen: (N+n-1 n) = 36 (11440)
- **Reihenfolge für SRS unwichtig!**

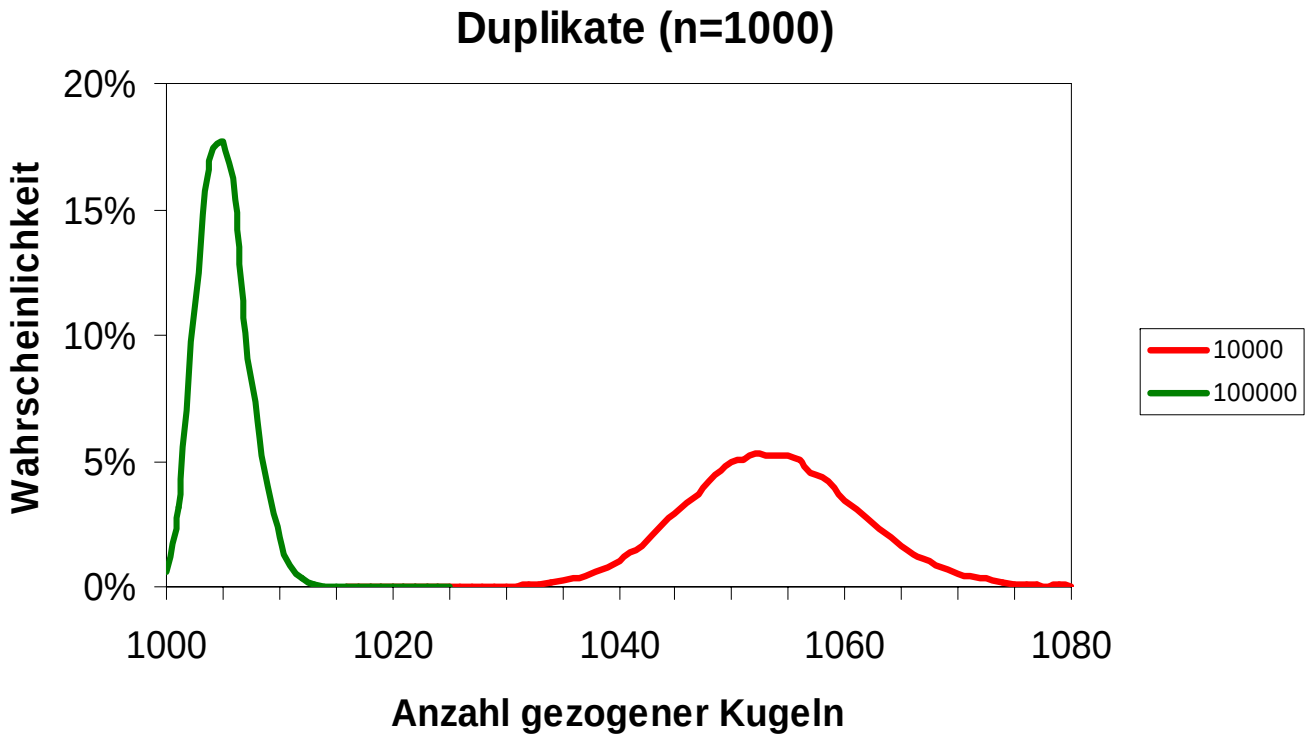
- **TABLESAMPLE-Klausel**
  - **SELECT** * **FROM** BIGTABLE **TABLESAMPLE** ***TECHNIQUE*** (1)



