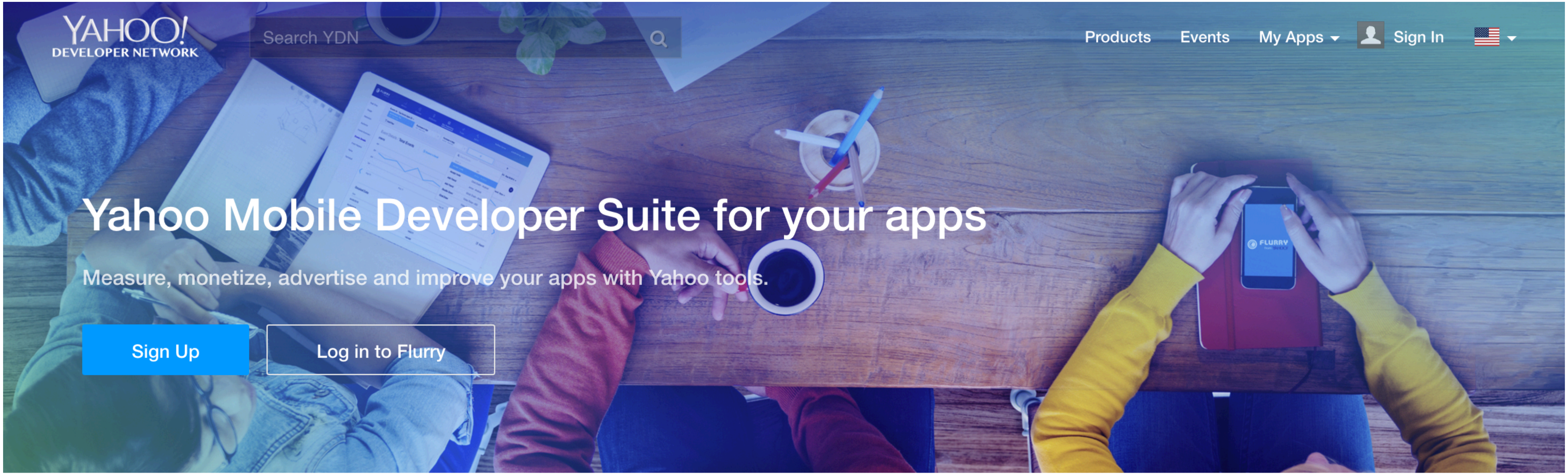
STRATIFIED SAMPLING MEETS MACHINE LEARNING

KEVIN LANG
KONSTANTIN SHMAKOV
EDO LIBERTY

YAHOO!



Flurry Analytics

Get free insights from the industry's leading mobile app analytics tool.

Yahoo App Publishing

Monetize your app with native and video ads from Yahoo, Flurry, and BrightRoll advertisers.

Yahoo App Marketing

Reach your target audience with the Gemini native and video marketplace.

YAHOO!

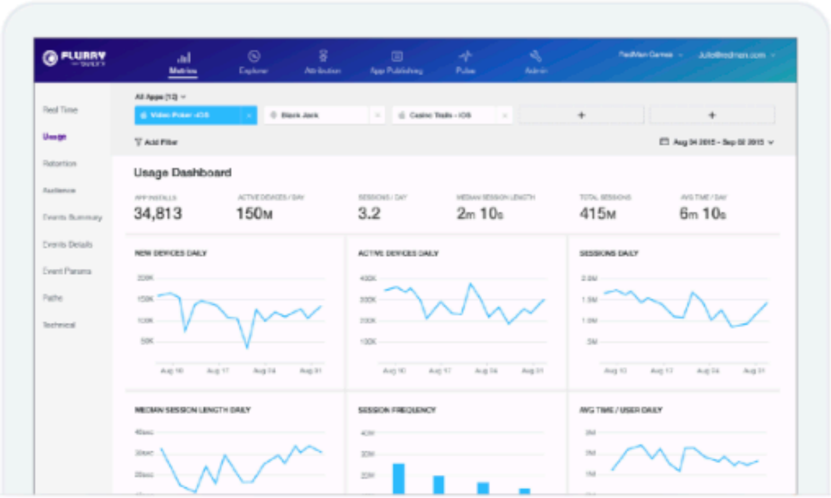
Introducing the all new Flurry Analytics

Measure and analyze activity across your app portfolio to answer your hardest questions and optimize your app experience.

Sign Up

Log in to Flurry

Documentation



Flurry Grows with You

As your business grows, we are committed to supporting you at scale for free. Empower all your employees to leverage analytics for data driven decision making.

