A Sample-and-Clean Framework for Fast and Accurate Query Processing on Dirty Data

Jiannan Wang, Sanjay Krishnan, Michael Franklin, Ken Goldberg, Tim Kraska,Tova Milo

Presented by: Jinglin Peng

Image you’re a data scientist…

Average citation of the papers published in 2016?

Simple! Run a SQL query.

Image you're a data scientist…

First, let's collect data from the Internet to create a citation database.

Microsoft
Academic

Image you're a data scientist…

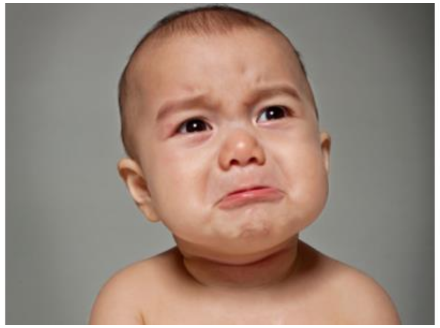
Wow! There are many errors in our collected data!

id	title	pub_year	citation
t1	CrowDB	11	18
t2	TinyDB	2005	1569
t3	YFilter	Feb,2002	298
t4	Aqua		106
t5	DataSpace	2008	107
t6	CrowER	2012	1
t7	Online Aggr.	1997	687
t8	Yfilter-ICDE	2002	298
…	…	…	…

Solution 1: No Cleaning

Directly run the query on the dirty data.

Low accuracy!



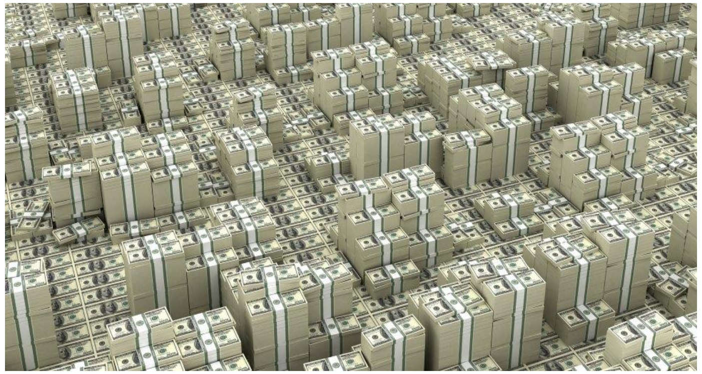
But this is what many data scientists do.

Solution 2: Full Cleaning

Clean the full data first, then make the query.

Very expensive!

Image you have TB even PB data.



Motivation

Comparison of two solutions

Solutions	Clean Time	Accuracy
No Cleaning	😄	😭
Full Cleaning	😭	😄

Can we balance the clean time and accuracy?

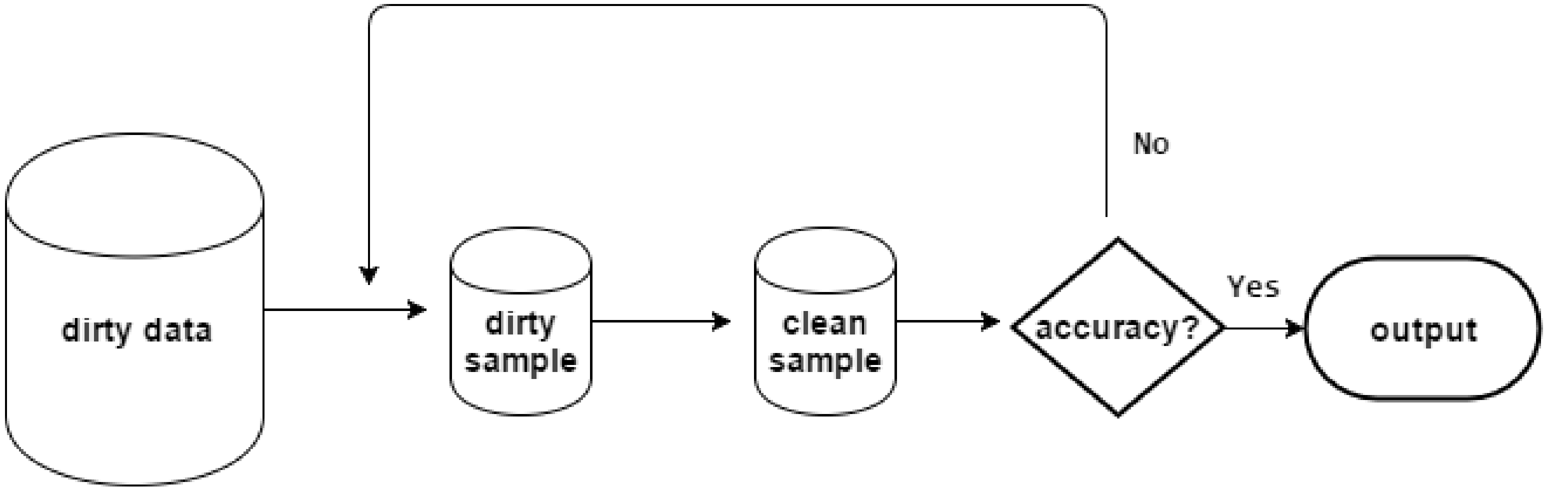
Just clean a sample!

TB, PB data

GB even MB sample

SampleClean Overview

Interactive data cleaning procedure



Our technique allows for interactive data analysis!

Problem Statement

Aggregation Queries

SELECT F(attr)
FROM table
WHERE condition
GROUP BY attrs

Supported Queries

SUM,COUNT, AVG, VAR, GEOMEAN, PRODUCT!!

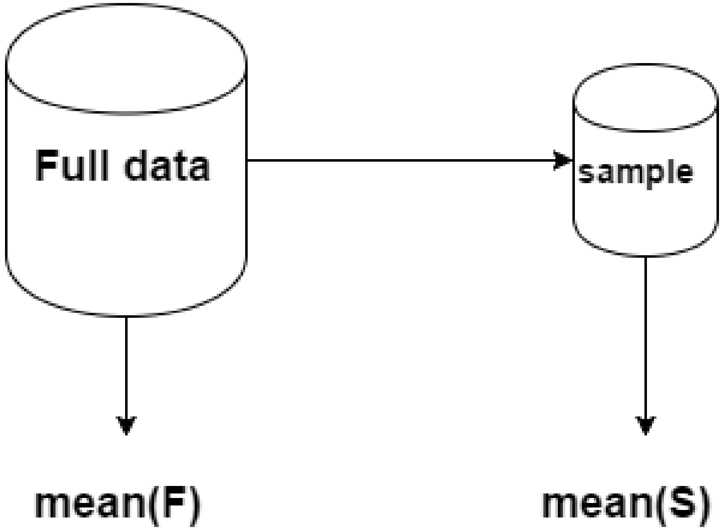
Uniform Sampling!

