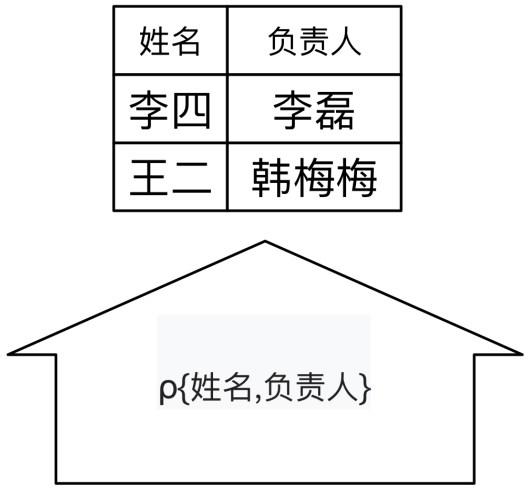
基于深度学习查询基数估计

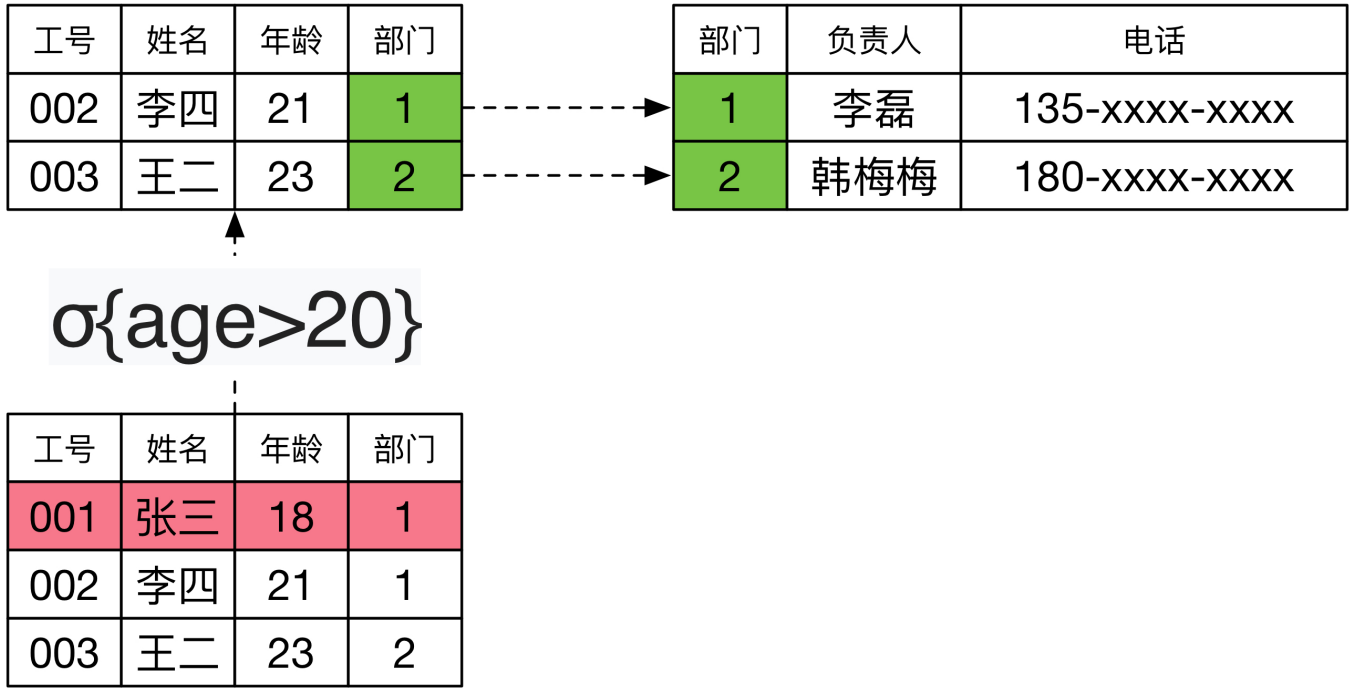
孙佶

查询基数

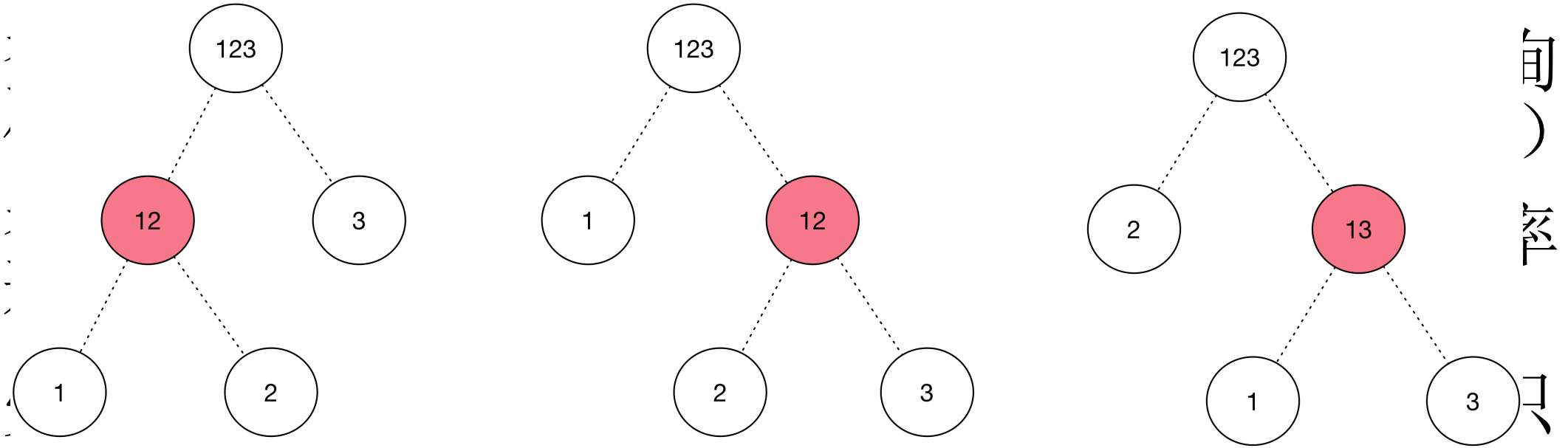
- 对于任意