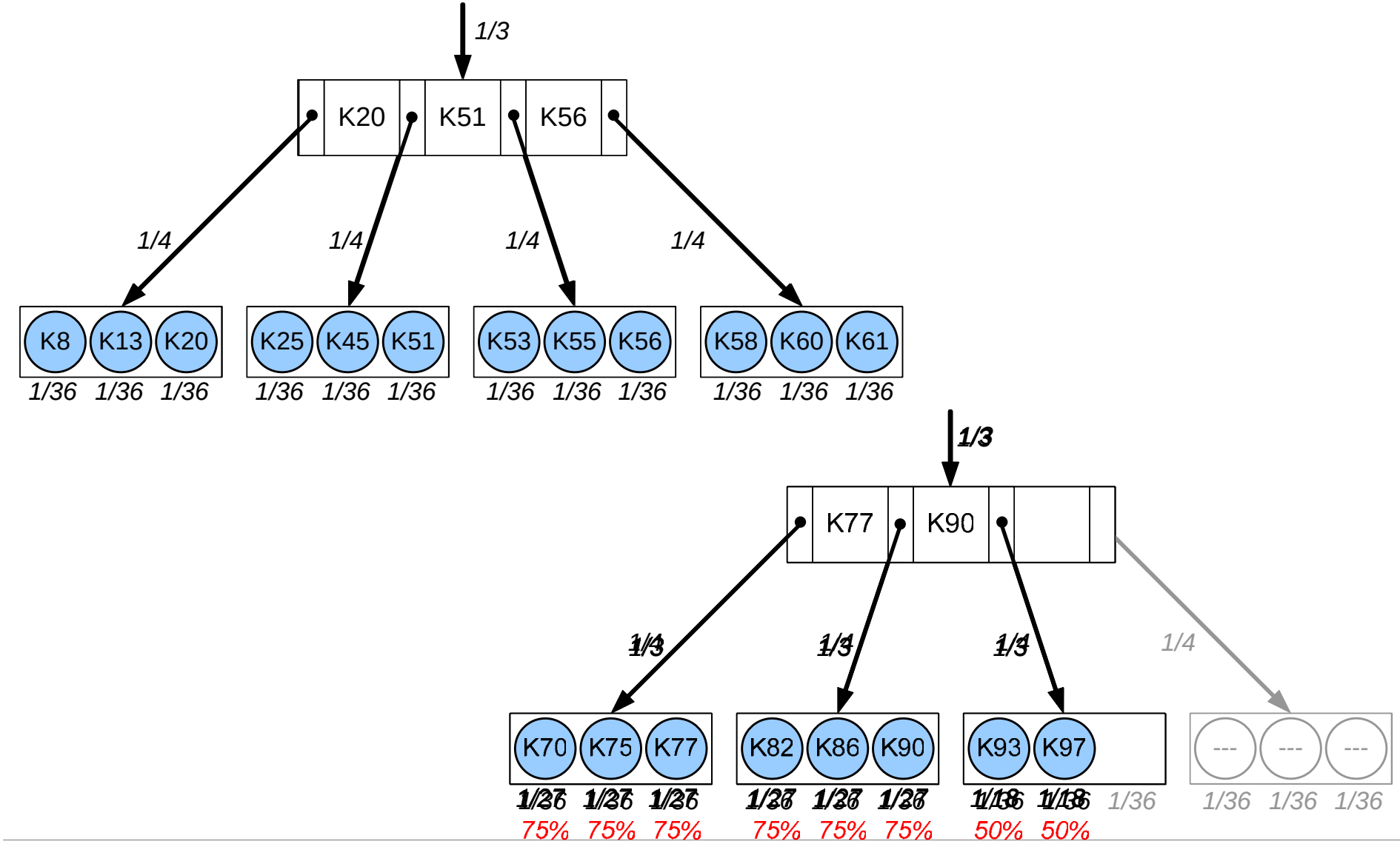
Gliederung

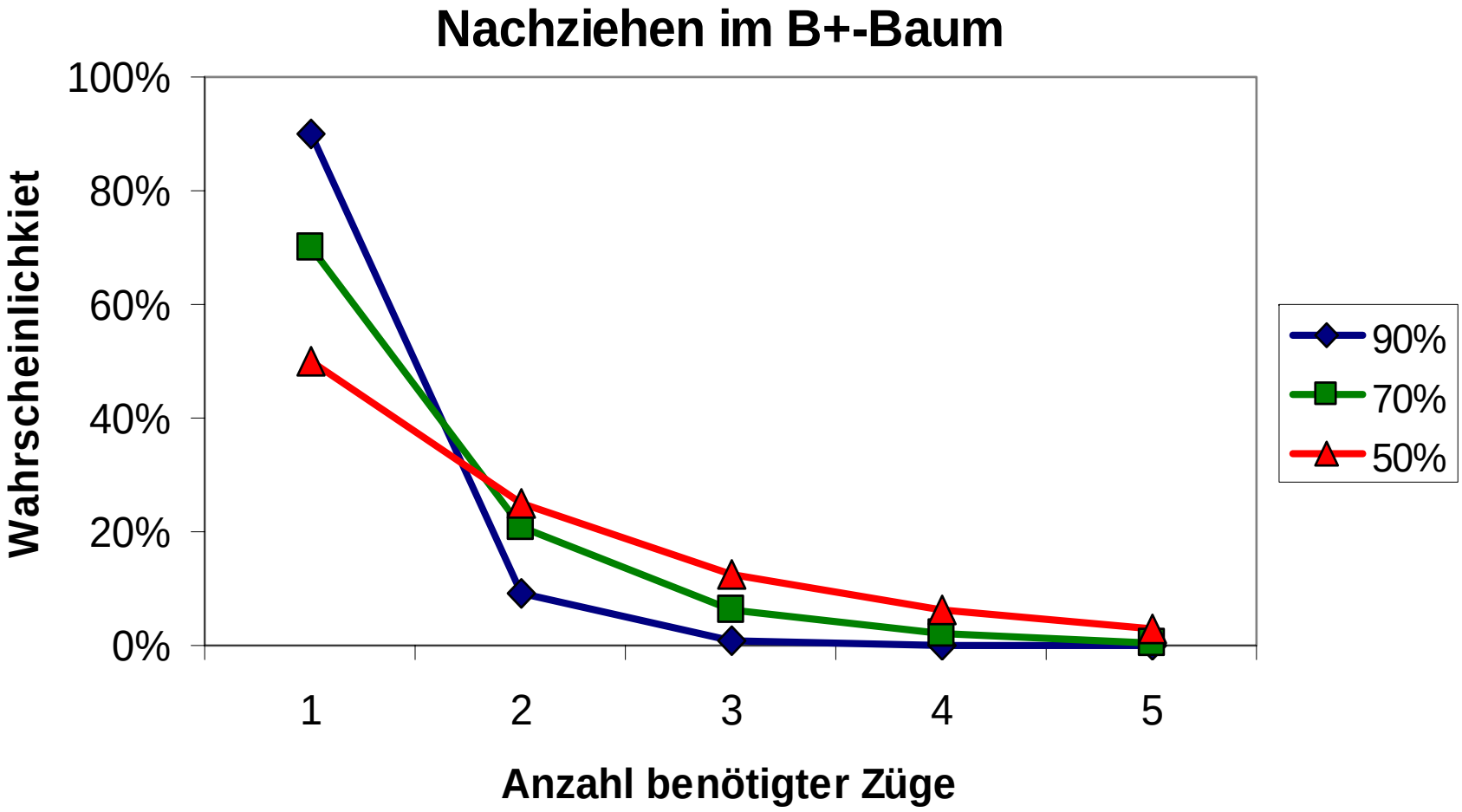
- **Kernidee**
  - Kugel aus Urne ziehen → Tupel aus Relation ziehen