Product Features

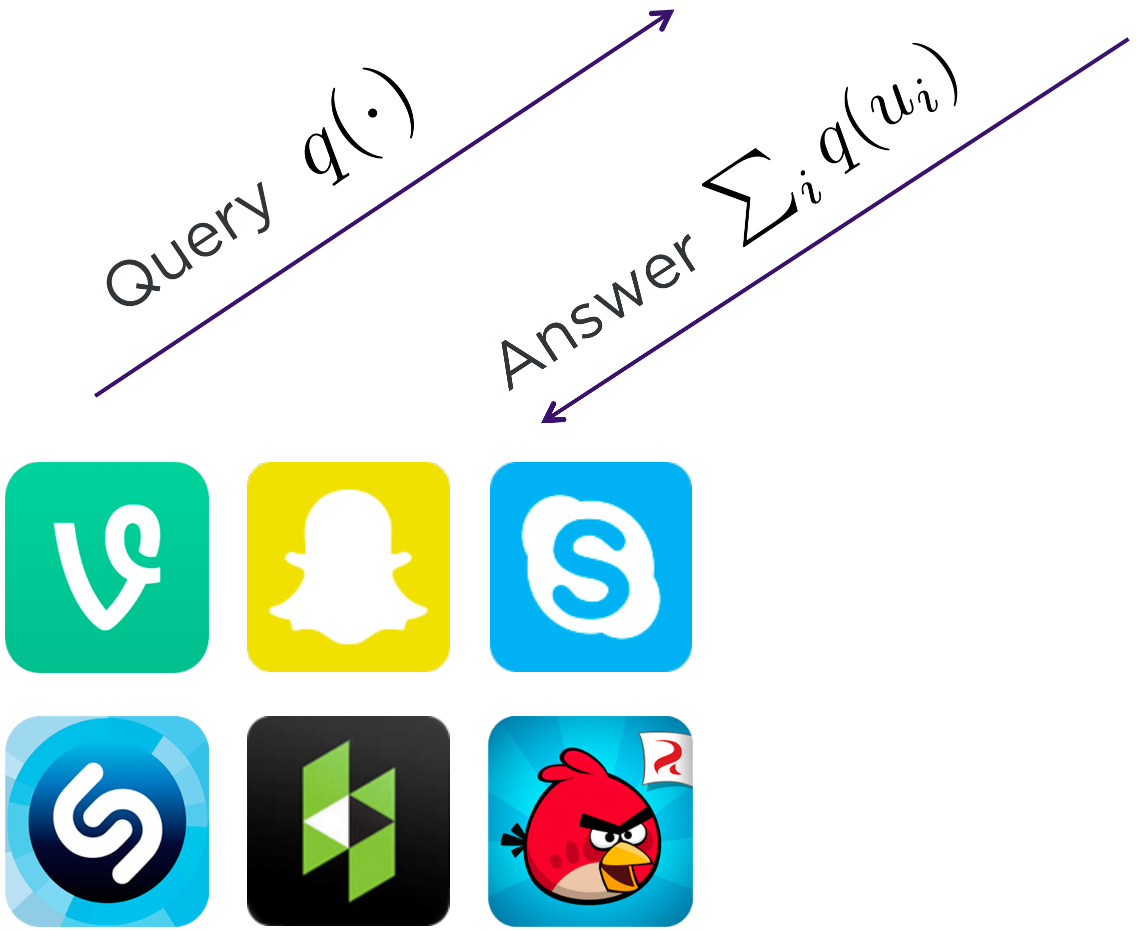
Flurry Analytics

Apps with Flurry SDK

Flurry Analytics

Events

Flurry Analytics



Examples:
- Number of event of a certain type
- Number of unique user
- Number of unique users in a specific day
- Total time spent in certain geo
- Average $ spent by age

SAMPLING

Challenges:

1. The data is very large. Computing $\sum_i q(u_i)$ exactly is too costly.
2. The function $q(\cdot)$ is user specified and completely unconstrained.

Good News:

And approximate answer is acceptable (if the error is small)

Solution:

Estimate the answer on a random subset of the records

YAHOO!

NOTATIONS

- $q_i := q(u_i)$ for brevity
- $y := \sum_i q_i$ the exact answer for the query q
- p_i the probability of choosing record i
- S the set of sampled records, each chosen with probability p_i
- $\tilde{y} = \sum_{i \in S} q_i / p_i$ the Horvitz-Thompson estimator for y

PROPERTIES

- $\mathbb{E}[\tilde{y} - y] = 0$ Horvitz-Thompson **estimator is unbiased**
- $\sigma[\tilde{y} - y] \leq y\sqrt{1/(\zeta \cdot \mathrm{card}(q))}$ **its standard deviation isn't large**

$$\zeta = \min_i p_i \qquad \mathrm{card}(q) := \sum |q_i| / \max |q_i|$$

- $\Pr[|\tilde{y} - y| \geq \varepsilon y] \leq e^{-O(\varepsilon^2 \zeta \cdot \mathrm{card}(q))}$ **probability for large error is small**

$$\boxed{\mathrm{card}(q) \sim \Omega(n) \quad \rightarrow \quad |S| \sim 1/\varepsilon^2}$$

(Olken, Rotem and Hellerstein 1986, and 1990) application to databases
(Acharya, Gibbons, Poosala 2000) uniform sampling is best in the worst case

STRATIFIED SAMPLING

- Sample = 100,000 US individuals.
- Query = Republicans vs. Democrats in American Samoa?

American Samoa is 0.02% of population → Only ~20 from Samoa in the sample → Survey error is very large!

If $\mathrm{card}(q)$ is small $|S|$ must be large

- Sample <u>different *strata*</u> (e.g. US territories) with <u>different probabilities</u>.

(Neyman, Jerzy 1934)

DBLP EXAMPLE

Choosing the right strata is hard!

- 2,101,151 papers
- 1000 most populous venues
- Query example

    title contains "learning" and # authors <= 3
    title contains "mechanism" and year > 2004

What is the right stratification here?

