Streaming Algorithms You Should Know

Ted Dunning

© 2014 MapR Technologies

Contact Information

Ted Dunning

Chief Applications Architect at MapR Technologies

Committer & PMC for Apache's Drill, Zookeeper & Mahout

VP of Incubator at Apache Foundation

Email tdunning@apache.org tdunning@maprtech.com

Twitter @ted_dunning

© 2014 MapR Technologies

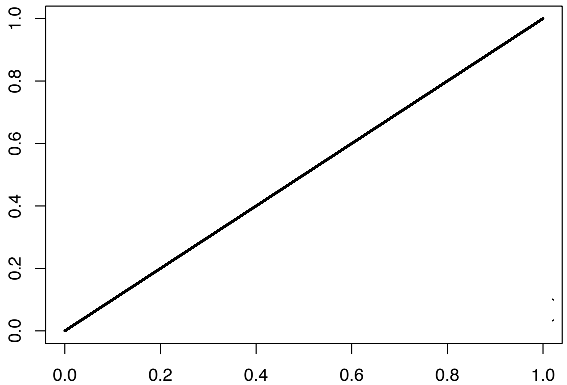
Must Approximate!

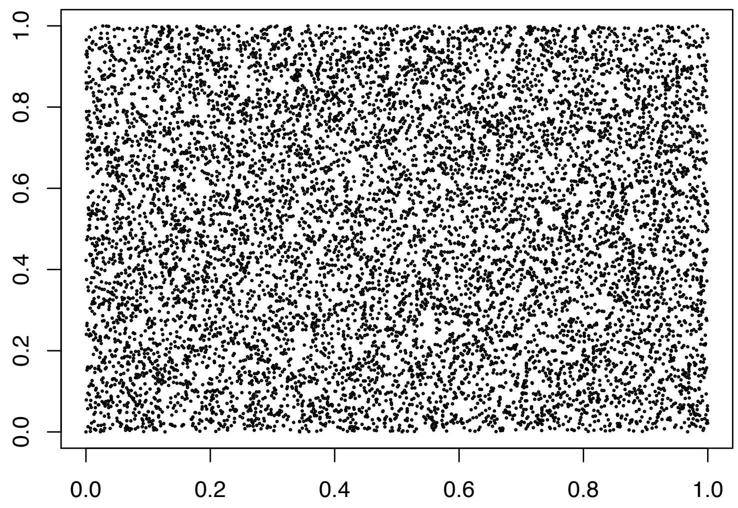
- Many important aggregates cannot be computed on-line
  - Memory cost proportional to data
  - Includes count distinct, heavy hitters, frequencies of top items, quantiles
  - Also k-means

- Most important aggregates can be approximated on-line
  - See all of the above

© 2014 MapR Technologies

Key Trick - Hashing





Key Trick - Sketching

- Often we can build a very small sketch of the data
- Carefully done, we can use the sketch to get values of interest

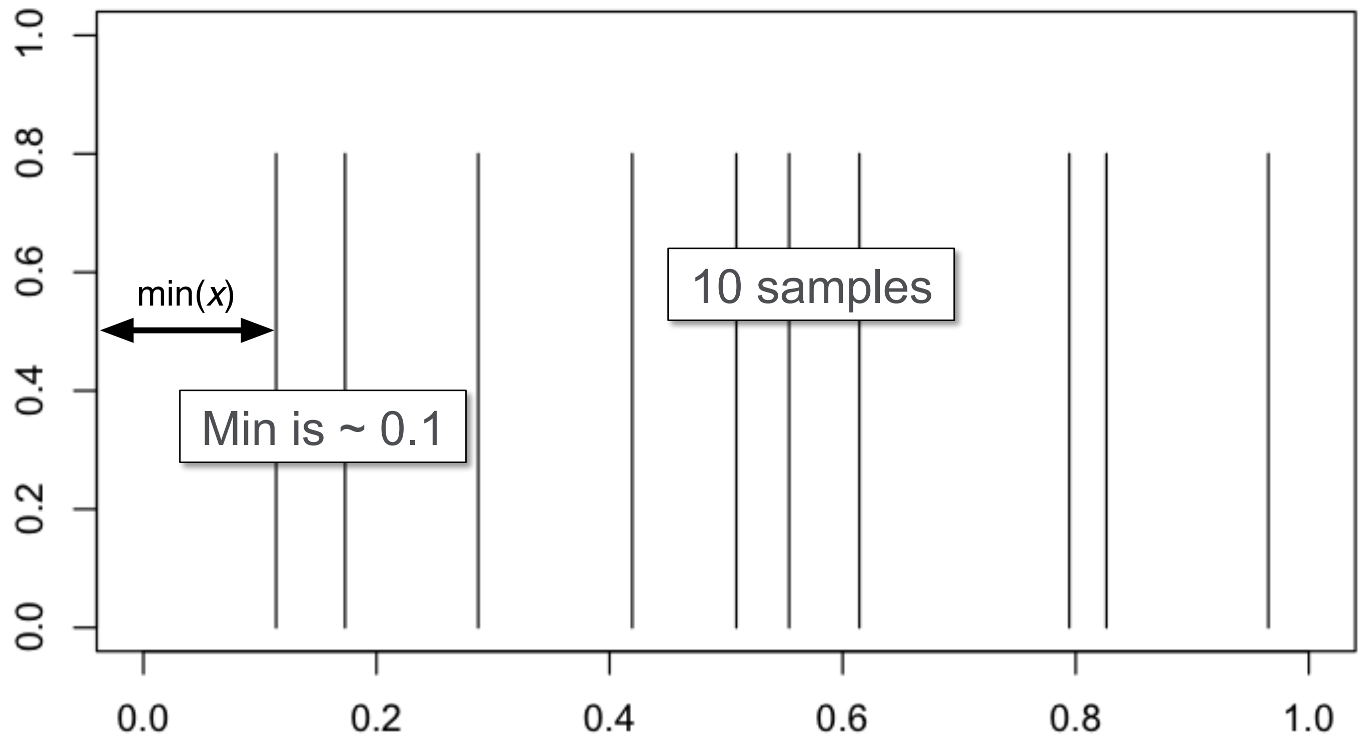
- All of the algorithms here use sketches of some kind