  行数就是



询结果的



基数估计的意义



有统计直方图或者极其有限的采样。

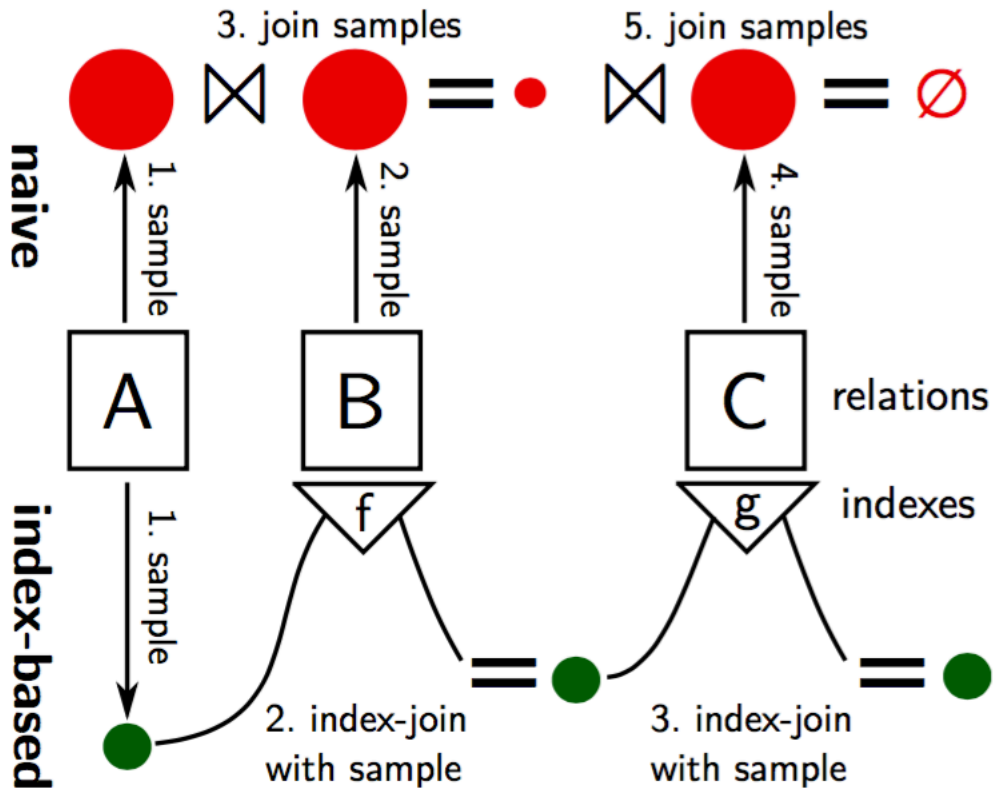
1. How Good Are Query Optimizers, Really?, Viktor etc., VLDB2015