- Stratifying by venue made things worse!
- Stratifying by year was better but still worse than uniform sampling.

YAHOO!

SAMPLING, STRATIFICATION, AND DATABASES

- Design strata that minimize worst case variance on possible queries
- Linearly combine strata based on record features
- Combine stratifies and uniform sampling: Congressional Sampling
  - Acharya, Gibbons, Poosala 2000:

<u>Important idea:</u> consider past queries to the database!

- Each stratum is a set of records that agree on all queries
  - Chaudhuri, Das and Narasayya 2007: optimize for the query log

- Split to two strata, per each query. Take linear combinations
  - Joshi, Jermaine, 2008: linear combinations of stratified probabilities

OUR APPROACH

- Assume queries are drawn from a distribution $\mathbb{Q}$
- Use the query log Q as a "training set" (assumed w.r.t. $\mathbb{Q}$)
- Allow each record to be sampled with a different probability p_i
- Minimize the Risk $\mathbb{E}[(\tilde{y} - y)^2]$
- This translates to $\mathbb{E}_{q \sim \mathbb{Q}} \sum_i q_i^2 (1/p_i - 1)$

unknown

YAHOO!

OUR APPROACH

- ERM: Minimize $$\sum_{q \in Q} \sum_i q_i^2 (1/p_i - 1)$$

Query log

- Sampling budget $$\sum_i p_i c_i \leq B$$ $(\sum_i c_i \ll B)$

- Regularization $\forall\, i \;\; p_i \in [\zeta, 1]$ $(\zeta \leq B / \sum_i c_i)$

YAHOO!

OUR APPROACH

- Solve with Lagrange multipliers

$$\max_{\alpha,\beta,\gamma}[\frac{1}{|Q|}\sum_{q\in Q}\sum_i q_i^2(1/p_i - 1) - \sum_i \alpha_i(p_i - \zeta)$$
$$- \sum_i \beta_i(1 - p_i) - \gamma(B - \sum_i p_i c_i)]$$

- By KKT conditions

$p_i = \zeta$ or $p_i \propto \sqrt{\frac{1}{c_i}\frac{1}{|Q|}\sum_{q\in Q} q_i^2}$ or $p_i = 1$

OUR APPROACH

input: training queries Q,
budget B, record costs c,
regularization factor $\eta \in [0, 1]$
$\zeta = \eta \cdot (B / \sum_i c_i)$
$\forall\, i \;\; z_i = \sqrt{\frac{1}{c_i} \frac{1}{|Q|} \sum_{q \in Q} q_i^2}$
Binary search for λ satisfying $\sum_i c_i \, \textsc{clip}_\zeta^1(\lambda z_i) = B$
output: $\forall\, i \;\; p_i = \textsc{clip}_\zeta^1(\lambda z_i)$

$$\mathrm{Risk}(p) \leq \mathrm{Risk}(p^*) \left(1 + O\left(\mathrm{skew} \sqrt{\frac{\log(n/\delta)}{|Q|}}\right)\right)$$

Alg' Risk

Best Risk

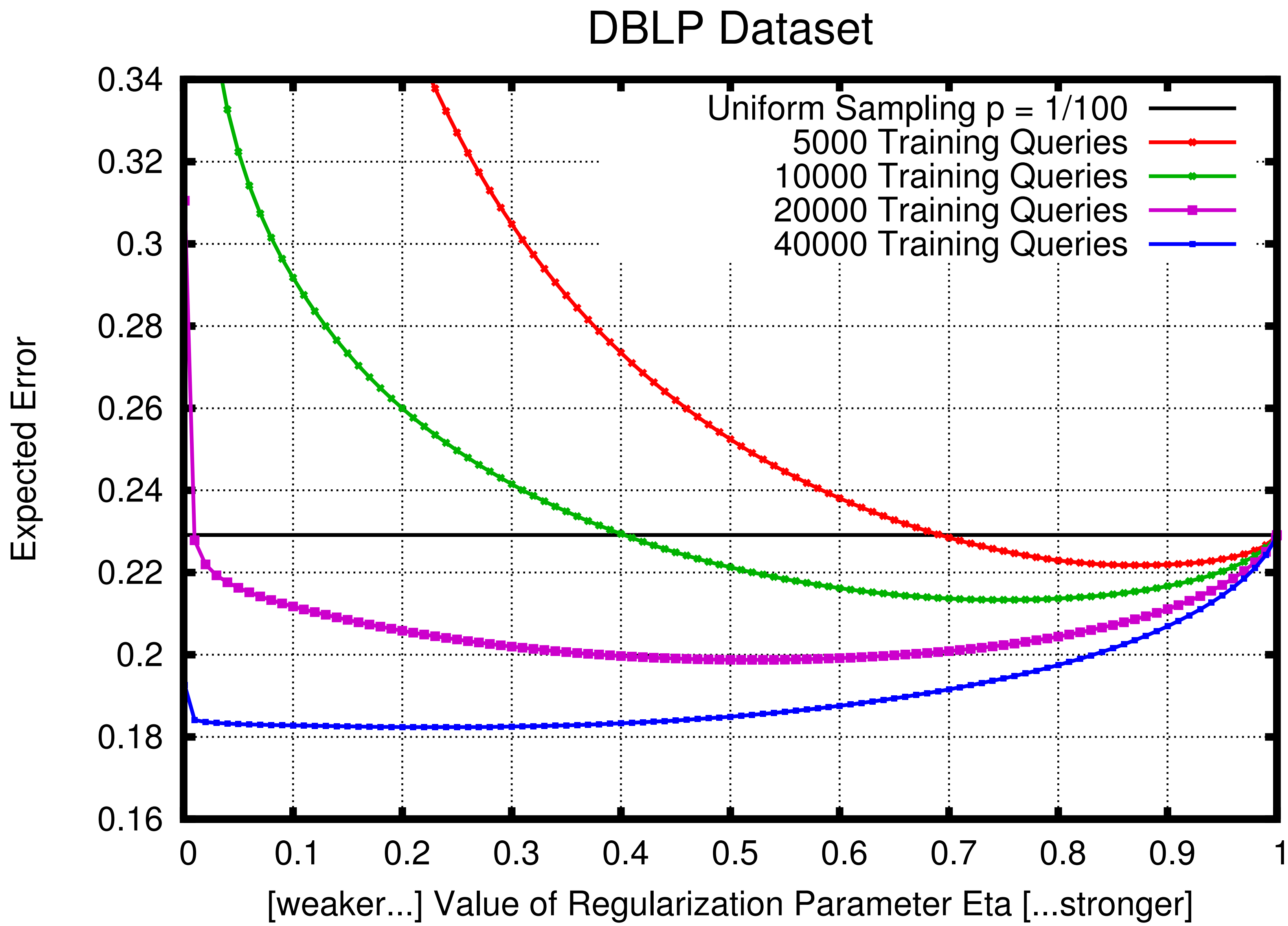
Database “badness”