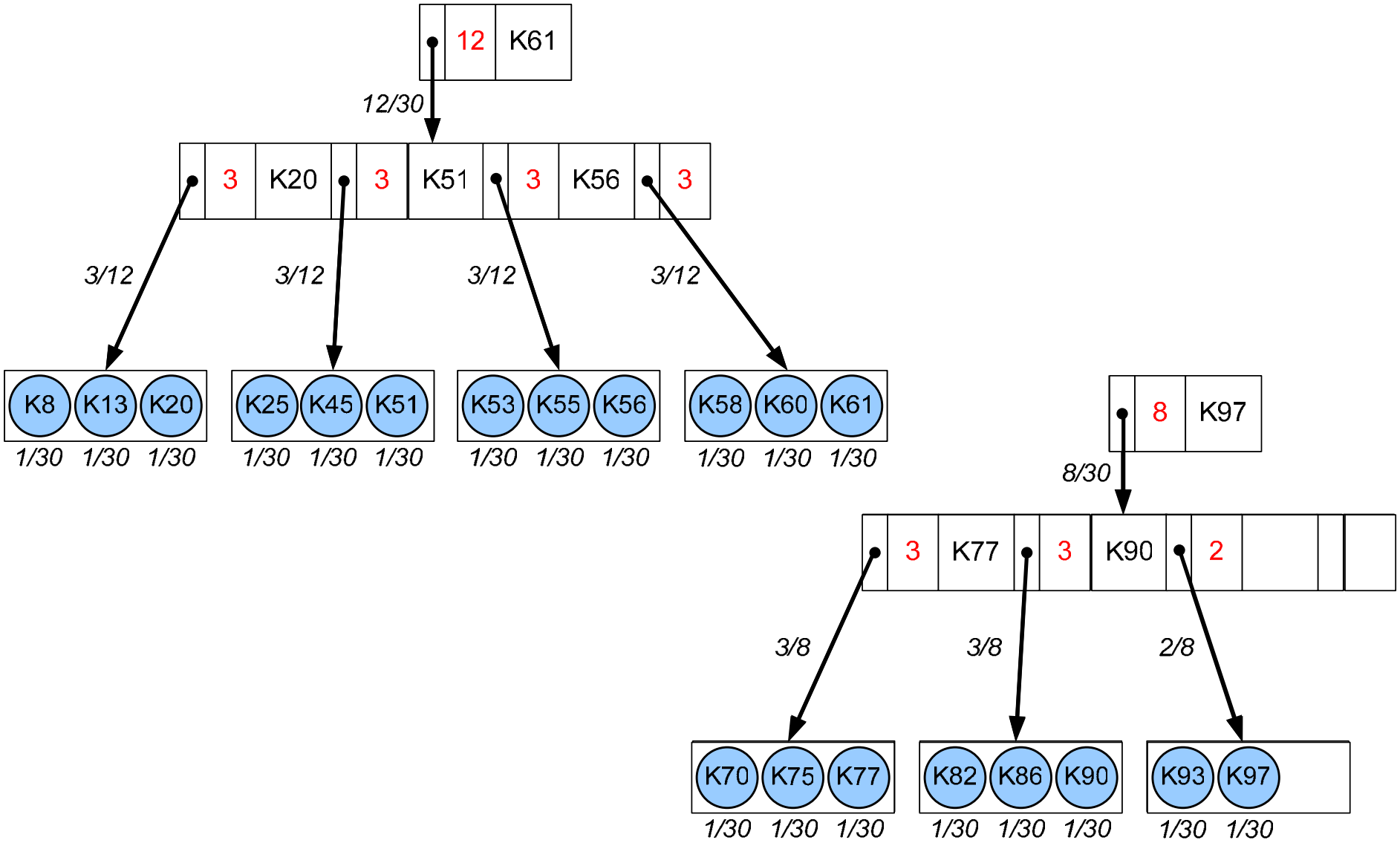


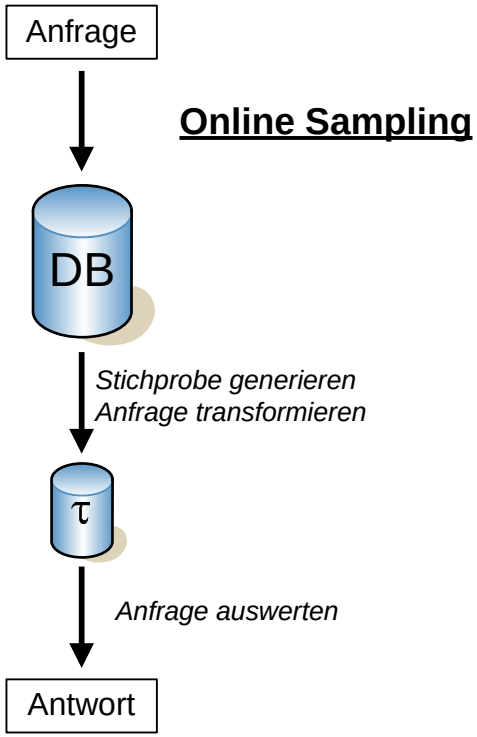
SRS aus B+-Baum




Nachziehen im B+-Baum
100%
80%
60%
40%
20%
0%
Wahrscheinlichkiet
1
2
3
4
5
Anzahl benötigter Züge
90%
70%
50%

SRS aus gewichtetem B+ Baum



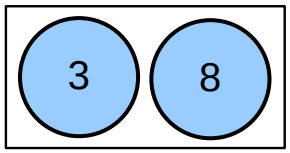


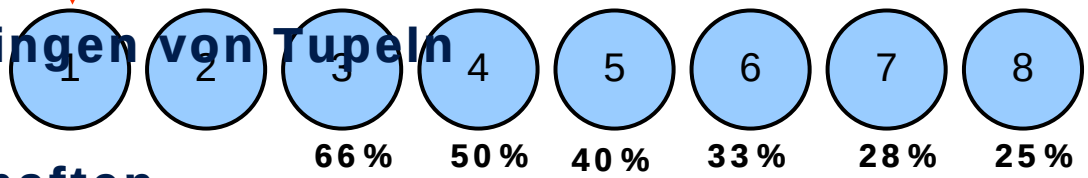
Fazit	
Pro	**Contra**
•Keine Wartung	•Aufwändig
•Progressiv	•Nur eine Relation
•Kein Speicheraufwand	•Umgang mit Datenverzerrung