Key Question

Key question: how to estimate the result using the cleaned sample?

Let’s make a review of how to estimate the result using a sample-based approximate query processing (SAQP) technique.

Use sample to estimate mean value



Estimation: mean(F) $\approx$ mean(S)

Uncertainty: $\lambda\sqrt{\frac{\mathrm{var}(S)}{K}}$

Input: sample
Output: estimation & uncertainty

Example

Estimation: 500
Uncertainty: 50 (with $\lambda = 1.96$)
Explanation: the mean value of full data will fall into [500-50,500+50] within 95% prob.

Use sample to estimate sum & count

How to estimate sum & count?

count is a special case of sum.

sum/count can be treated as estimating a mean value after some transformation.

Use $\phi(t)$ to transform tuple t.

Example of estimating sum

Query: sum of the citations of the papers published after 2007.

$$\phi_{sum}(t) = \Pr edicate(t) \bullet N \bullet t[a]$$

Full data

id	title	pub_year	citation	predicate	ϕ
t1	CrowDB	2011	144	True	144*6
t2	TinyDB	2005	1569	False	0
t3	YFilter	2002	298	False	0
t4	Aqua	1999	106	False	0
t5	DataSpa ce	2008	107	True	107*6
t6	CrowER	2012	34	True	34*6

Sample

id	title	pub_year	citation	predicate	ϕ
t2	TinyDB	2005	1569	False	0
t5	DataSpa ce	2008	107	True	107*6
t6	CrowER	2012	34	True	34*6

Real result

mean(144*6+0+0+0+107*6+34*6**)**

Estimation

mean(0+107*6+34*6**)**

Uncertainty

$$1.96\sqrt{\frac{\mathrm{var}(0,107*6,34*6)}{3}}$$

Challenging Problem

If data has no errors, the sampling method gives an unbiased estimation.

What if data has errors?

Three Type of Errors

Query: average citation of paper published after 2000.

Dirty Data

Condition Error

Value Error

Duplication Error

P	id	title	pub_year	citation
1/6	t1	CrowDB	11	144
1/6	t2	TinyDB	2005	1
1/6	t3	YFilter	2002	298
1/6	t4	Aqua	1999	106
1/6	t5	Yfilter-ICDE	2002	298
1/6	t6	CrowER	2012	34

Duplication increases the prob. of ‘Yfilter’ to be sampled!

Correction of Errors

We need to correct the impact of duplication error!

Down weight of duplication tuples.

Derive equation:

$$\phi(t) = \frac{c[t]}{d_t} \Pi(condition)$$

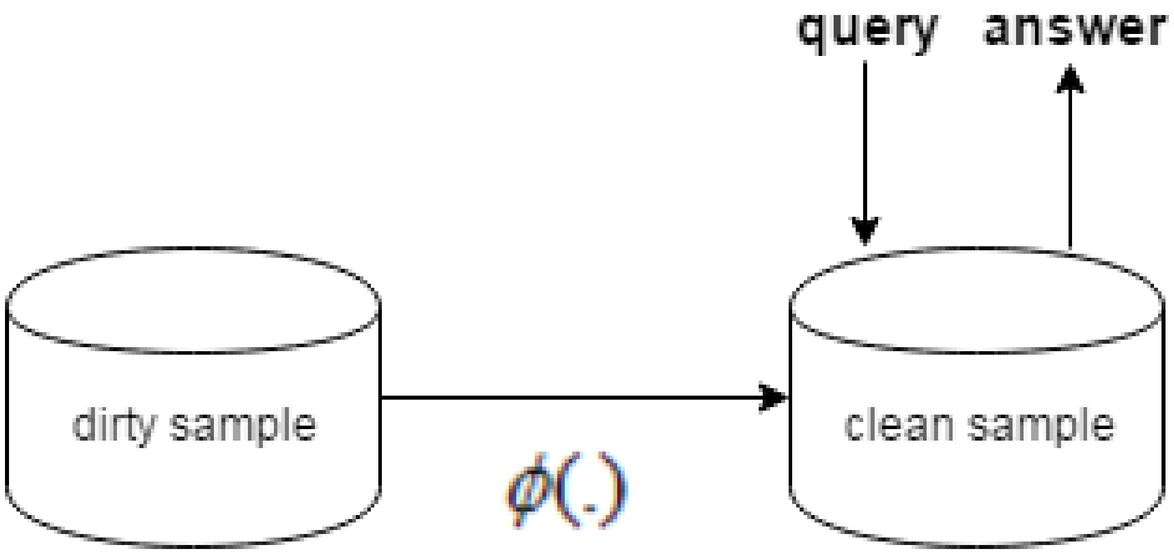
Value Error

Duplication Error

Condition Error

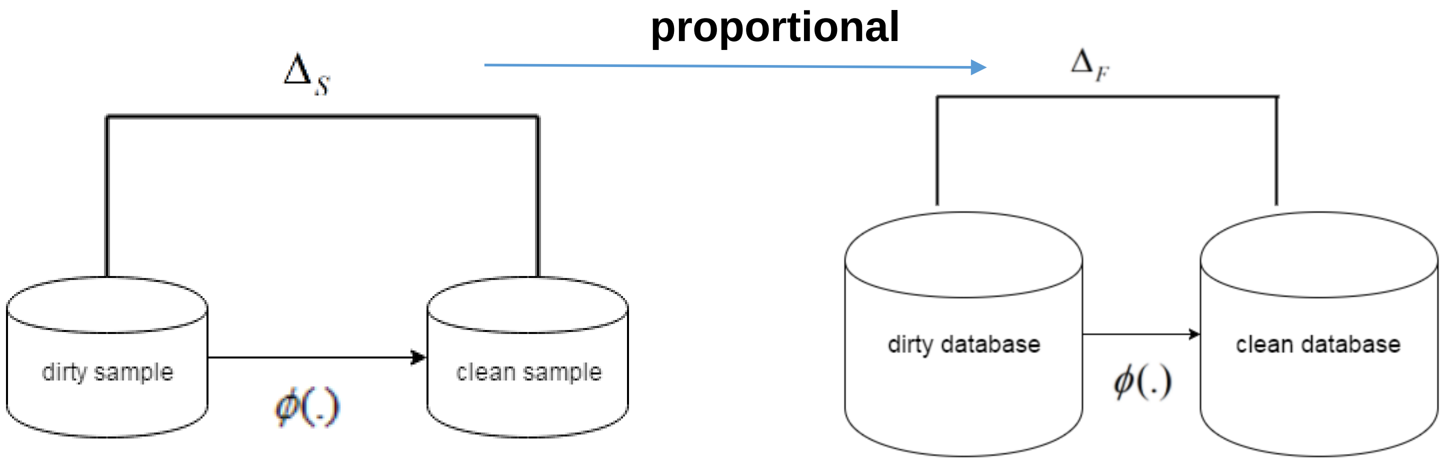
Algo. 1 RawSC Estimation

Query on the cleaned sample to get the estimation



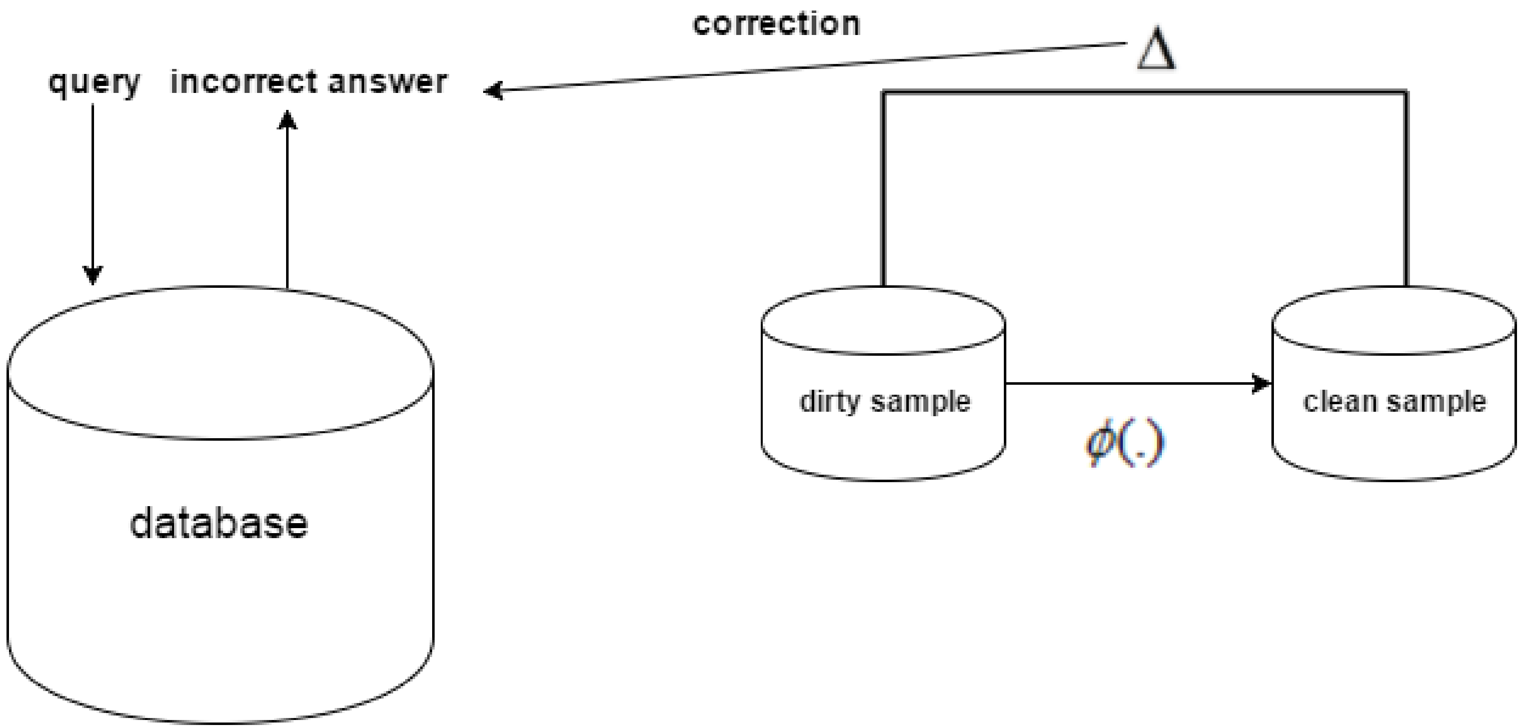
Algo. 2 NormalizedSC Estimation

How much did the cleaning change the data?



Can we query on full dirty data and use cleaned sample to correct the result?