Training Set size

RESULTS

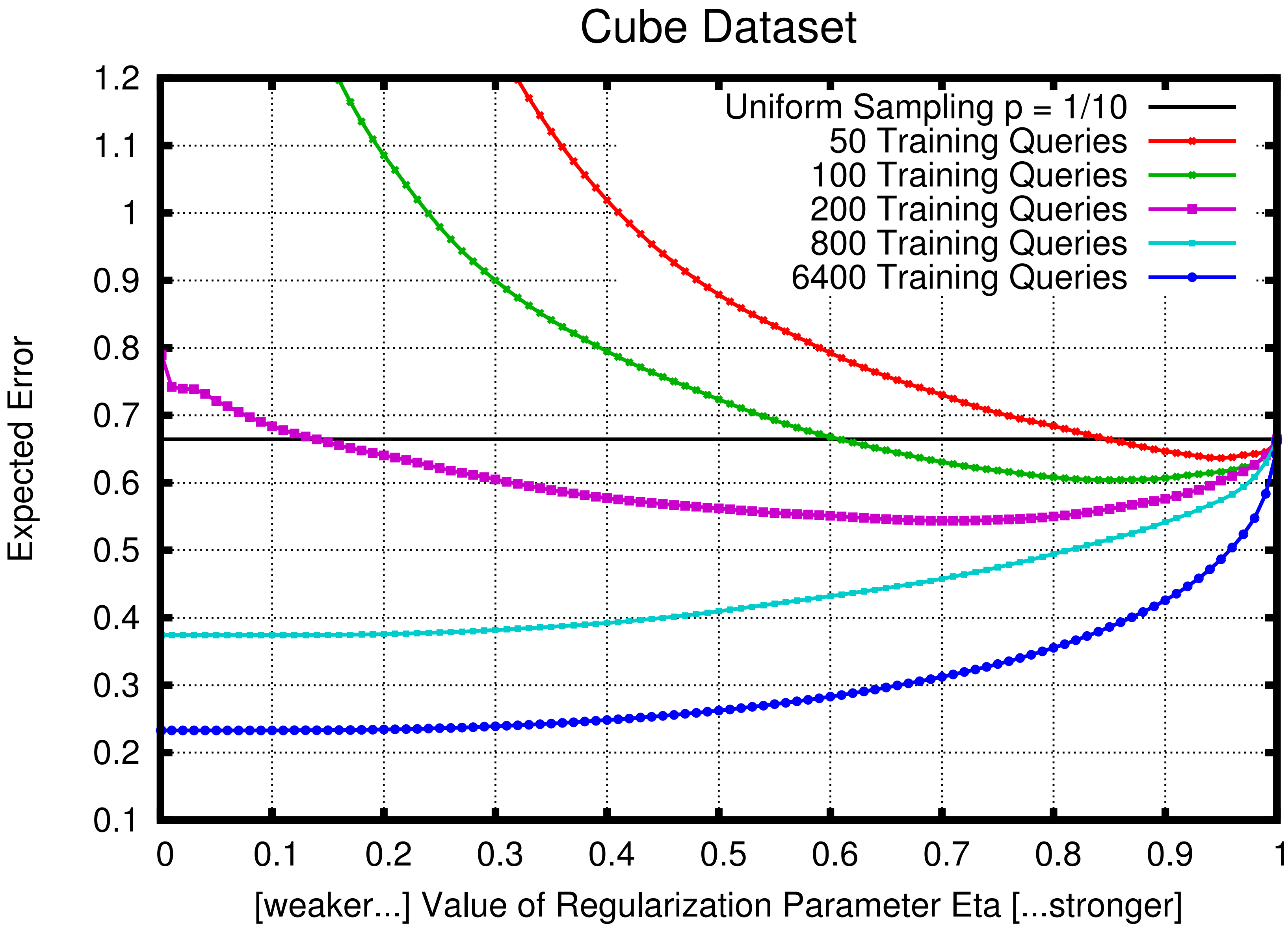
Dataset Sampling Rate	Cube 0.1	DBLP 0.01	YAM+ 0.01
Uniform Sampling	0.664	0.229	0.104
Neyman Allocation	0.643	0.640	0.286
Regularized Neyman	0.582	0.228	0.102
ERM-η, small training set	0.637	0.222	0.096
ERM-ρ, small training set	0.623	0.213	0.092
ERM-η, large training set	**0.233**	0.182	0.064
ERM-ρ, large training set	**0.233**	**0.179**	**0.059**

