

DBDBD 2016

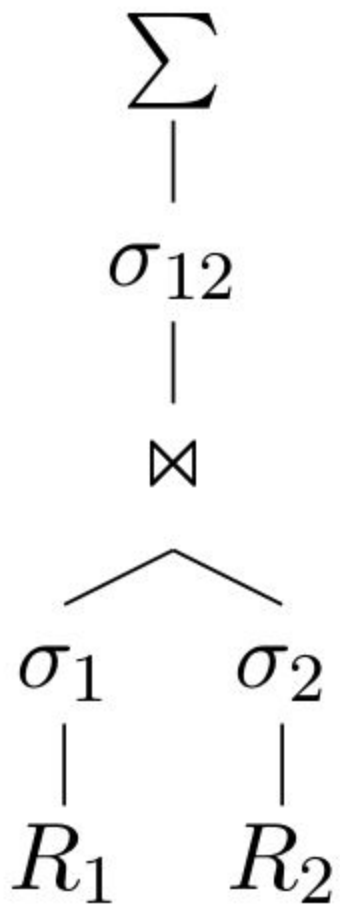
Estimating Aggregates Over Joins

— Abe Wits —

with thanks to my supervisors

Hannes Mühleisen & Lefteris Sidirourgos

Setting



Estimating Sums

$$S \subset R$$

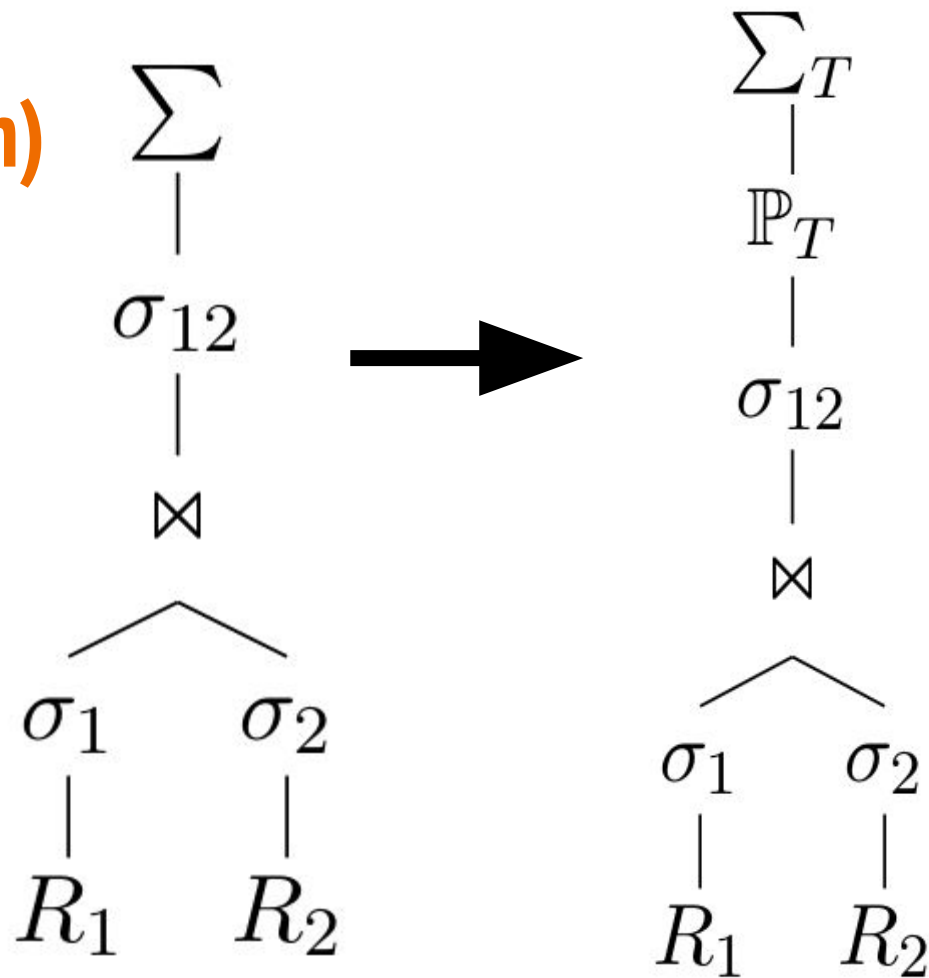
$$\sum R \approx \frac{|R|}{|S|} \sum S$$