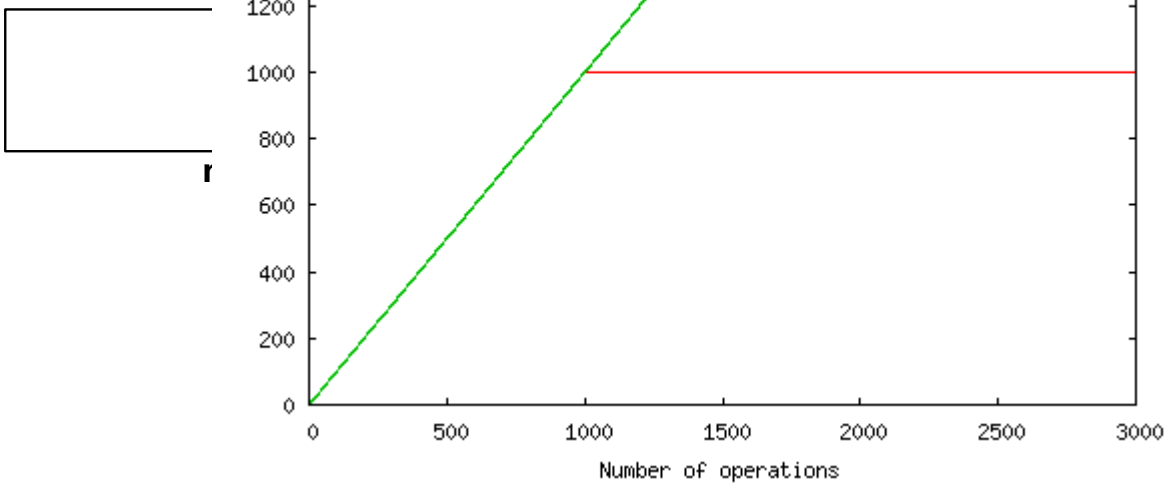
Gliederung

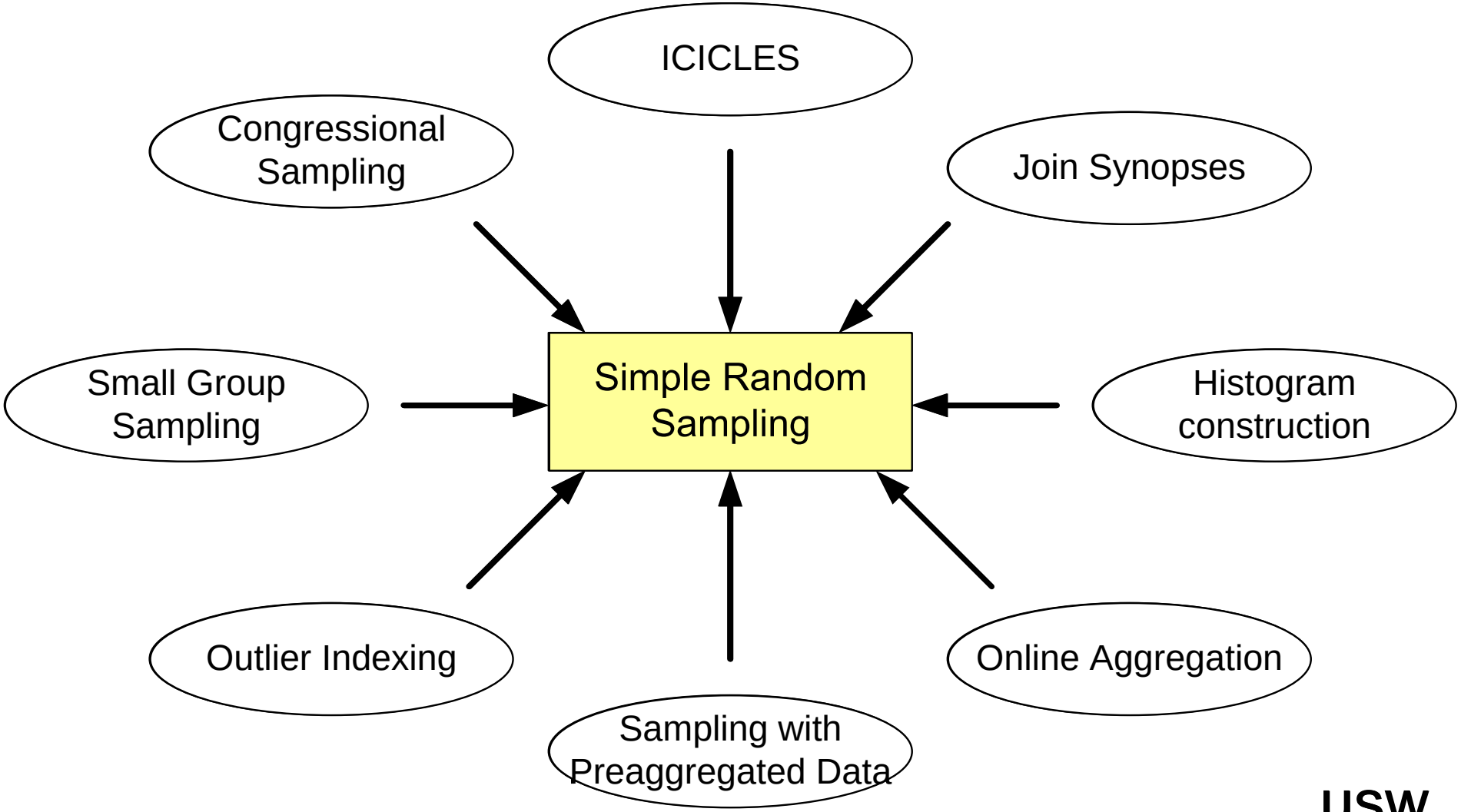
Erhebung des SRS vor der Abfrage

- **Reservoir Sampling**
  - Stichprobenerhebung durch Relationenscan
  - Aufnahmewahrscheinlichkeit: n / (N+1)
- **Überspringen von Tupeln**
- **Eigenschaften**
  - effizient