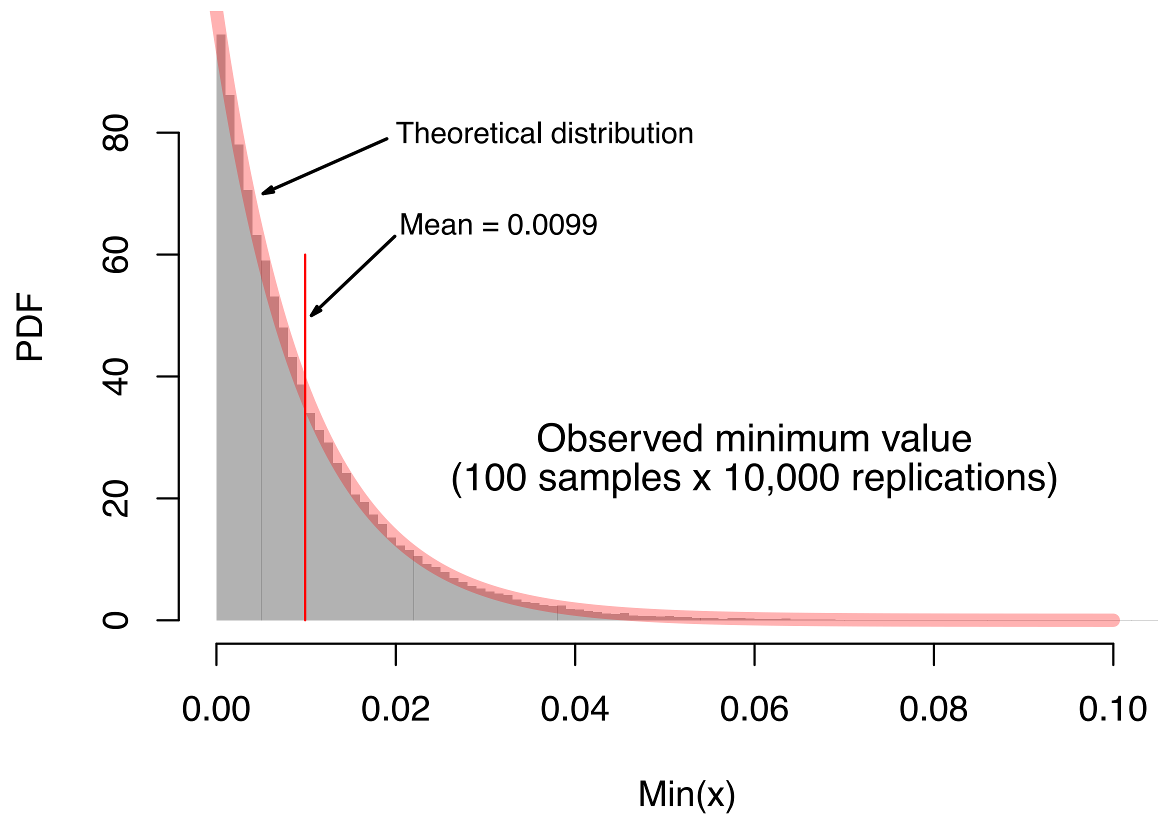
© 2014 MapR Technologies

Hyper log log

© 2014 MapR Technologies

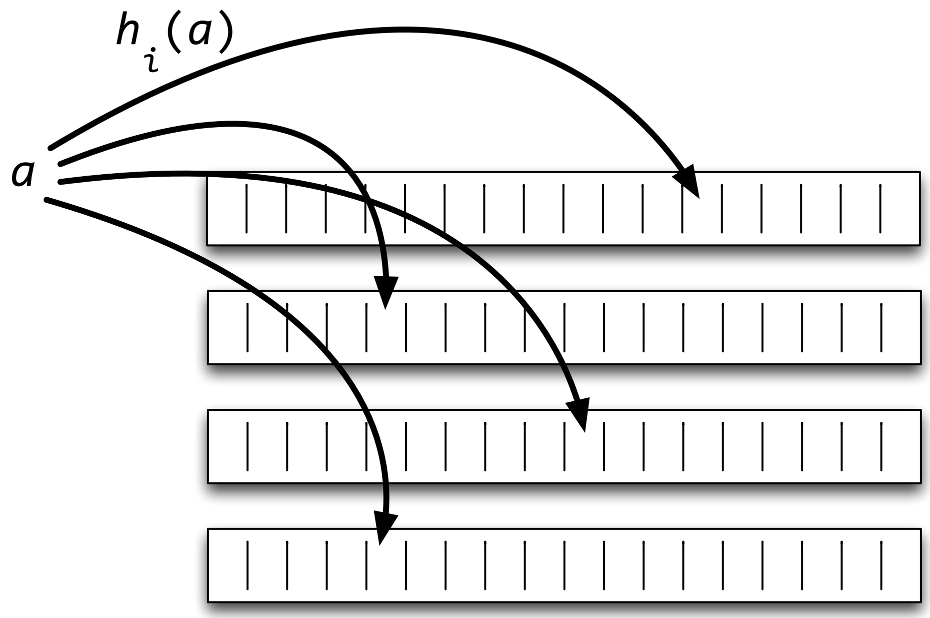
Repeated Minimum



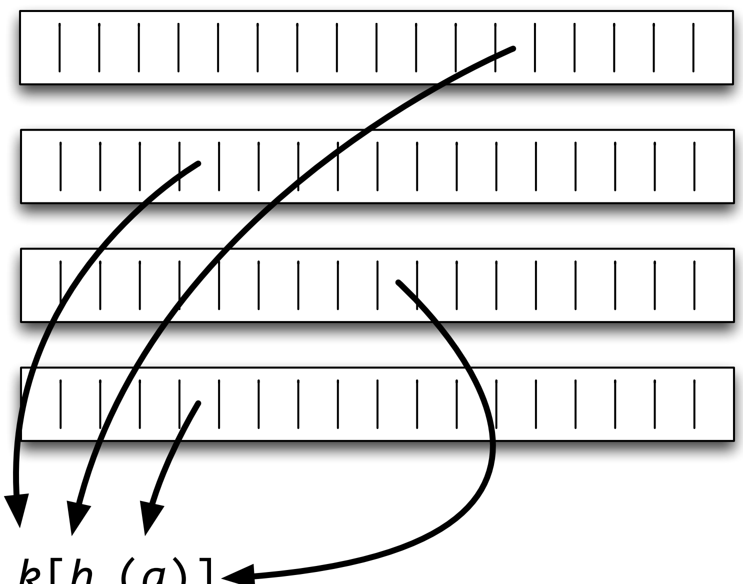

Theoretical distribution
Mean = 0.0099
Observed minimum value
(100 samples x 10,000 replications)
PDF
0
20
40
60
80
0.00
0.02
0.04
0.06
0.08
0.10
Min(x)