$$\sum_{R}$$

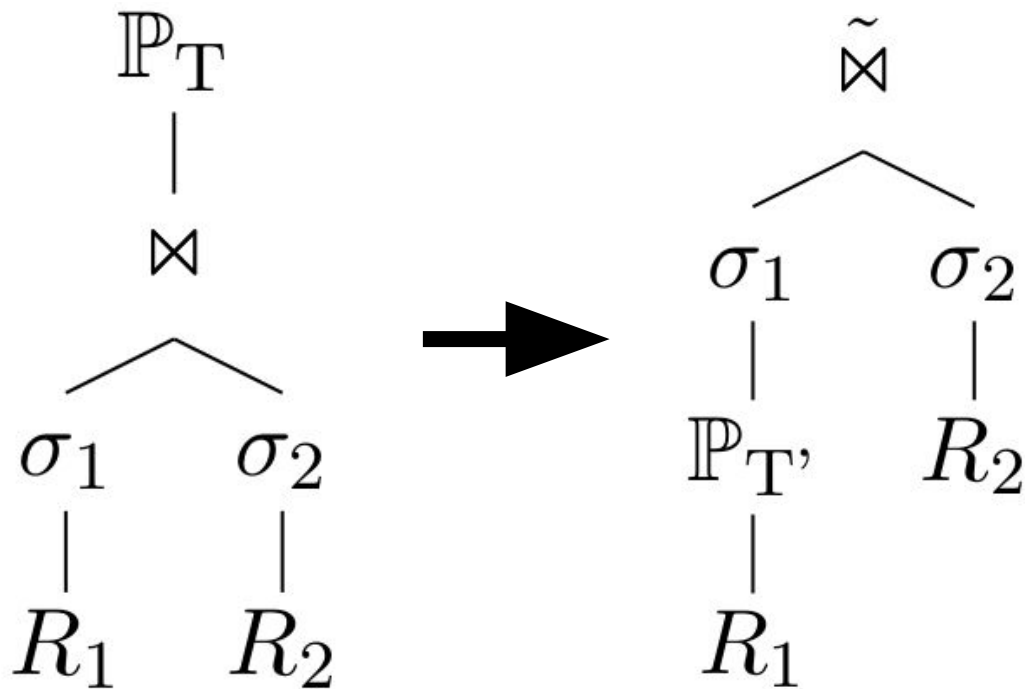


$$\sum_{\mathbb{P}_T} R$$

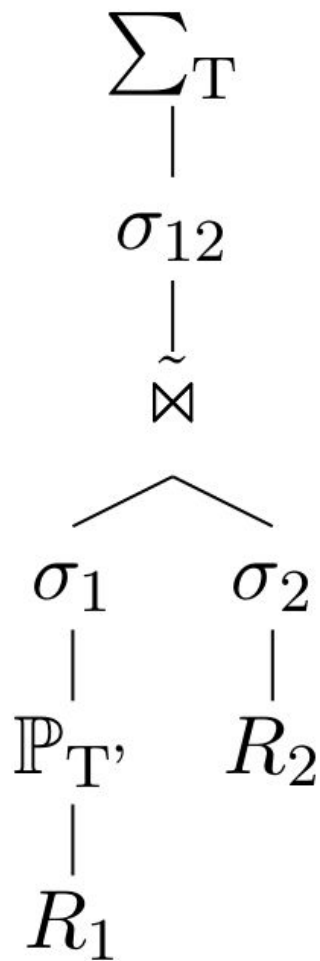
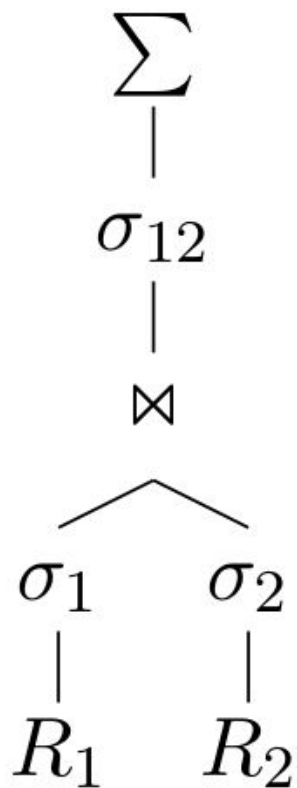
Setting (Estimate Sum)



Pushing Sampling down Join (SSJ)



Setting (Estimate Sum + SSJ)



Improve on this!

1) $\sum R \approx \sum_T S \subset R$ can be improved

2) \mathbb{P}_T
|
 R_1 takes $O(|R_1|)$

Improve on this!