  - beliebige Eingabedaten
  - Teilstichproben
  - inkrementell wartbar

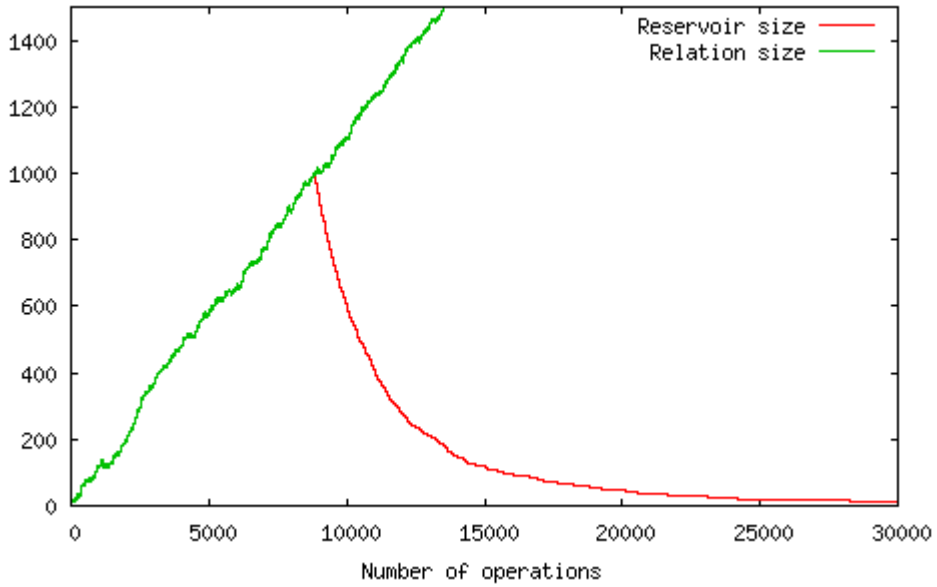







ICICLES
Congressional Sampling
Join Synopses
Small Group Sampling
Simple Random Sampling
Histogram construction
Outlier Indexing
Online Aggregation
Sampling with Preaggregated Data
USW.