Count min sketch

Increment Hashed Locations to Insert

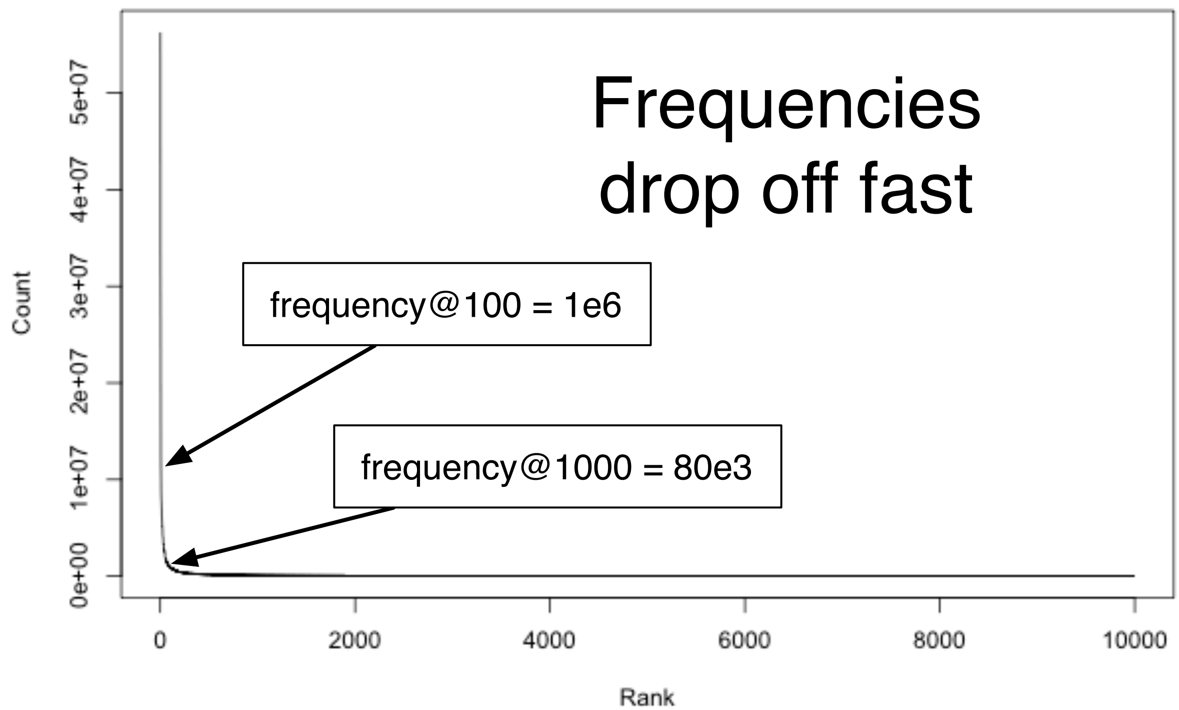


Probe Using min of Counts



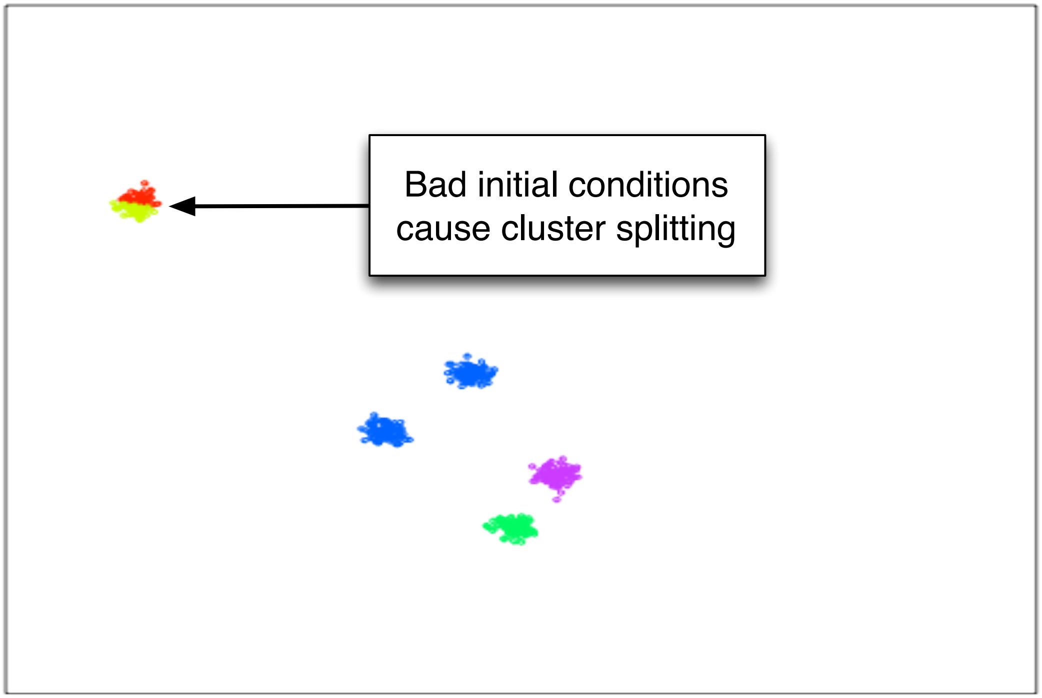
$$\min_i \, k[h_i(a)]$$