RESULTS

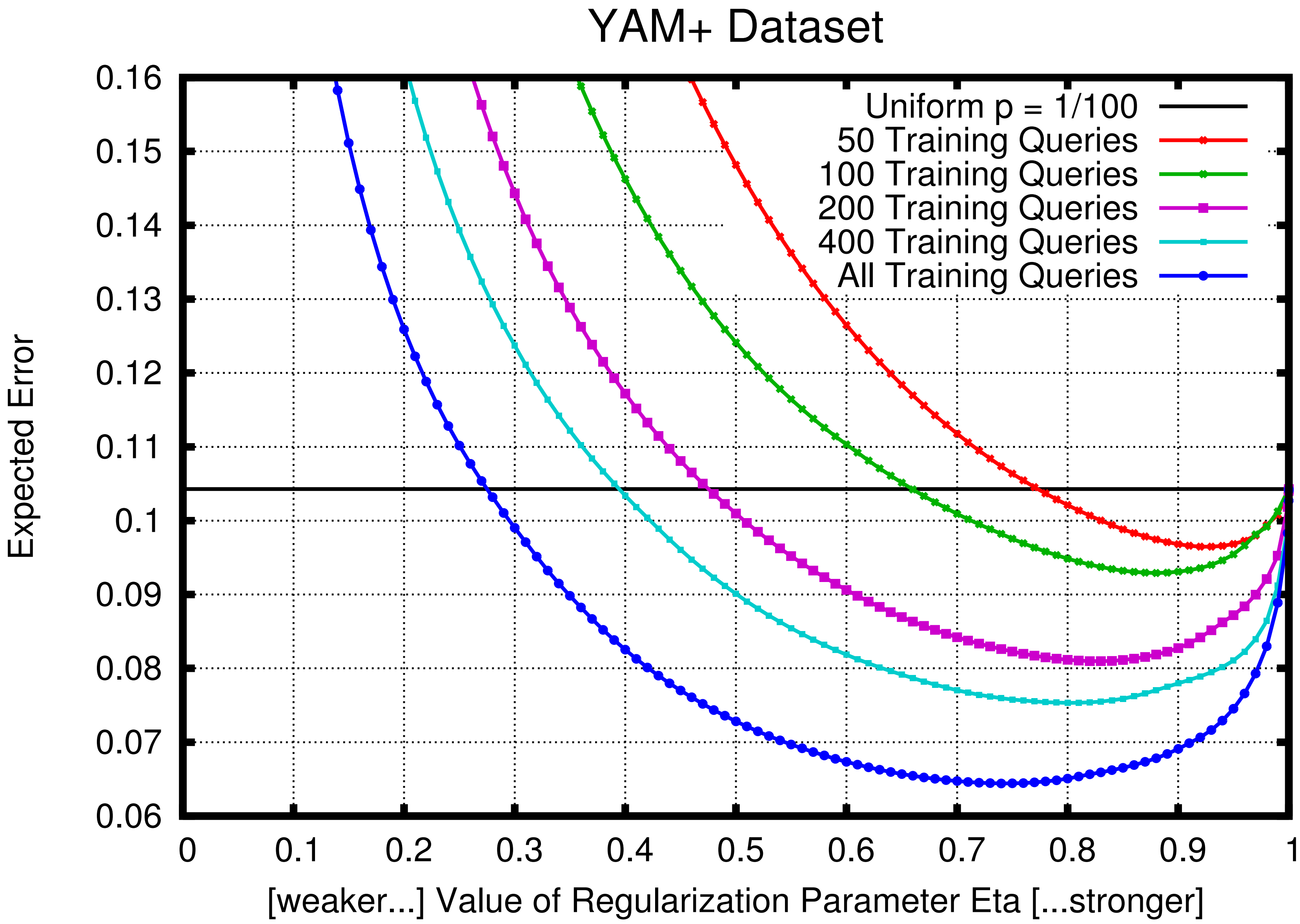


DBLP Dataset
Uniform Sampling p = 1/100
5000 Training Queries
10000 Training Queries
20000 Training Queries
40000 Training Queries
Expected Error
0.34
0.32
0.3
0.28
0.26
0.24
0.22
0.2
0.18
0.16
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
[weaker...] Value of Regularization Parameter Eta [...stronger]