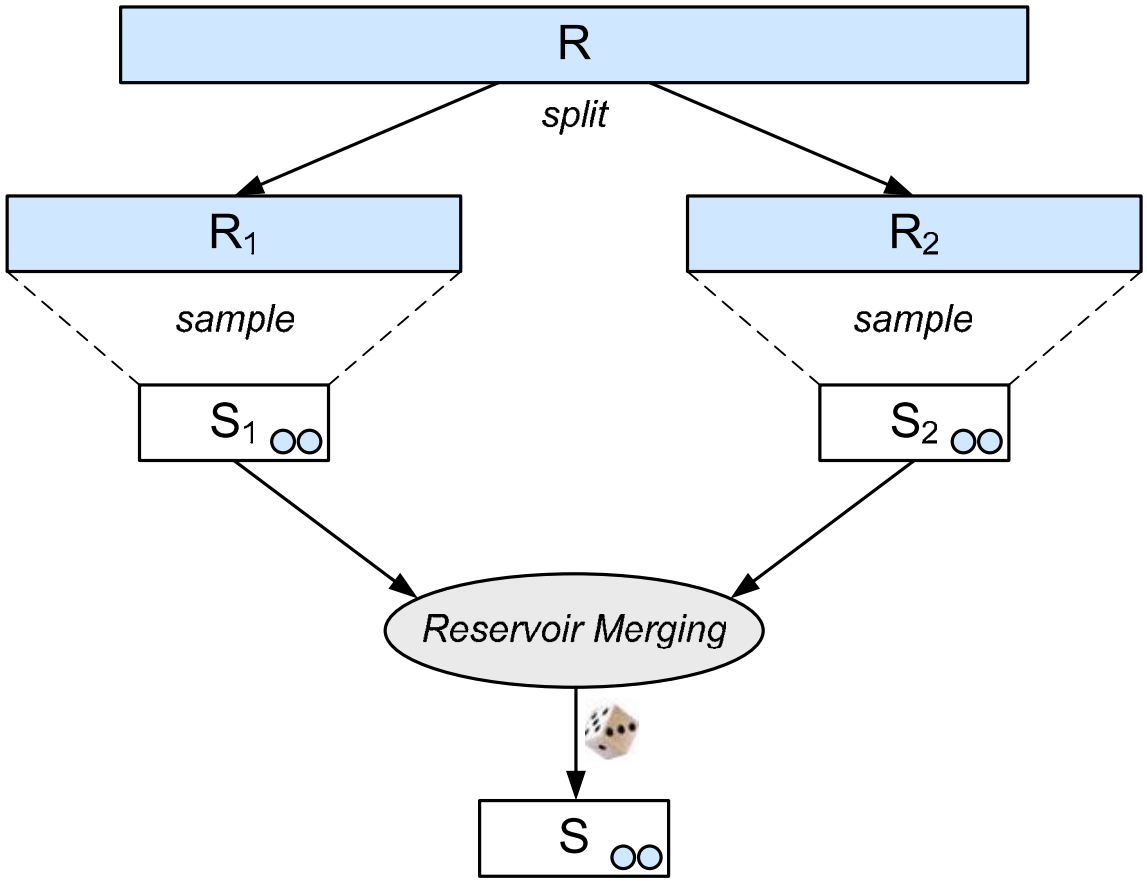
- **Inkrementelle Wartbarkeit**
  - Einfügen → kein Problem
  - Löschen → Stichprobengröße fällt!