Leaky counters

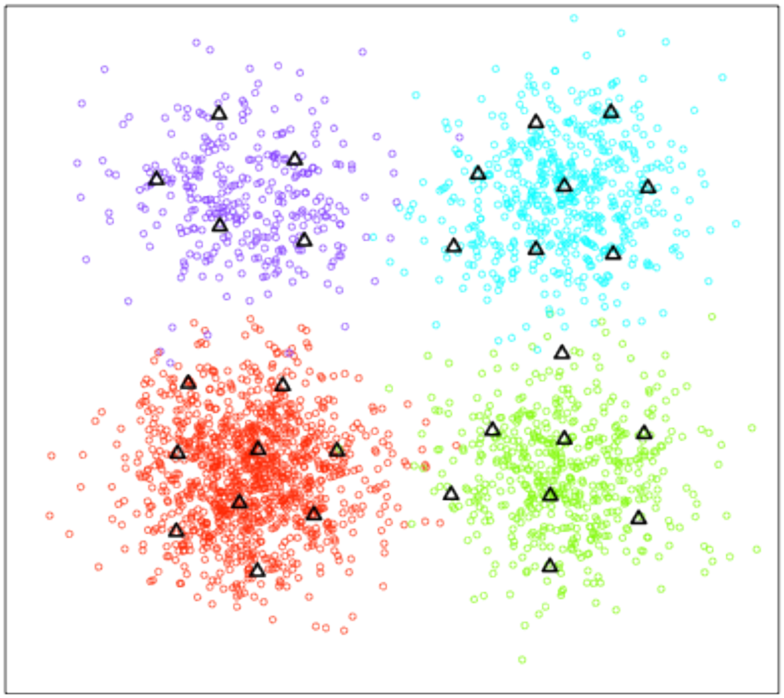

Frequencies
drop off fast
frequency@100 = 1e6
frequency@1000 = 80e3
Count
0e+00
1e+07
2e+07
3e+07
4e+07
5e+07
Rank
0
2000
4000
6000
8000
10000