Algo. 2 NormalizedSC Estimation



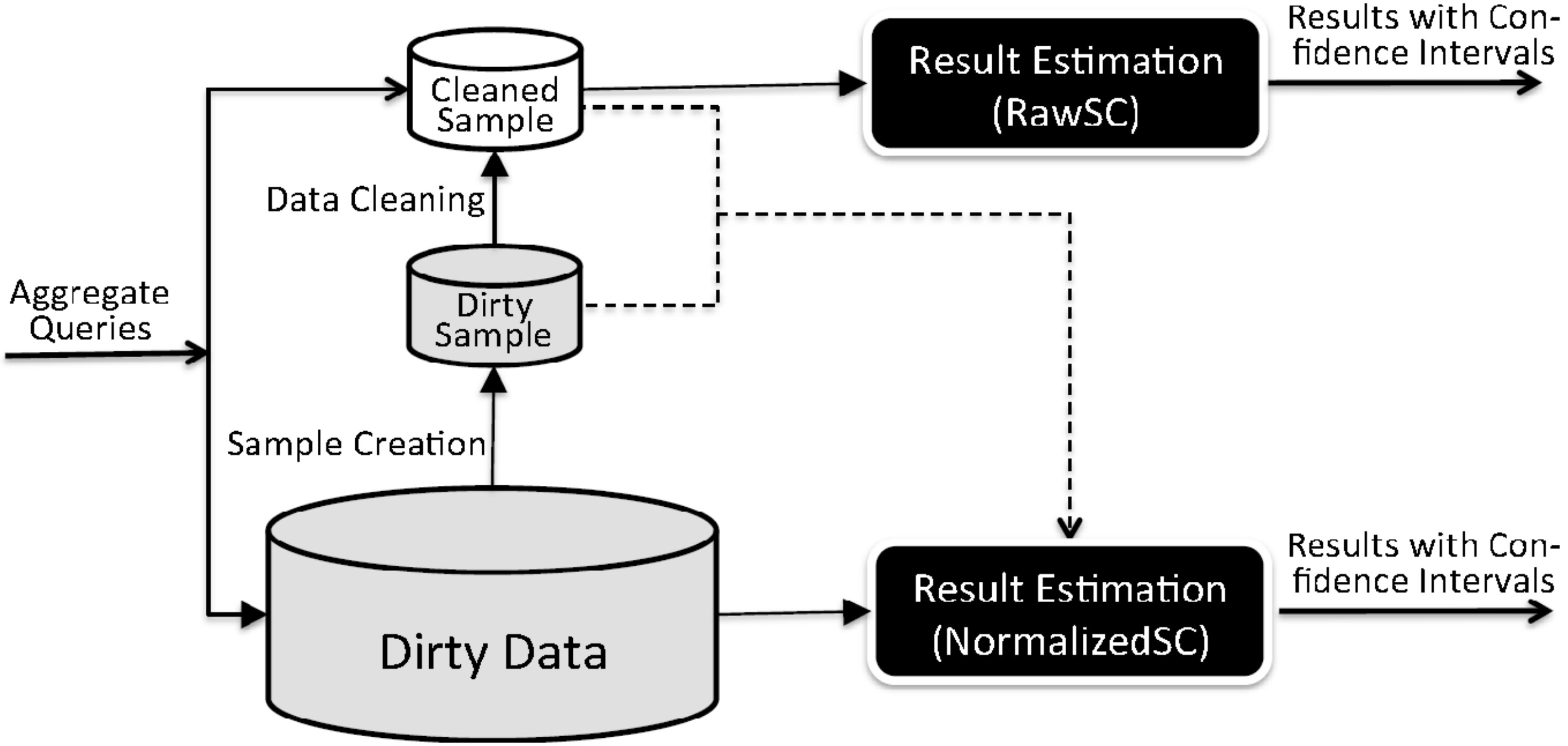
RawSC vs. NormalizedSC

Comparison of Two Methods

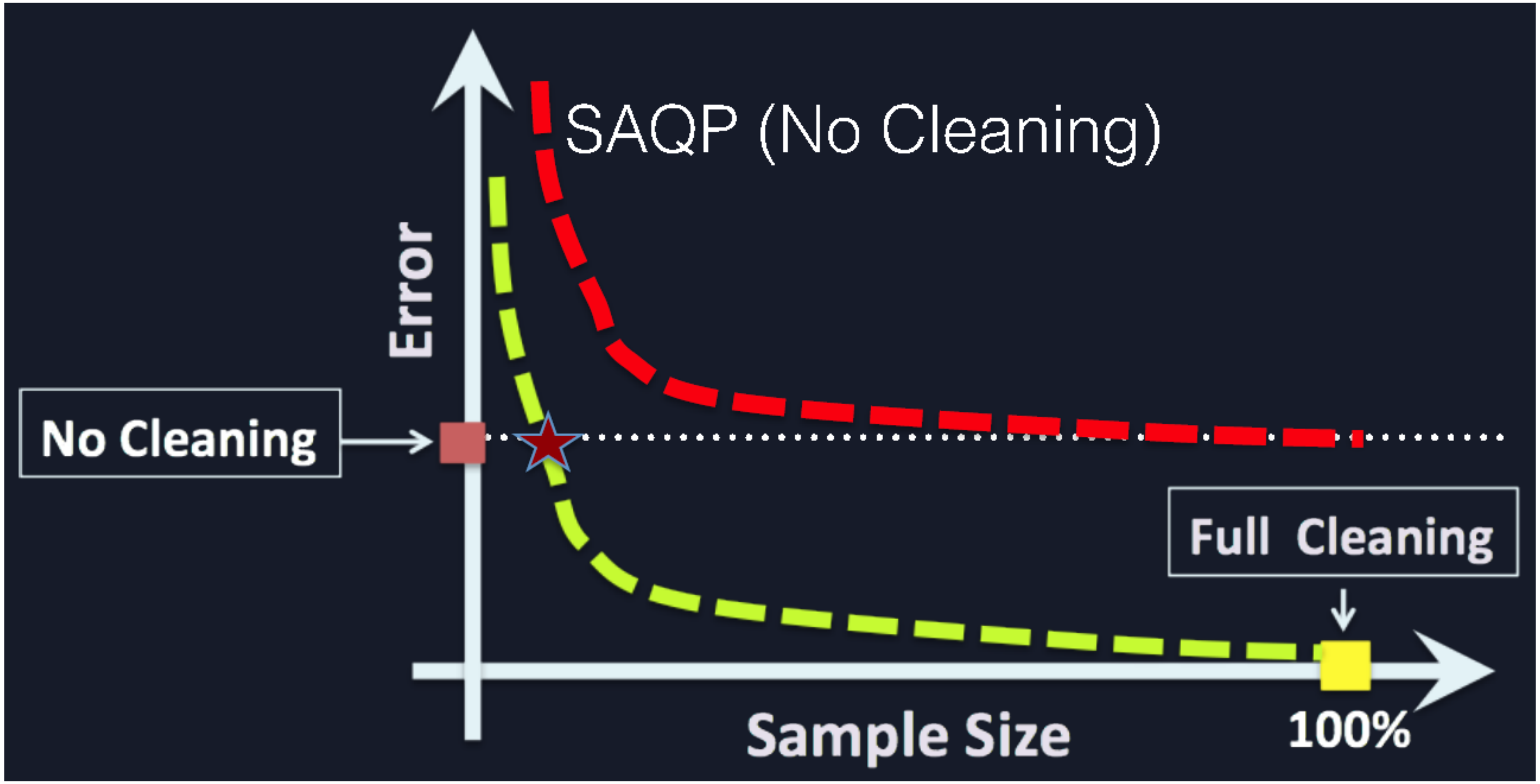
Method	RawSC	NormalizedSC
Idea	Clean Estimation	Dirty Correction
Error	$\frac{\mathrm{var}(\phi)}{k}$	$\frac{\mathrm{var}(\Delta)}{k}$
Runtime	$O(k)$	$O(n)$
Query Data	Sample	Full Data

SampleClean Framework

SampleClean will chose the better result from RawSC and NormalizedSC as final estimation.