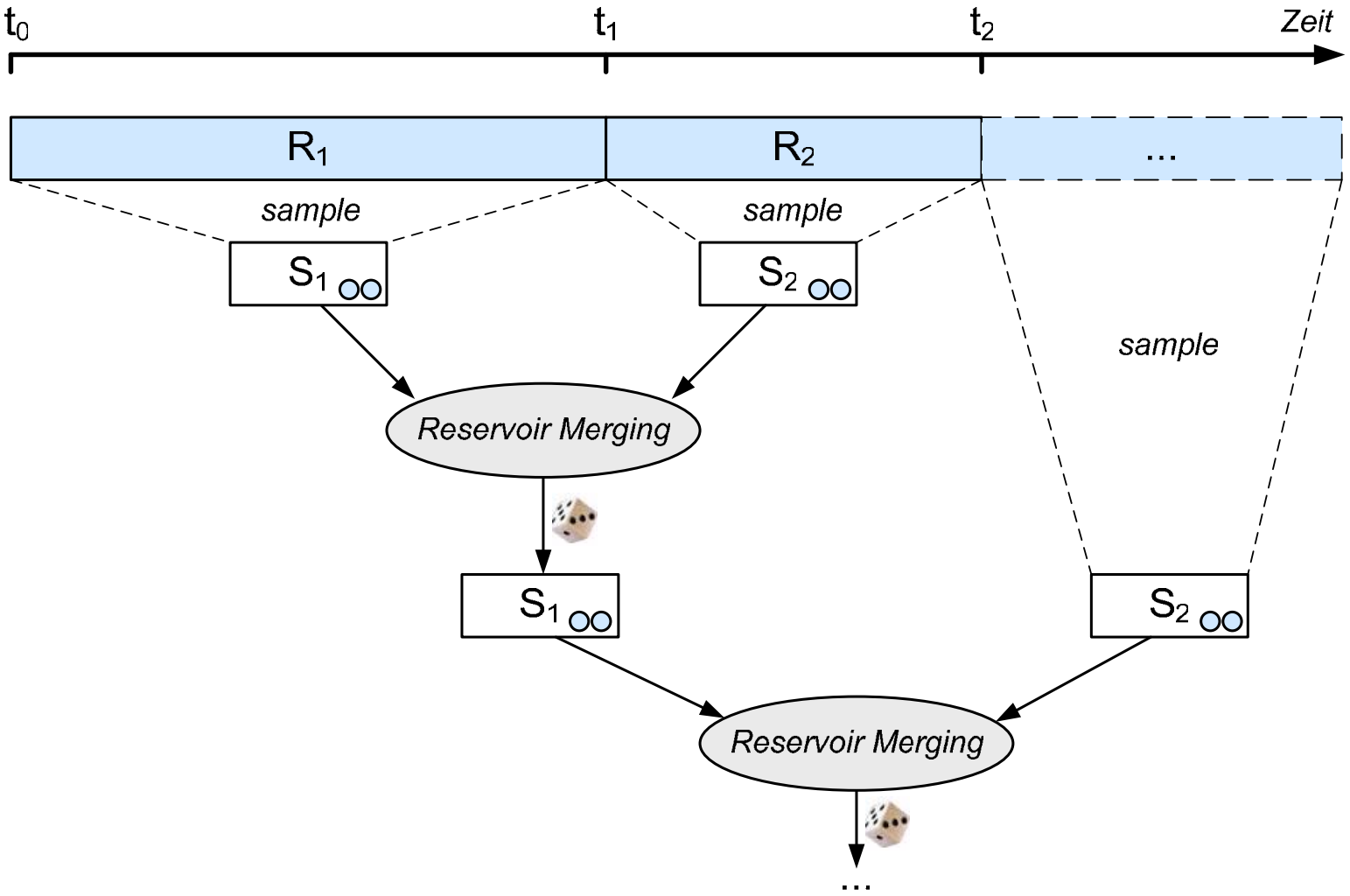
- **Lösung**
  - Nachziehen von Tupeln → teuer
  - Nutzung von neu eingefügten Tupeln