1) $\sum R \approx \sum_T S \subset R$ can be improved

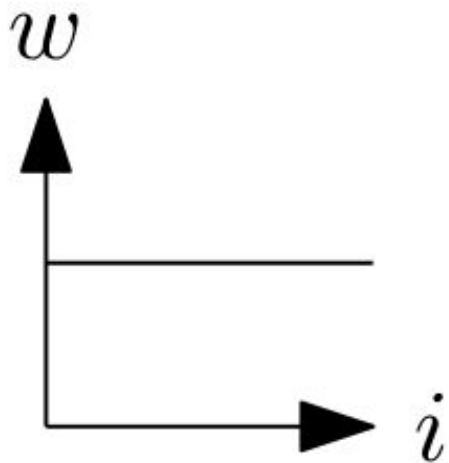
2) \mathbb{P}_T
|
 R_1 takes $O(|R_1|)$

Estimating Sums $\sum R \approx \sum_T S \subset R$

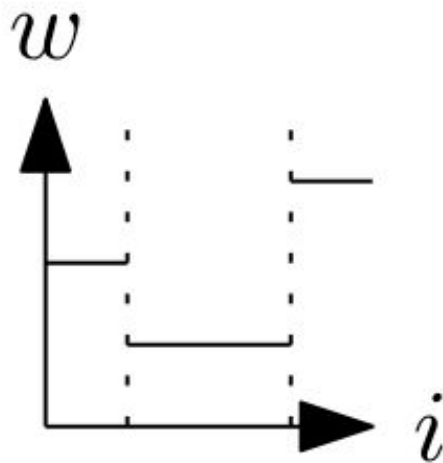
uniform

stratified

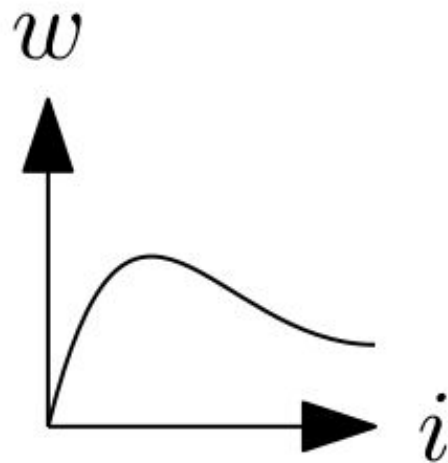
weighted



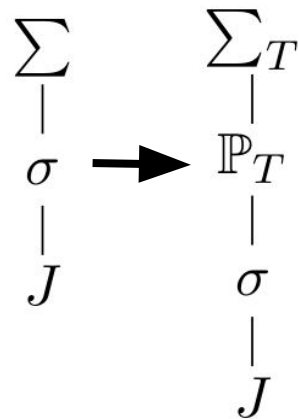
$$\frac{|R|}{|S|} \sum S$$



$$\sum_{i=1}^{\text{\#strata}} |R_i| \text{avg}(S_i)$$



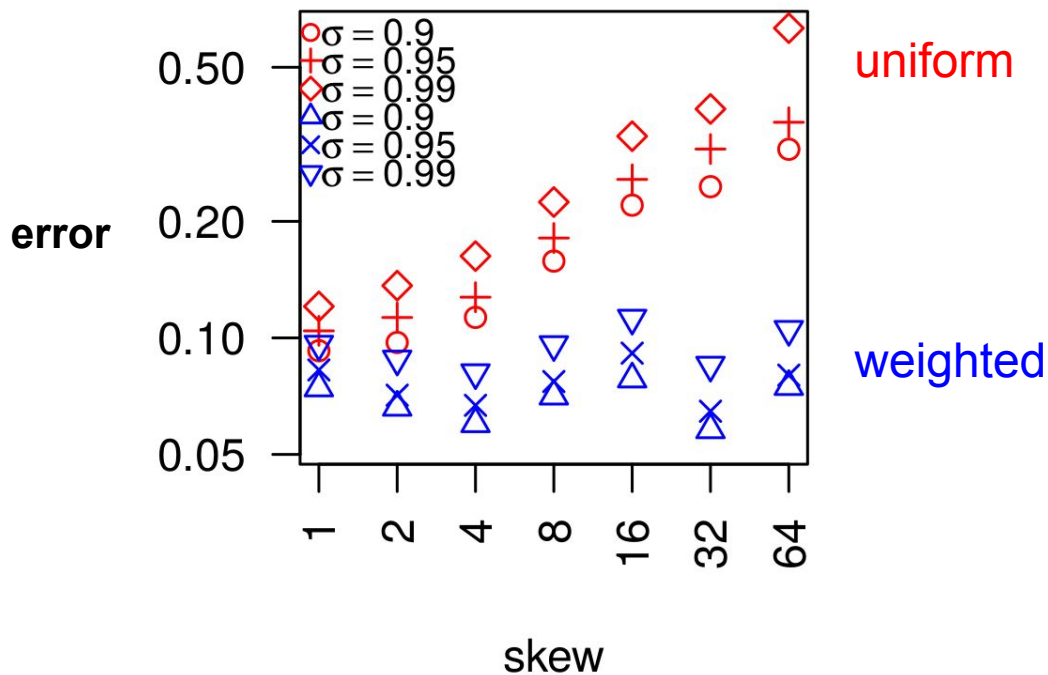
$$\frac{1}{|S|} \sum_{i=1}^{|S|} \frac{s_i}{p_i}$$