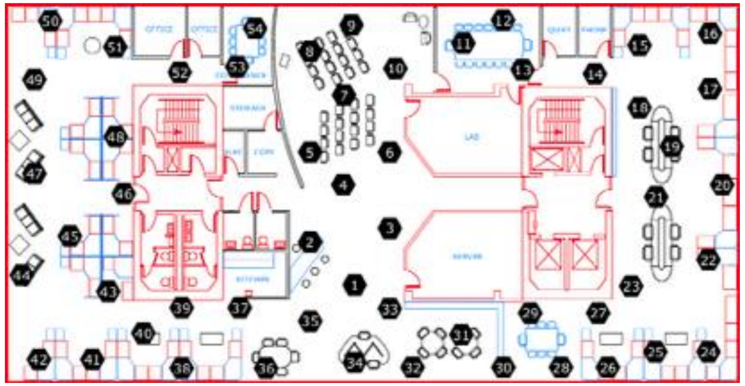


SampleClean: tradeoff



Experiments

- Microsoft Academic Search (1374 records)
- Intel Wireless Sensor Dataset (44,460 records)
- TPC-H: Simulated Errors (6M records)



Exp. 1 Academic Ranking

What’s the ranking of three authors?

Rakesh Agrawal
Microsoft
Publications: 353 | Citations: 33537
Fields: Databases, Data Mining, World Wide Web
Collaborated with 365 co-authors from 1982 to 2012 | Cited by 24220 authors

Jeffrey D. Ullman
Stanford University
Publications: 460 | Citations: 43431
Fields: Databases, Algorithms & Theory, Scientific Computing
Collaborated with 317 co-authors from 1961 to 2012 | Cited by 31987 authors

Michael Franklin
University of California Berkeley
Publications: 561 | Citations: 15174
Fields: Databases, Pharmacology, Data Mining
Collaborated with 3451 co-authors from 1974 to 2012 | Cited by 15795 authors

Exp. 1 Academic Ranking

Microsoft Academic Search Dataset

Total: 1374 Records

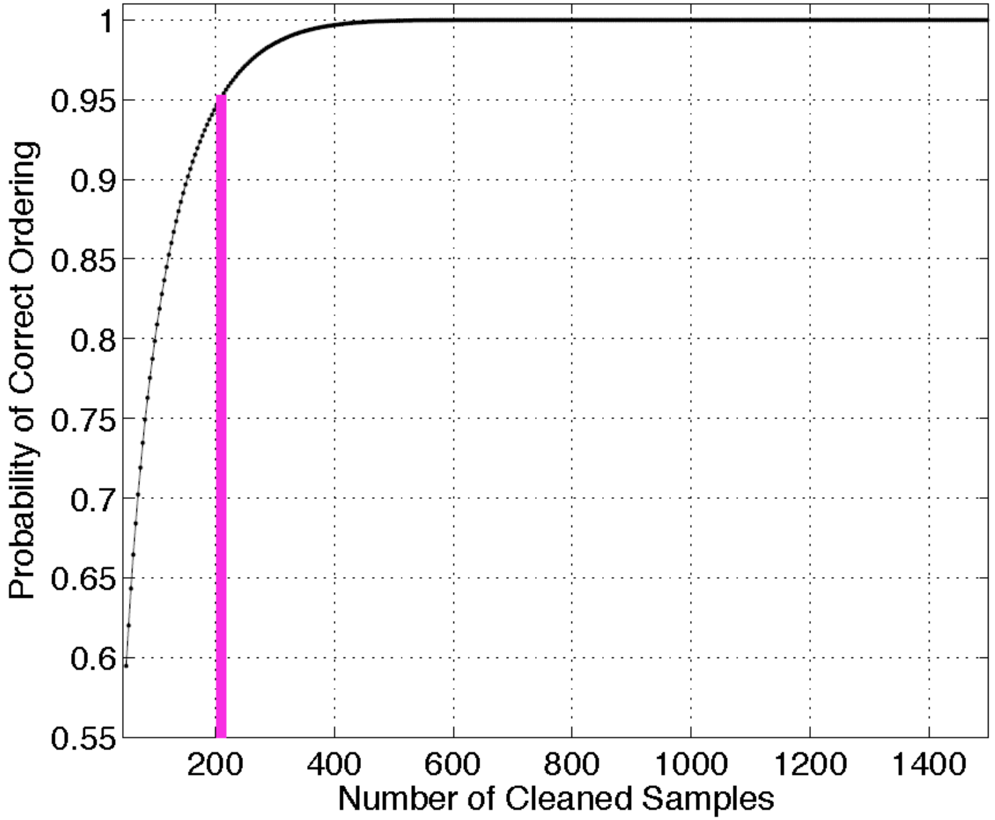
Author	Dirty	Clean
Rakesh Agarwal	353	211
Jeffrey Ullman	460	255
Michael Franklin	561	173

Ranking based on dirty data: Michael, Jeffrey, Rakesh

Ranking based on clean data: Jeffrey, Rakesh, Michael

Exp. 1 Academic Ranking

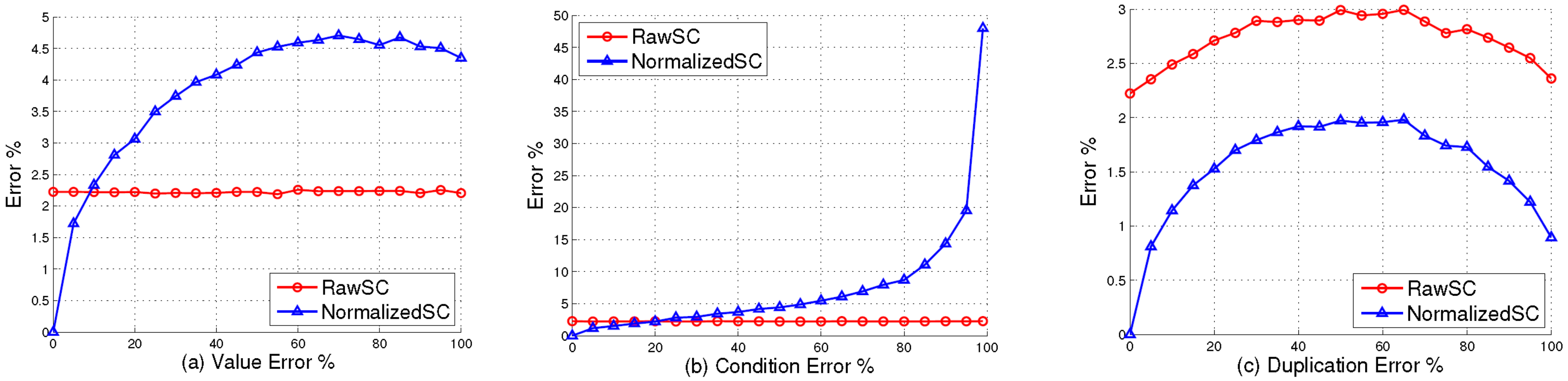
Dataset: Microsoft Academic Search (1374)
Query type: COUNT
Sample counts: 10,000