  → *Adaptive Reservoir Sampling*

Reservoir Merging

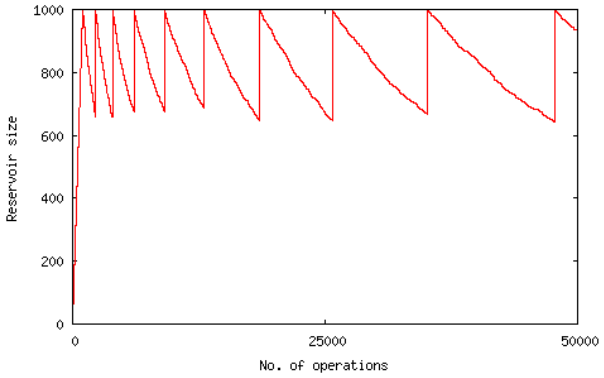


Adaptive Reservoir Sampling

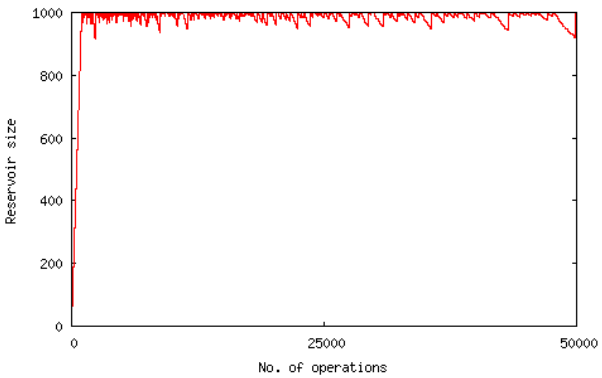


Vereinigungszeitpunkt

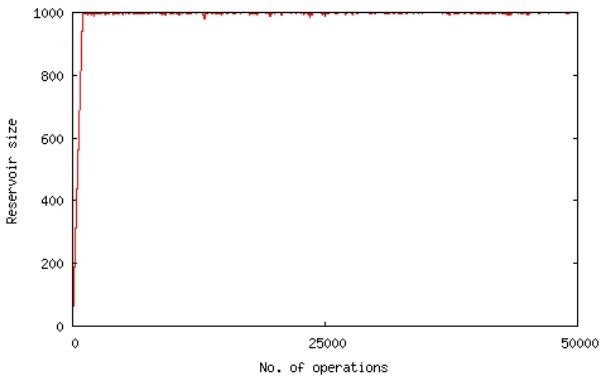
STATIC



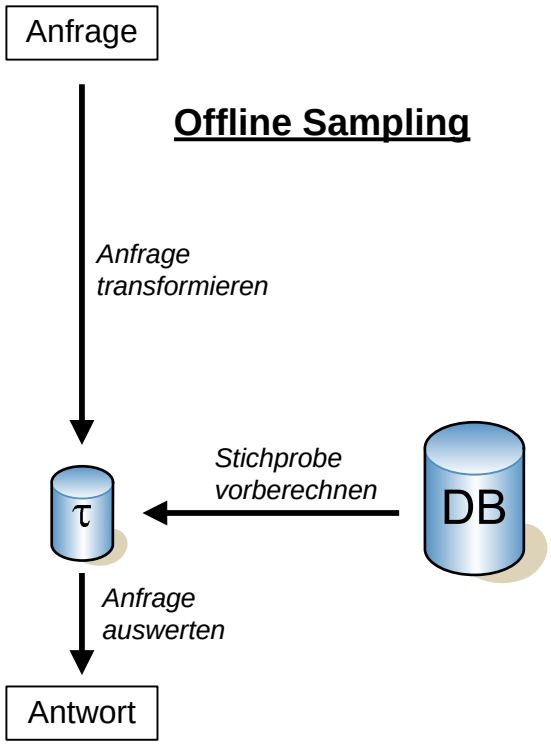
PROB



SIZE



Offline Sampling



Fazit	
Pro	**Contra**
•Schnell	•Speicheraufwand
•Beliebige Daten	•Wartungsaufwand
•Umgang mit Datenverzerrung	
•Progressiv	

Fazit

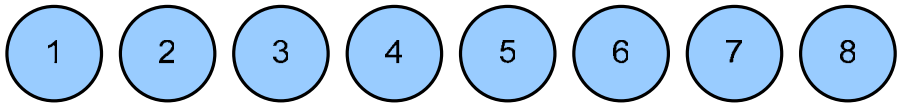
Zusammenfassung	Online	Offline
Speicherplatz	+	
Wartung	+	-
Geschwindigkeit	-	+
Genauigkeit		+
Vielseitigkeit		+
Progressivität	+	

- **Sampling an der TU Dresden**
  - DFG-Projekt: Adaptives Offline Sampling
  - IBM Joint Study Agreement

Vielen Dank!

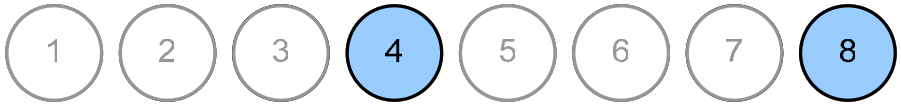
Fragen?

- **Systematische Stichprobe**
  - Beispiel: 2 aus 8 (ohne Zurücklegen)



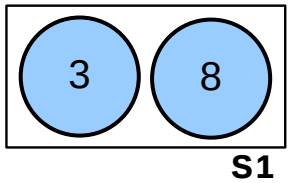
  - zufälliger Startpunkt in 1…k, dann jedes k-te Tupel nehmen

k=4

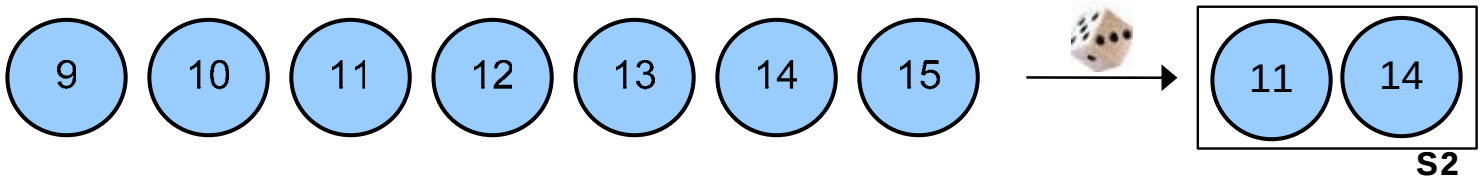


  - Eigenschaften
    - jedes Tupel gleichwahrscheinlich
    - liefert *N/k* Tupel
    - *k* verschiedene Stichproben

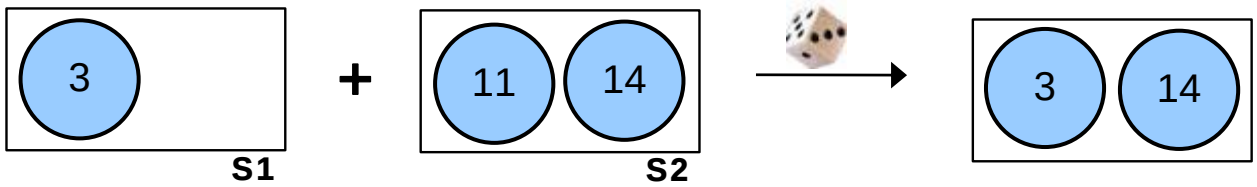
- **Idee**
  - nach Löschen: Stichprobe einfrieren



  - neu eingefügte Tupel zwischenspeichern (*stratified sampling*)



  - Zusammenführen (*reservoir merging*)



    - hier: Erfolgswahrscheinlichkeit 77%