Streaming *k*-means

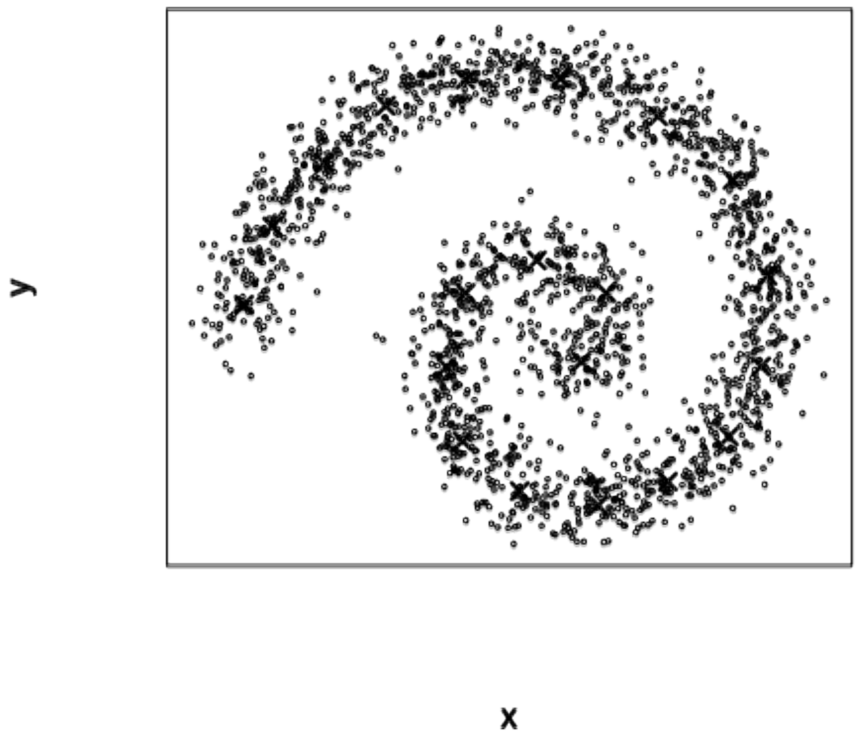
Typical k-means Failure



Many Clusters in Sketch is Fine



Lots of Clusters Approximate Anything

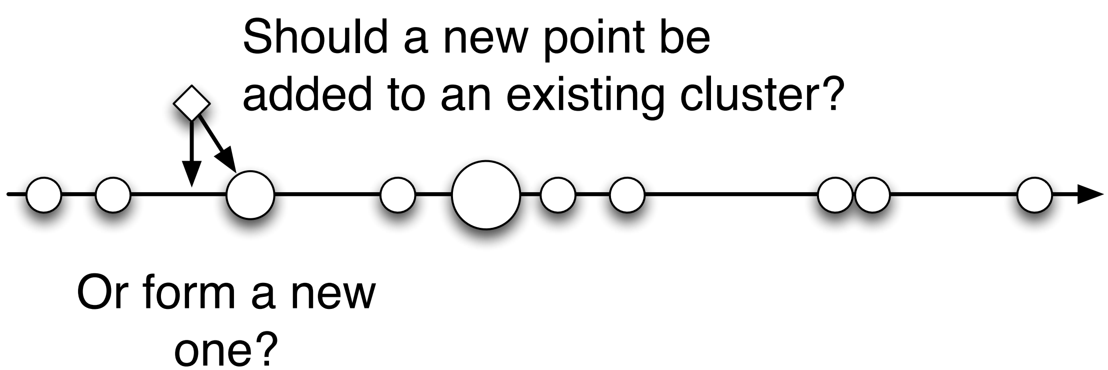


Clustering Summary

- Sketch is just lots of clusters
- Sketching can be done very approximately in one pass
- High quality clustering of the sketch is a high quality clustering of the data