Cleaning 210 out of 1374 can rank correctly within 95% prob.

Exp. 2 RawSC vs. NormalizedSC

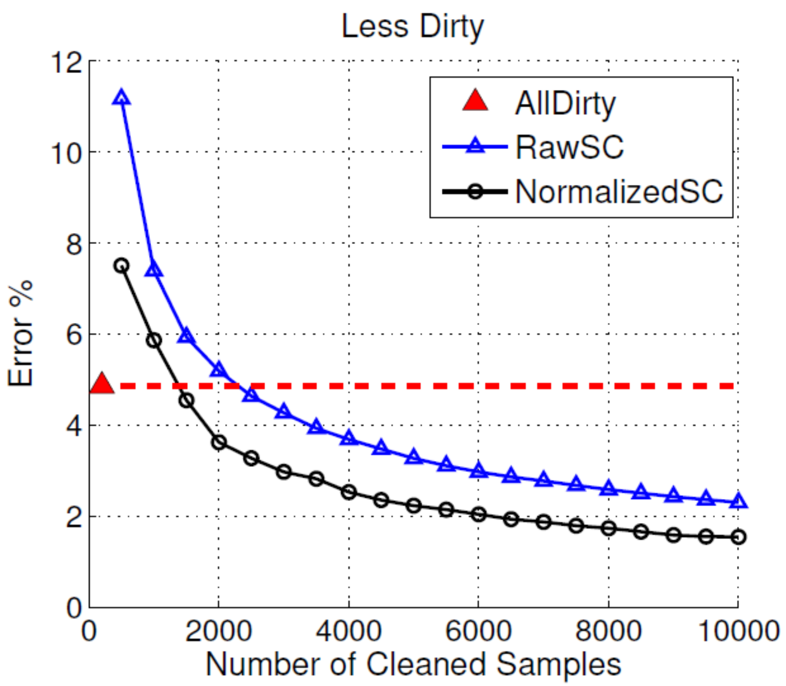
Dataset: TPC-H benchmark (6M)
Query type: AVG
Sample size: 0.01M, 0.17% of 6M

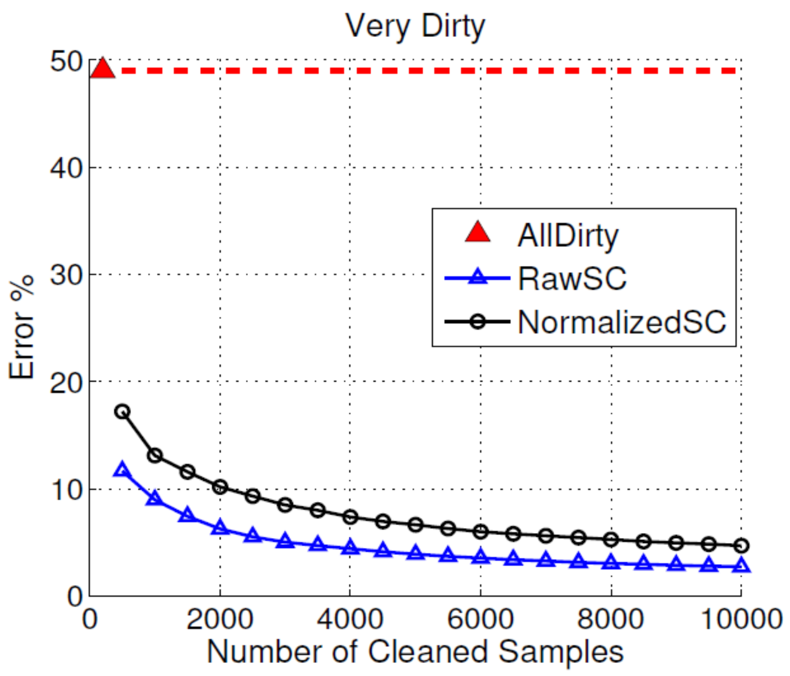


1. RawSC works better when value error or condition error is large.
2. NormalizedSC works better when value error or condition error is small, or when data has duplication error.

Exp. 3 Clean Cost vs. Result Quality

Dataset: TPC-H benchmark (6M)
Query type: AVG
Less Dirty: 3% value, 1% condition, and 2% duplication errors
Very Dirty: 30% value, 10% condition, and 20% duplication errors