RESULTS

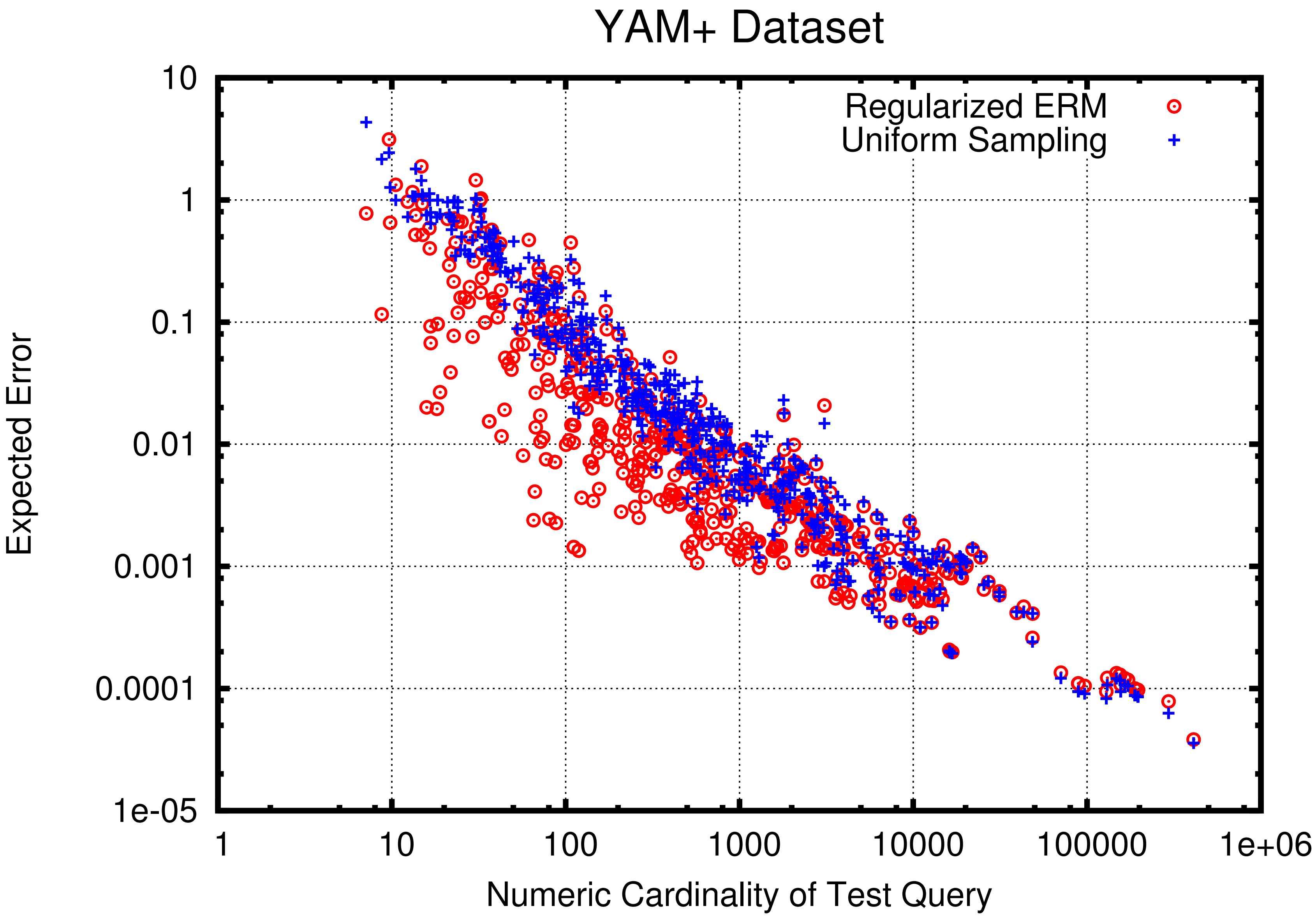

Cube Dataset
Uniform Sampling p = 1/10
50 Training Queries
100 Training Queries
200 Training Queries
800 Training Queries
6400 Training Queries
Expected Error
1.2
1.1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
[weaker...] Value of Regularization Parameter Eta [...stronger]