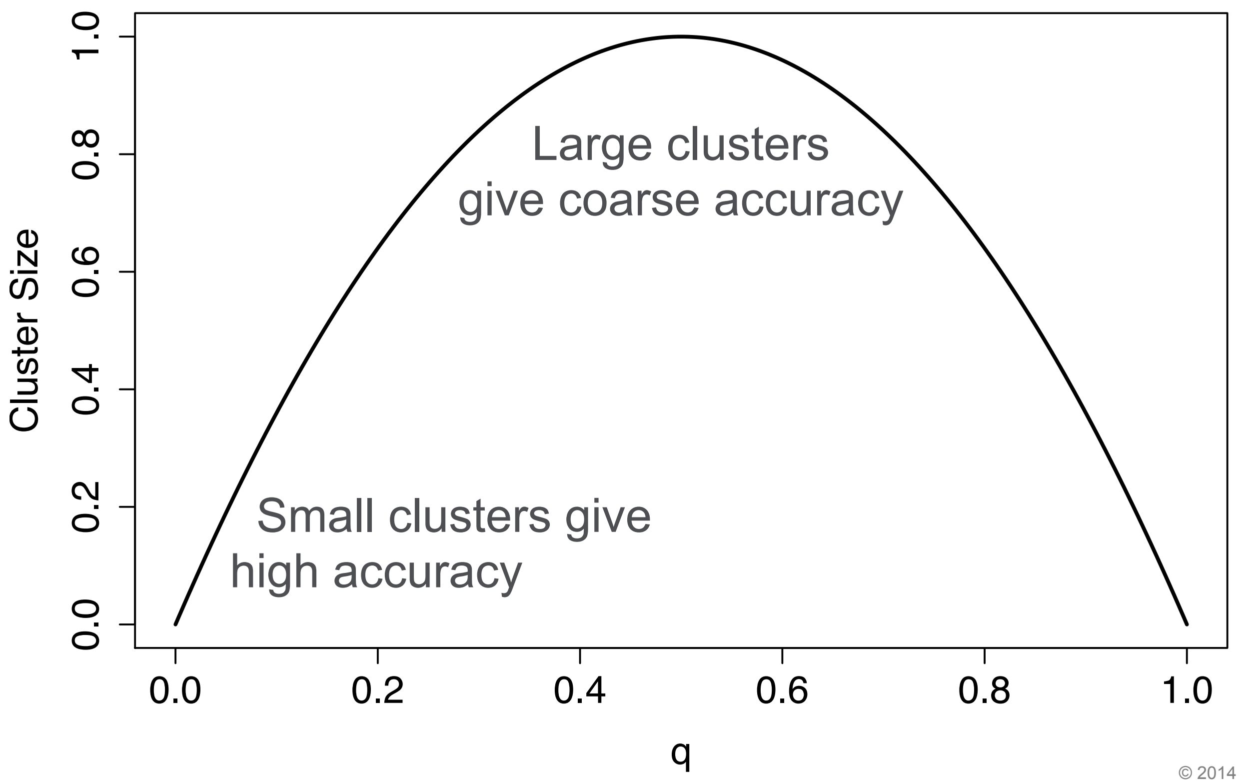
© 2014 MapR Technologies

Finally, *t*-digest

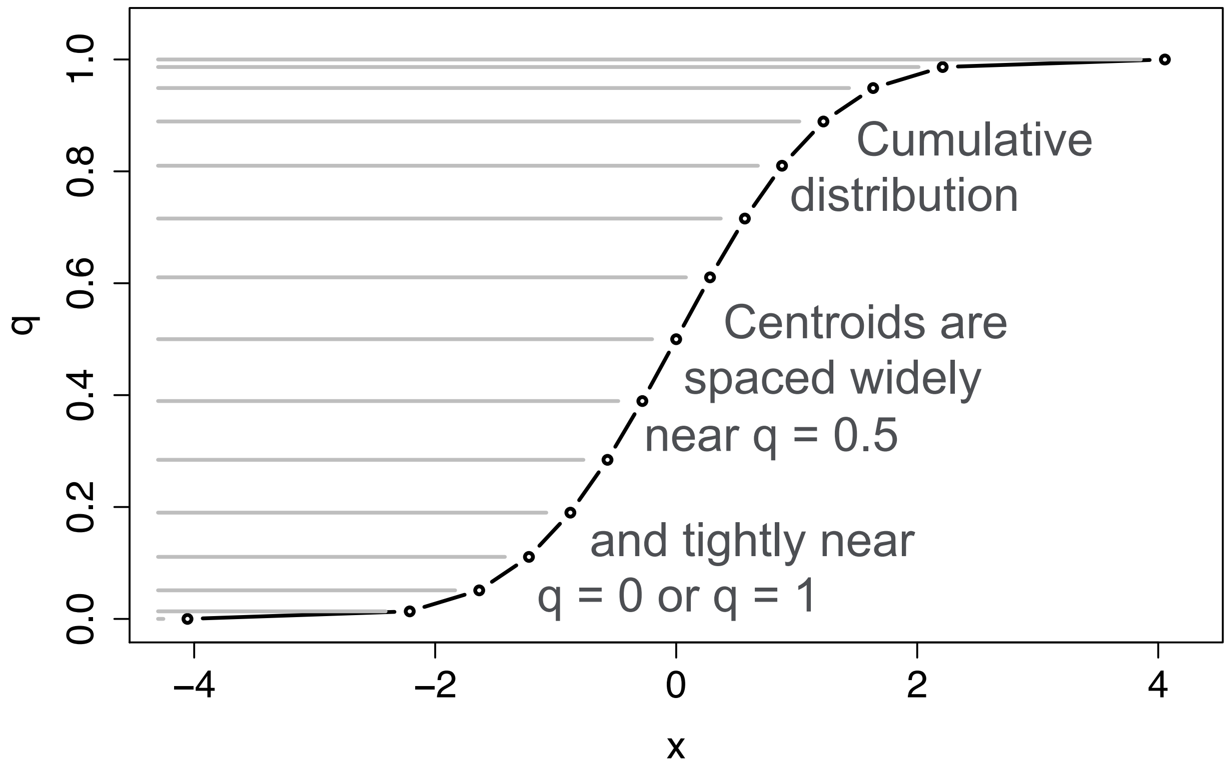