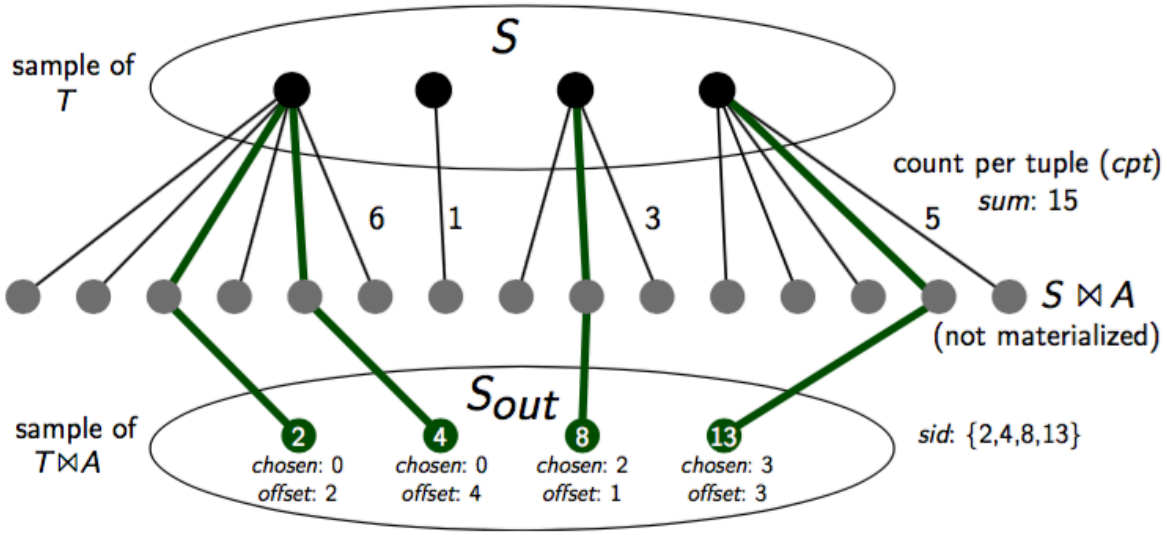
传统基数估计方法

- 基于统计直方图的估计(Database Profile)
  - 均匀性和独立性假设导致低估
  - 获得精确分布不现实

- 基于采样的估计(Sampling Techniques)
  - 有限的采样
  - 多表连接快速退化

一种基于索引的基数估计





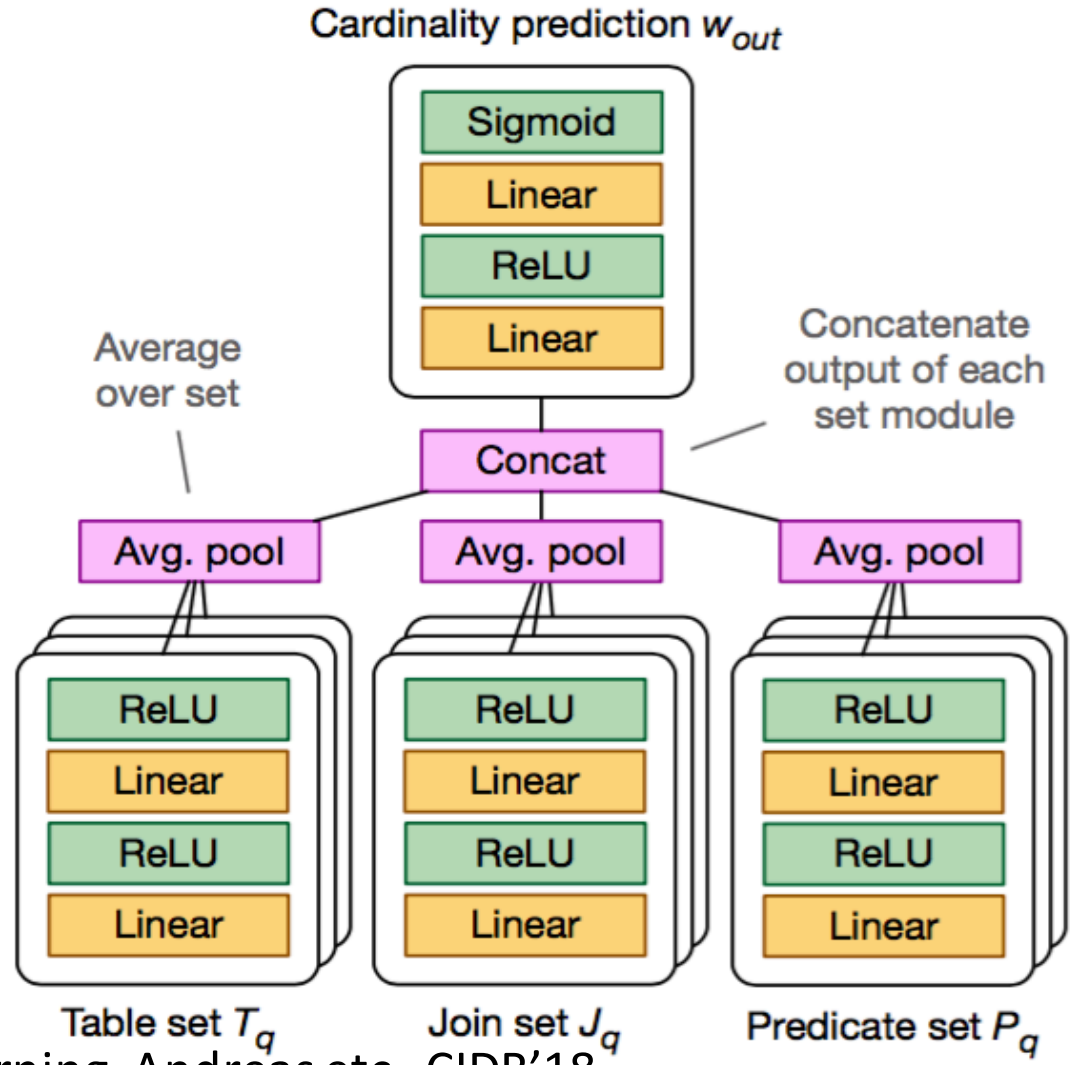
Cardinality Estimation Done Right: Index-Based Join Sampling, Viktor etc., CIDR'17