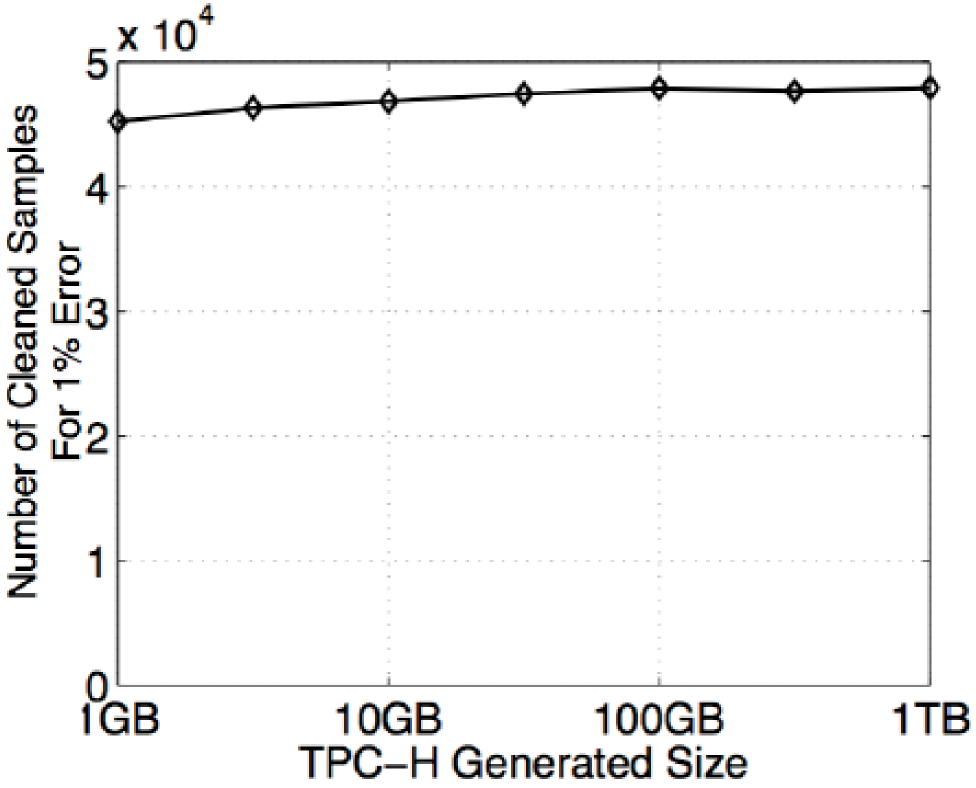




1. **Both methods converge at a rate $\frac{1}{\sqrt{K}}$.**
2. **There will always be a single *better* choice between two methods.**
3. **Both methods are better than *AllDirty* by cleaning a really small sample.**

Exp. 4 Scalability of Cleaning Cost

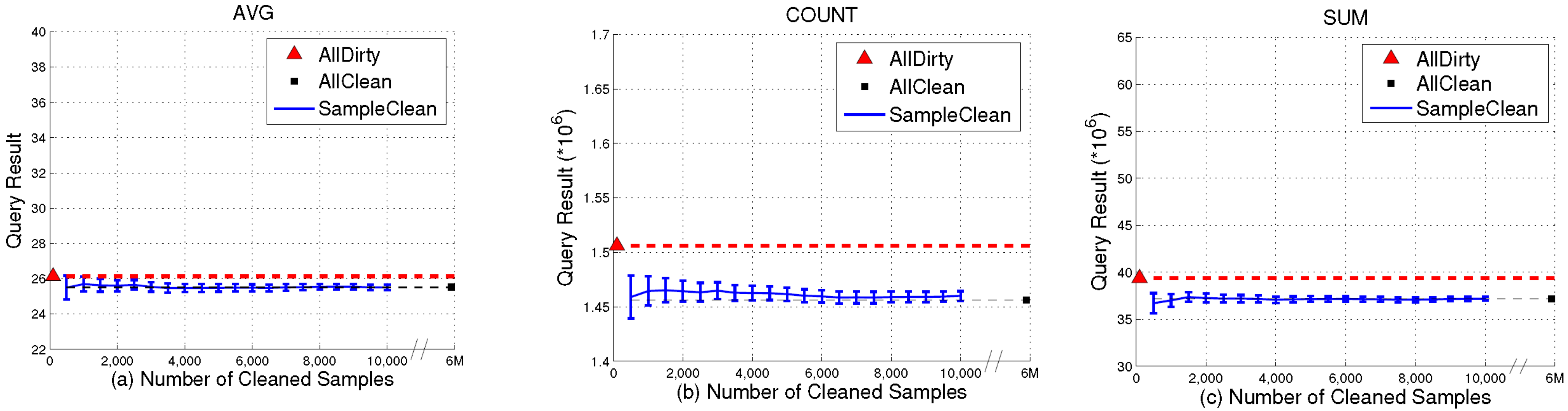
Dataset: TPC-H benchmark (6M)
Query type: AVG
Error: 30% value, 10% condition, and 20% duplication errors



The number of cleaned tuples needed to achieve a certain error doesn't increase with data size.

Exp. 5-1 End-to-End (Less Dirty)

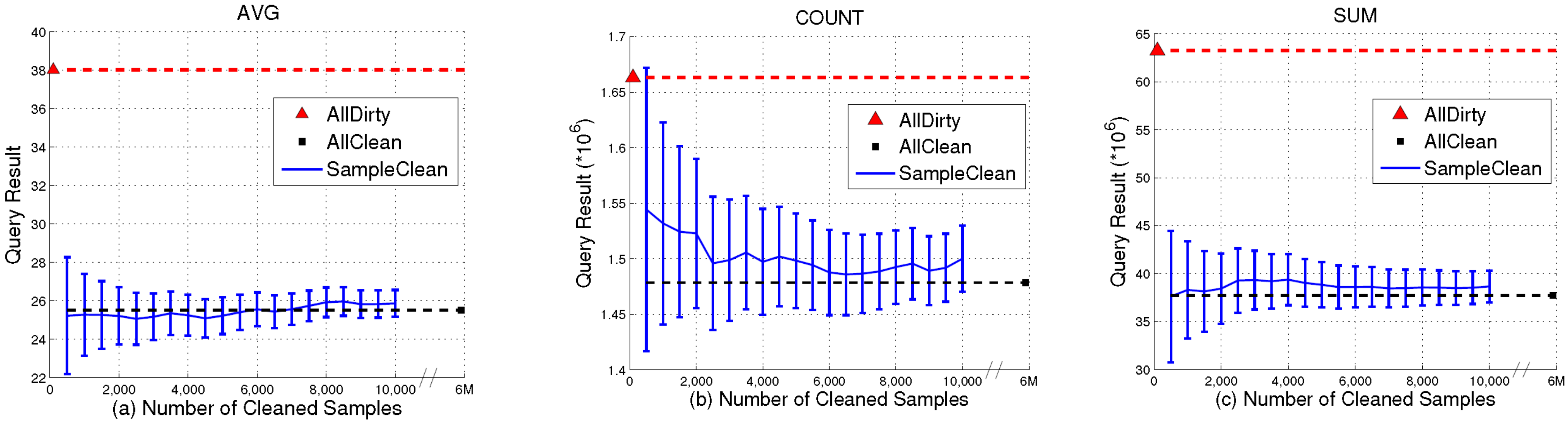
Dataset: **TPC-H benchmark (6M)**
Query type: **AVG, COUNT and SUM**
Error: **3% value, 1% condition, and 2% duplication errors**