© 2014 MapR Technologies

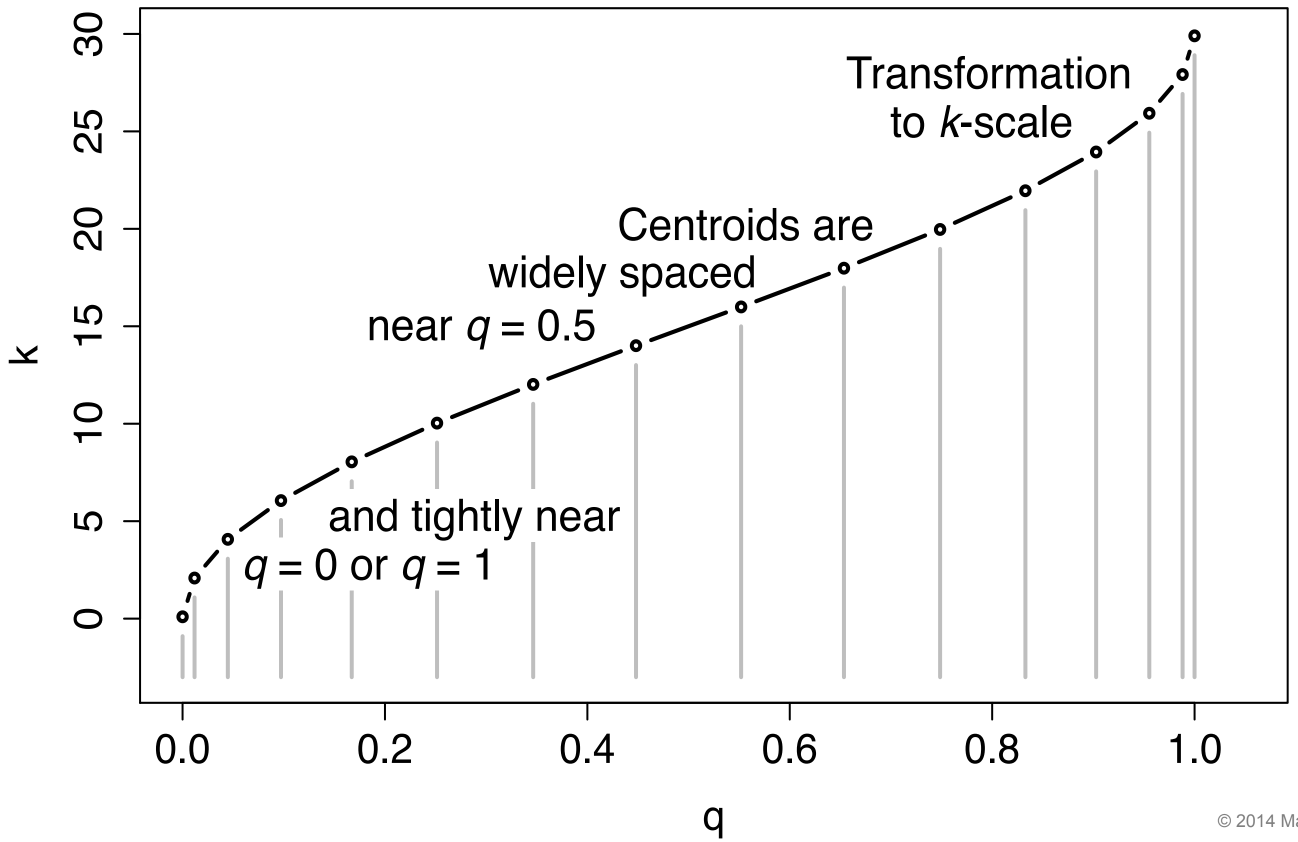
Variable Cluster Size for Constant Relative Accuracy