基于深度学习的基数估计

- 多集合卷积网络

SELECT * FROM A,B,C WHERE
A.pk = B.fk1 and B.fk2 = C.pk and
A.Attr > 10;

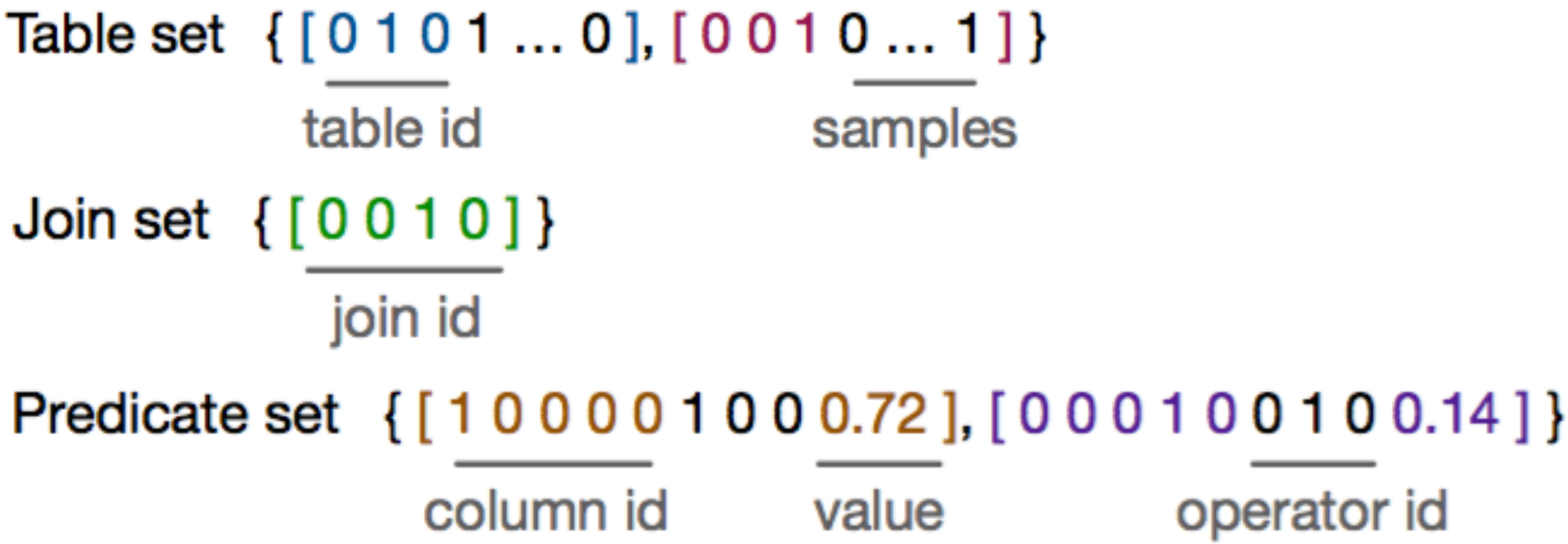
Table Set = {A,B,C}
Join Set = {A.pk-B.fk1, B.fk2-C.pk}
Predicate Set = {A.Attr > 10}



Learned Cardinalities: Estimating Correlated Joins with Deep Learning, Andreas etc., CIDR'18

基于深度学习的基数估计

- 输入向量表示



Learned Cardinalities: Estimating Correlated Joins with Deep Learning, Andreas etc., CIDR'18

训练集生成方法

- Random Select TableNum
- Random Select One Table as start point
- Count ← 0 repeat until count >= TableNum
  - Random Select One Table Joinable with selected one
- For Each Selected Tables:
  - Random Select PredNum
  - Random PredNum Columns
  - Random Select Operator {>,=,<} For Each Column
  - Random Select One Literal Value From Table For Each Column

Learned Cardinalities: Estimating Correlated Joins with Deep Learning, Andreas etc., CIDR'18