1. **After cleaning only 1000 tuples (0.016%), *SampleClean* is better than *AllDirty*.**
2. ***SampleClean* quickly converges to the right answer.**
3. ***SampleClean* provides a tradeoff of cleaning time & result quality.**

Exp. 5-2 End-to-End (Very Dirty)

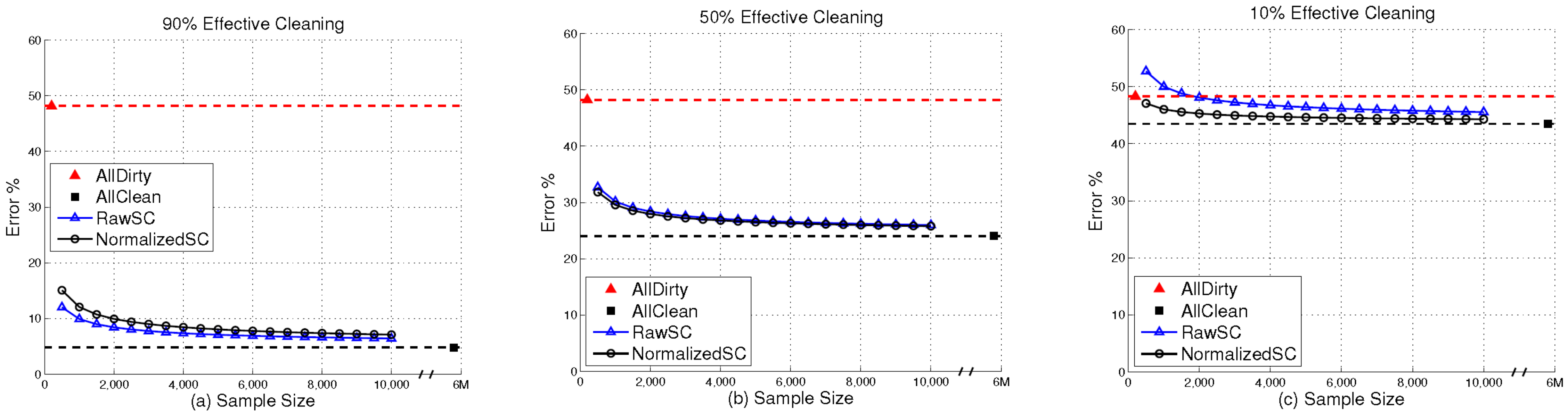
Dataset: **TPC-H benchmark (6M)**
Query type: **AVG, COUNT and SUM**
Error: **30% value, 10% condition, and 20% duplication errors**



1. ***SampleClean*** **works well when data error is large.**
2. **For all queries, the estimation is within 5% of *AllClean* after cleaning only 5000 tuples (0.08%).**

Exp. 6 Imperfect Cleaning

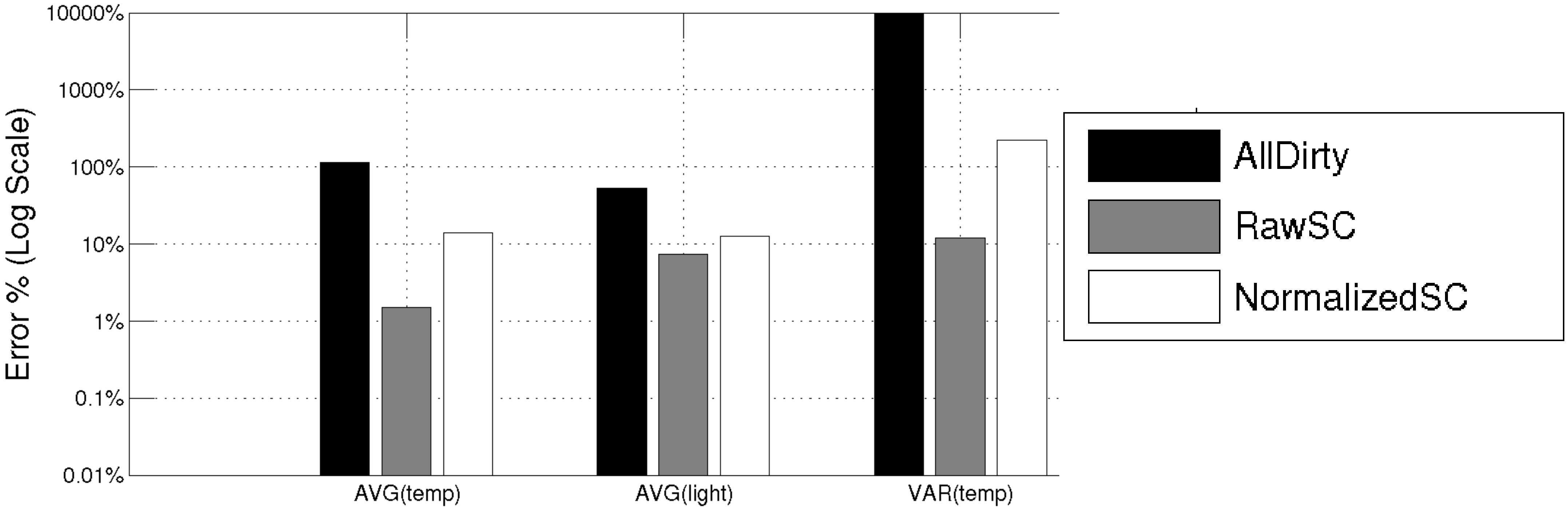
Dataset: **TPC-H benchmark (6M)**
Query type: **AVG**
Error: **30% value, 10% condition, and 20% duplication errors**



1. ***SampleClean*** **converges to real value quickly.**
2. **A 10% effective cleaning module can be accurate than AllDirty after cleaning 2000 tuples (0.03%).**

Exp. 7 Evaluation on Sensor Dataset

Dataset: **Sensor Dataset (44,460)**
Sample size: **500 (1.12%)**



1. **The query quality of *AllDirty* is really bad.**
2. **Error of our method is less 10% even when data error is orders of magnitude higher.**

Conclusion

- **SampleClean can improve query quality by cleaning a small sample.**

- **SampleClean provides an unbiased estimation for full clean data.**

- **SampleClean allows for interactive analysis on dirty data.**

Thank you!