RESULTS



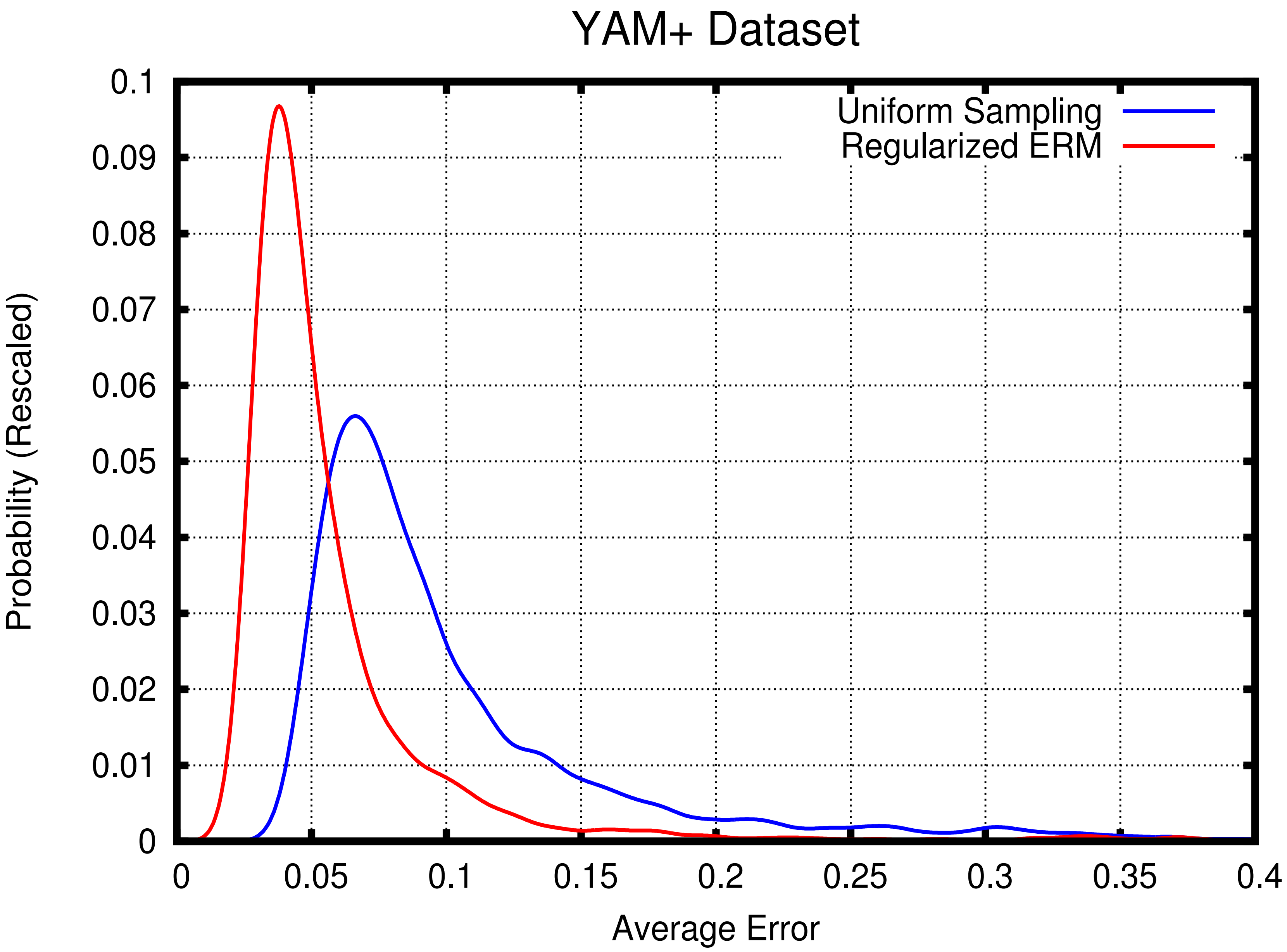
YAHOO!

RESULTS



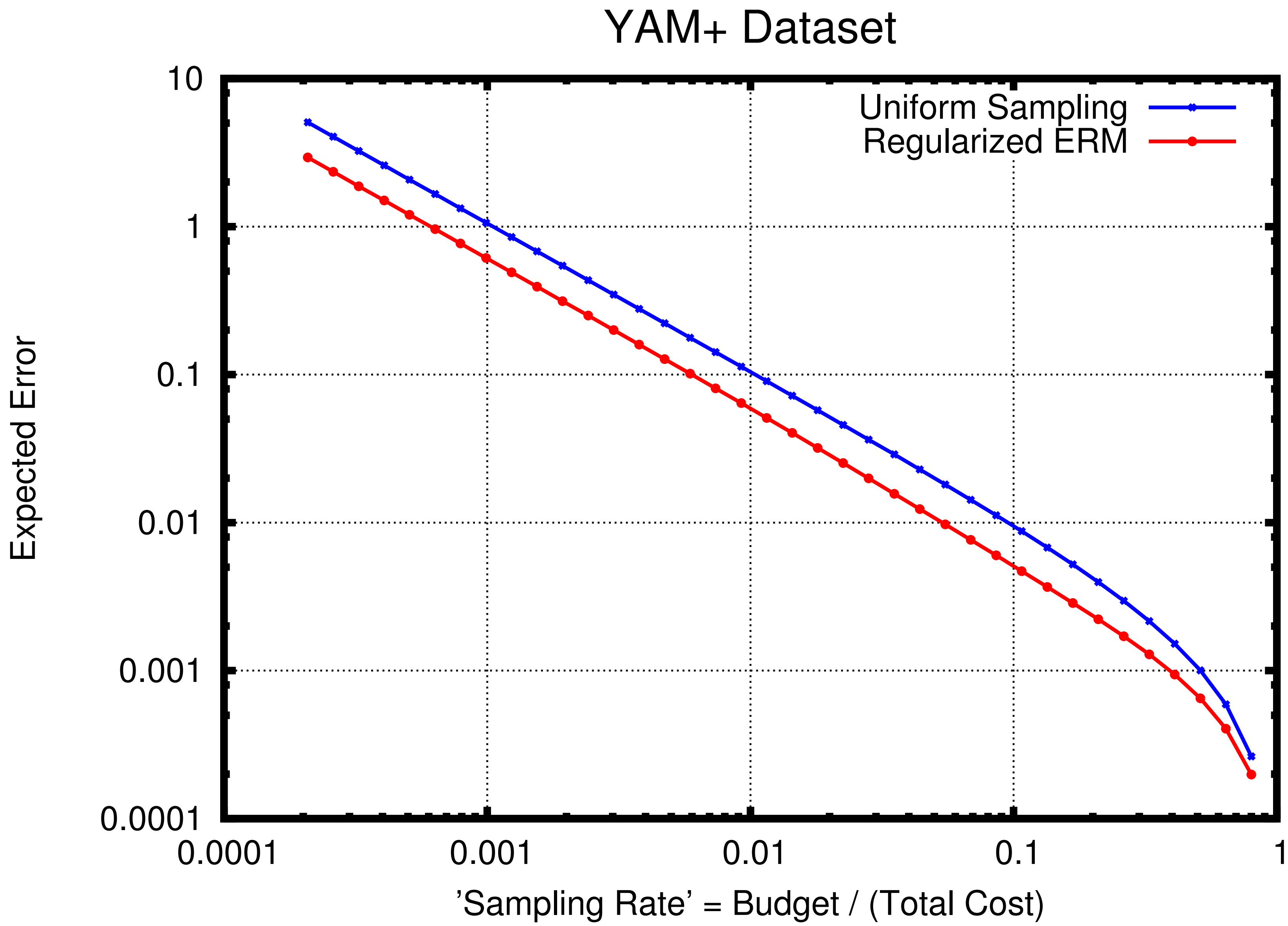
YAHOO!