实验结果

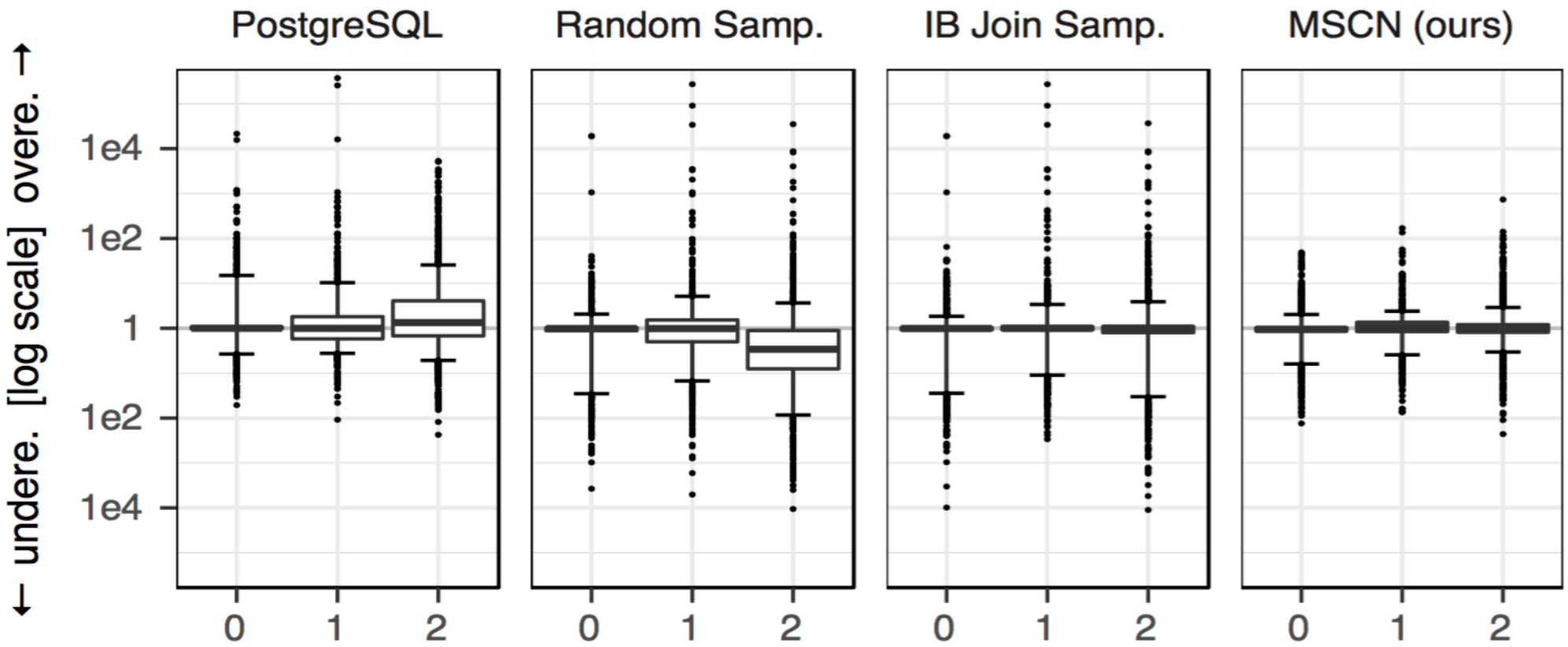
- 测试数据集

number of joins	0	1	2	3	4	overall
synthetic	1636	1407	1957	0	0	5000
scale	100	100	100	100	100	500
JOB-light	0	3	32	23	12	70

Learned Cardinalities: Estimating Correlated Joins with Deep Learning, Andreas etc., CIDR'18

实验结果

- 合成数据集上的测试



Learned Cardinalities: Estimating Correlated Joins with Deep Learning, Andreas etc., CIDR'18

实验结果

- 合成数据集上的测试

	median	90th	95th	99th	max	mean
PostgreSQL	1.69	9.57	23.9	465	373901	154
Random Samp.	1.89	19.2	53.4	587	272501	125
IB Join Samp.	**1.09**	9.93	33.2	295	272514	118
MSCN (ours)	1.19	**3.50**	**7.22**	**34.9**	**735**	**2.88**

Learned Cardinalities: Estimating Correlated Joins with Deep Learning, Andreas etc., CIDR'18