This matters!

Adjusting Weights for Aggregation

sample 1000 elements out of 10^7



Improve on this!

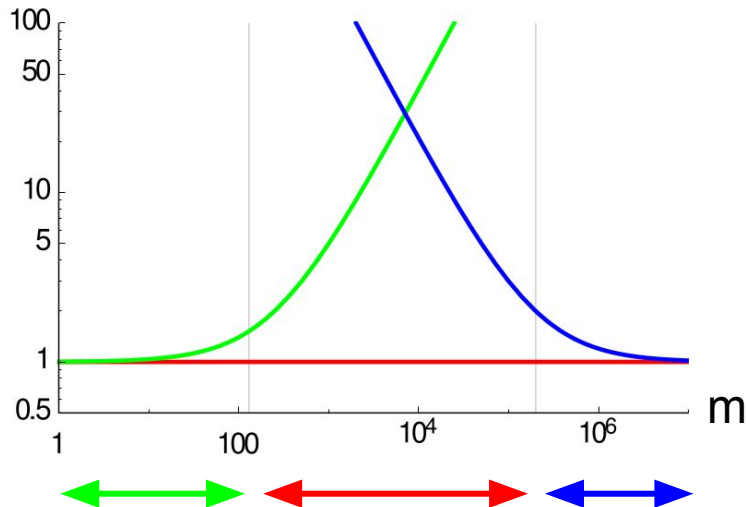
1) $\sum R \approx \sum_T S \subset R$ can be improved

2) \mathbb{P}_T
|
 R_1 takes $O(|R_1|)$

Faster Sampling

$$\mathbb{P}_T \mid R_1$$

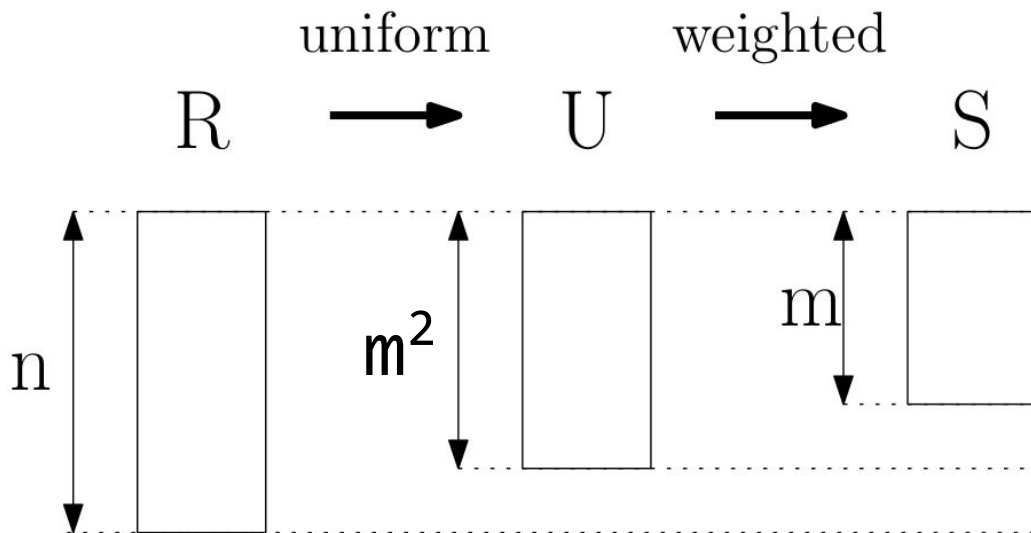
times slower than **Uniform**
(when including join and aggregation)



Cost of taking size m sample:

- Default $O(|R_1|)$
- Heuristic Sampling $O(m^2)$
- Use CDF of R_1 $O(m \log |R_1|)$
- Uniform $O(m \log |R_1|)$

Heuristic Sampling



Questions?