RESULTS



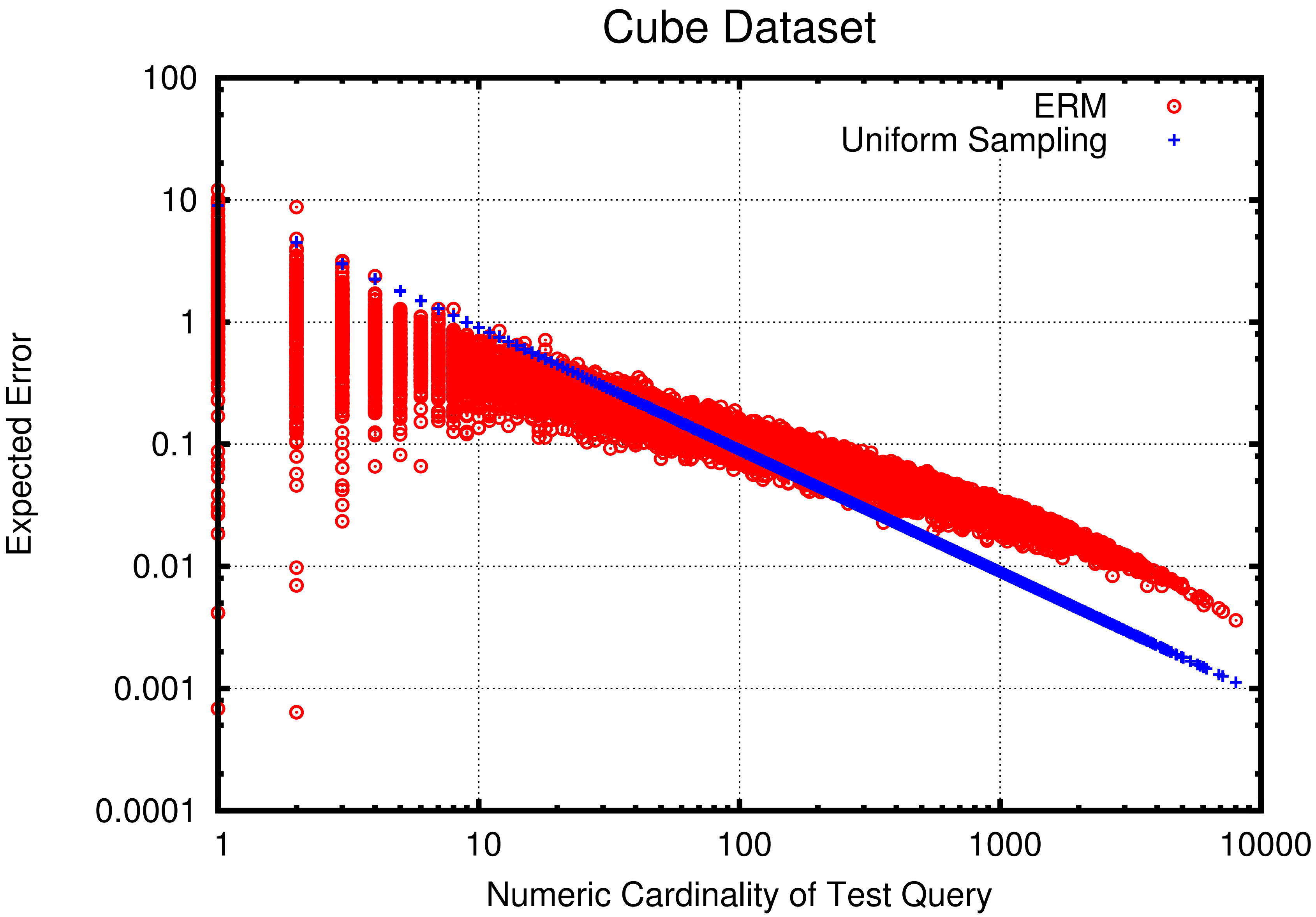
YAM+ Dataset
Uniform Sampling
Regularized ERM
Probability (Rescaled)
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
Average Error

RESULTS



YAHOO!

RESULTS



YAHOO!

YAHOO!