实验结果

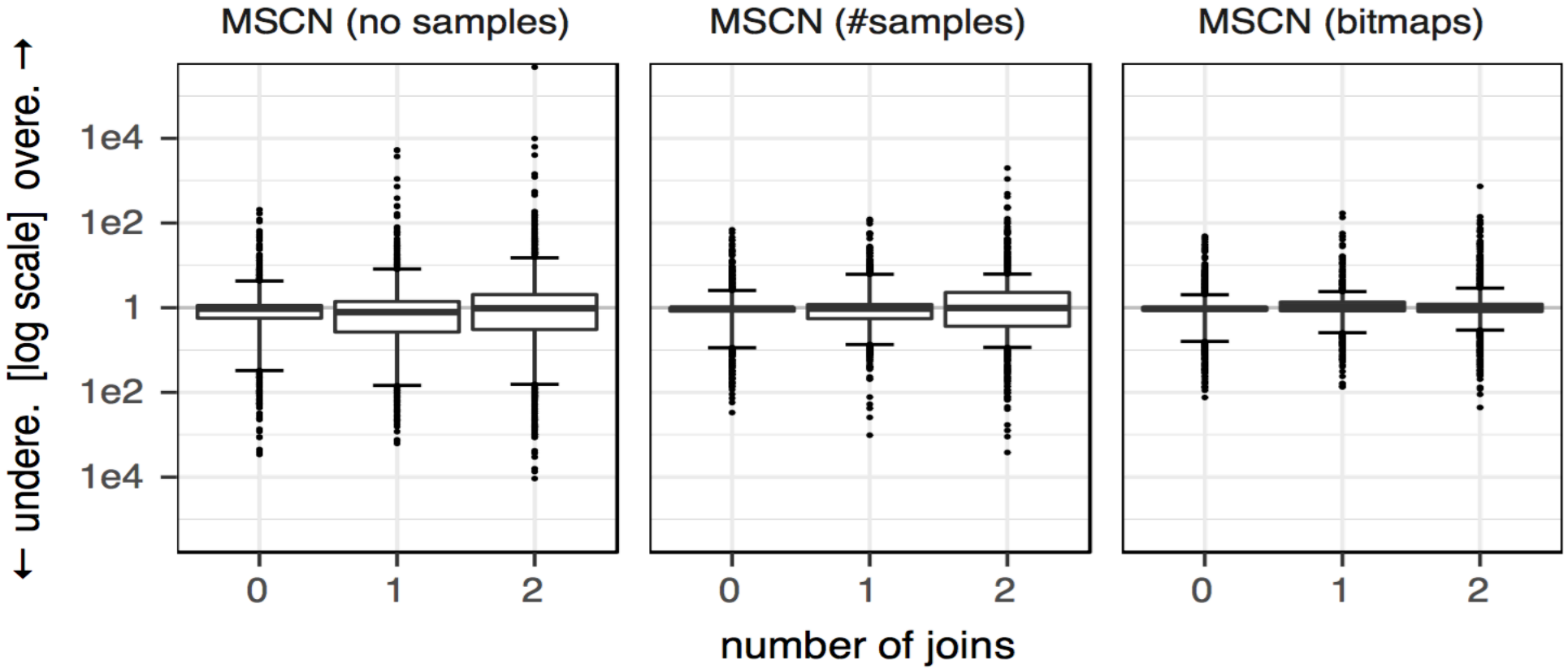
- 合成数据集上0-Tuple测试

	median	90th	95th	99th	max	mean
PostgreSQL	4.78	62.8	107	1141	21522	133
Random Samp.	9.13	80.1	173	993	19009	147
MSCN	**3.17**	**17.2**	**28.8**	**55.5**	**96.2**	**7.20**

Learned Cardinalities: Estimating Correlated Joins with Deep Learning, Andreas etc., CIDR'18

实验结果

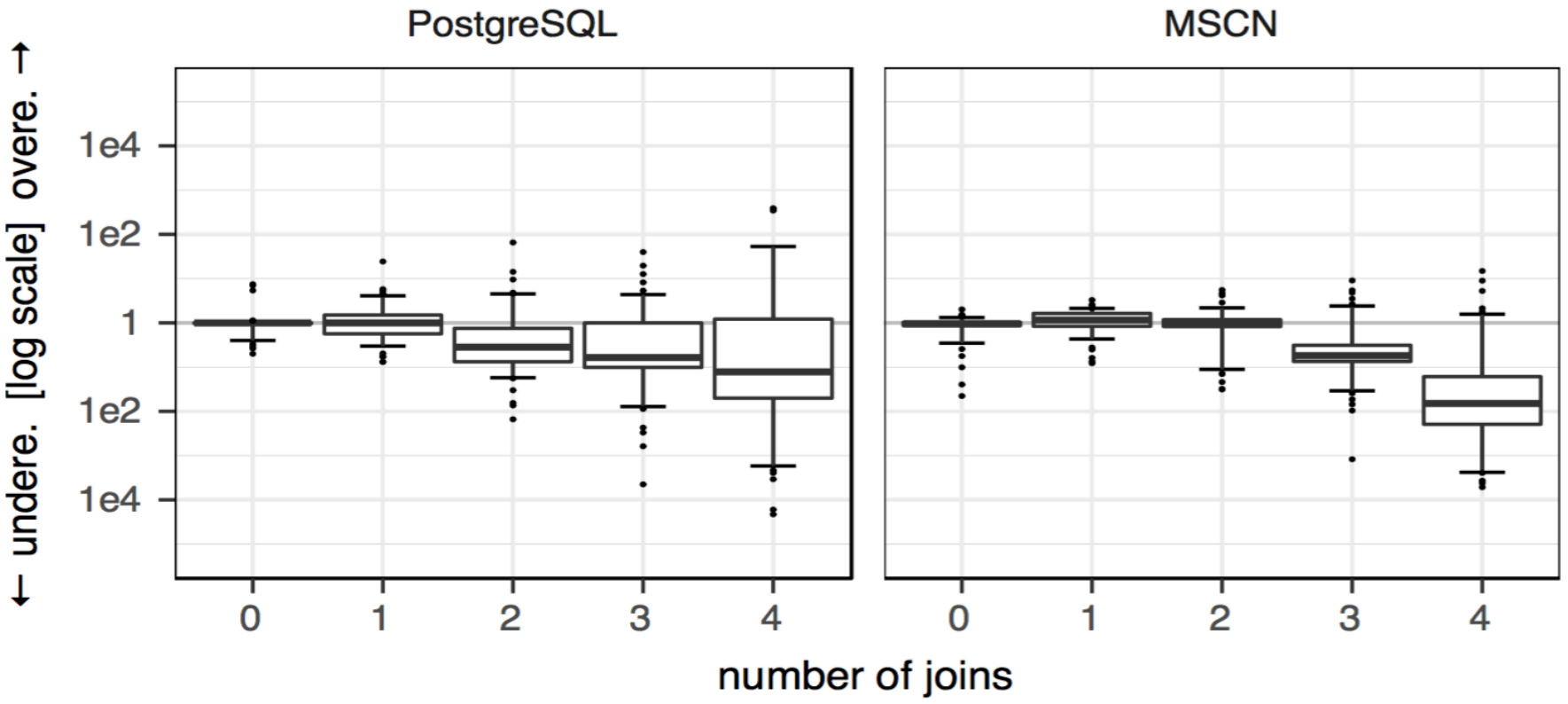
- 合成数据集上有无Sample的测试



Learned Cardinalities: Estimating Correlated Joins with Deep Learning, Andreas etc., CIDR'18

实验结果

- 在比例数据集上泛化到更多连接表



Learned Cardinalities: Estimating Correlated Joins with Deep Learning, Andreas etc., CIDR'18

实验结果

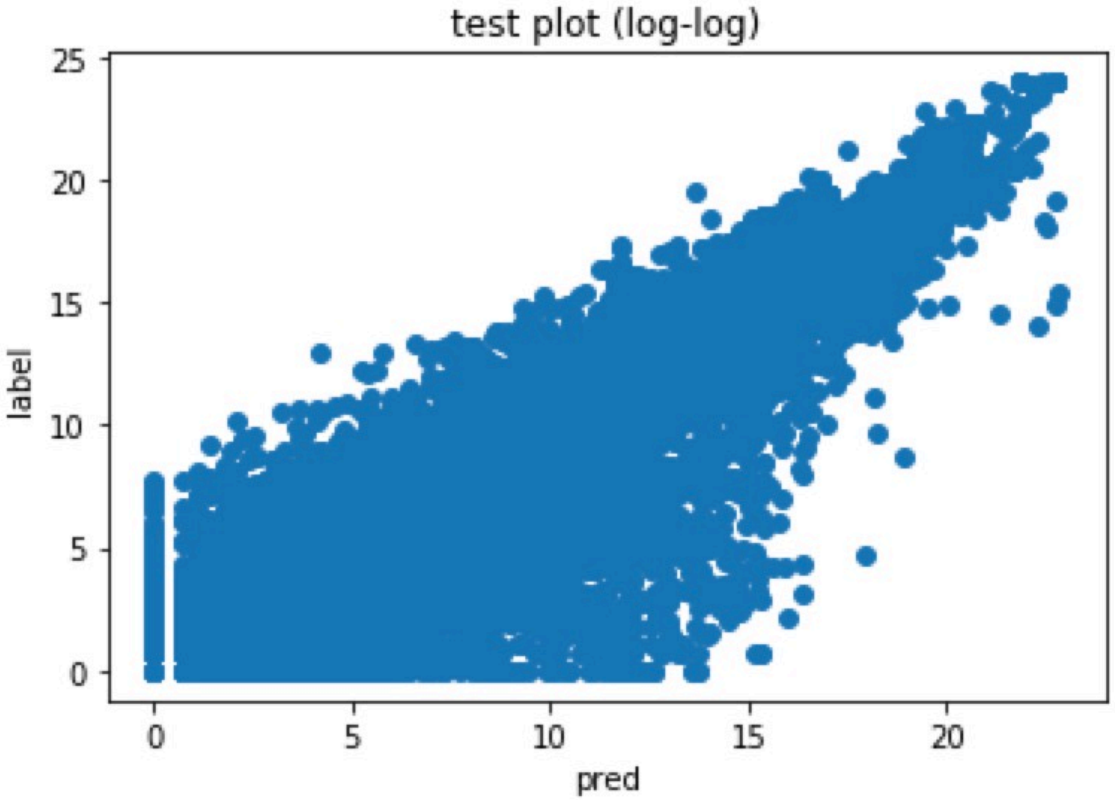
- 在JOB数据集上的测试

	median	90th	95th	99th	max	mean
PostgreSQL	7.93	164	1104	2912	3477	174
Random Samp.	11.5	198	4073	22748	23992	1046
IB Join Samp.	**1.59**	150	3198	14309	15775	590
MSCN	3.56	**31.5**	**77.6**	**415**	**676**	**27.5**

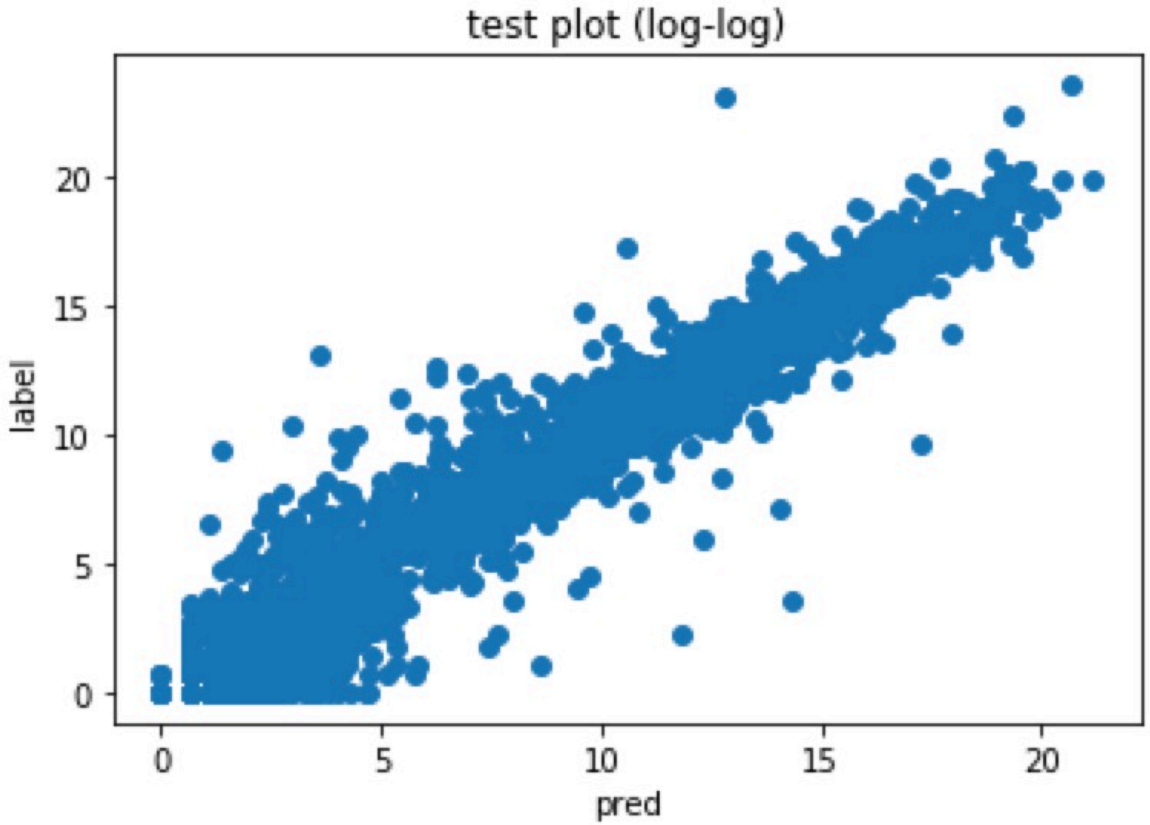
Learned Cardinalities: Estimating Correlated Joins with Deep Learning, Andreas etc., CIDR'18

实验结果

- 我的数据集



Postgres



MSCN

实验总结

- Positive
  - 效果优于单纯采样算法和Postgres中的估计算法。
  - 具有一定的泛化能力
- Negative
  - 多表泛化能力不理想。
  - JOB数据上效果不理想。
  - 0-Tuple问题不能完全解决。
  - 不支持字符串Like查询。

方案尝试

- 使用一维卷积神经网络训练，获得原始数据表高维信息以解决单纯Sample导致的0-Tuple问题，提高多表泛化能力。
- 寻找字符串类型数据的编码方式使其支持字符串查询。
- 探究模型有效的原因（学到了什么），寻找最佳的训练集生成方式，使用更少的训练集获得更好的效果。