Second-Order Accuracy via Interpolation

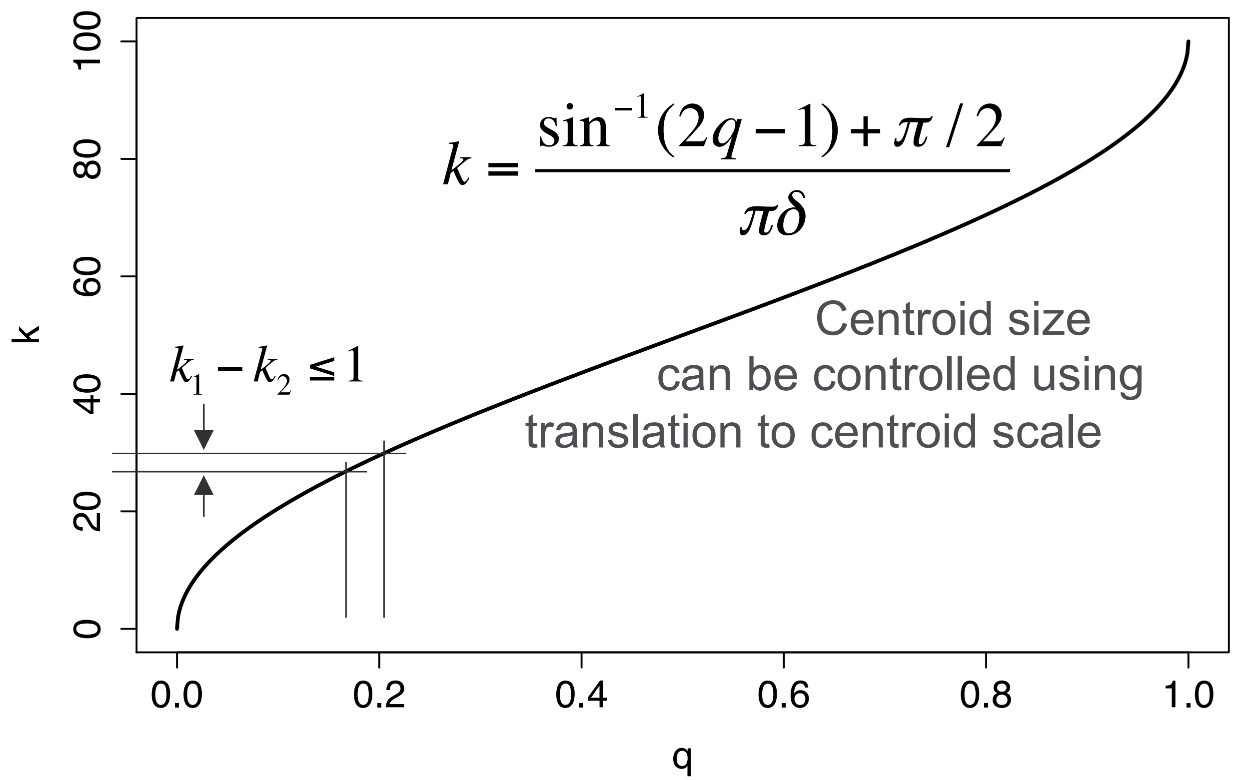


Size Limits via Constant Steps in *k*-scale

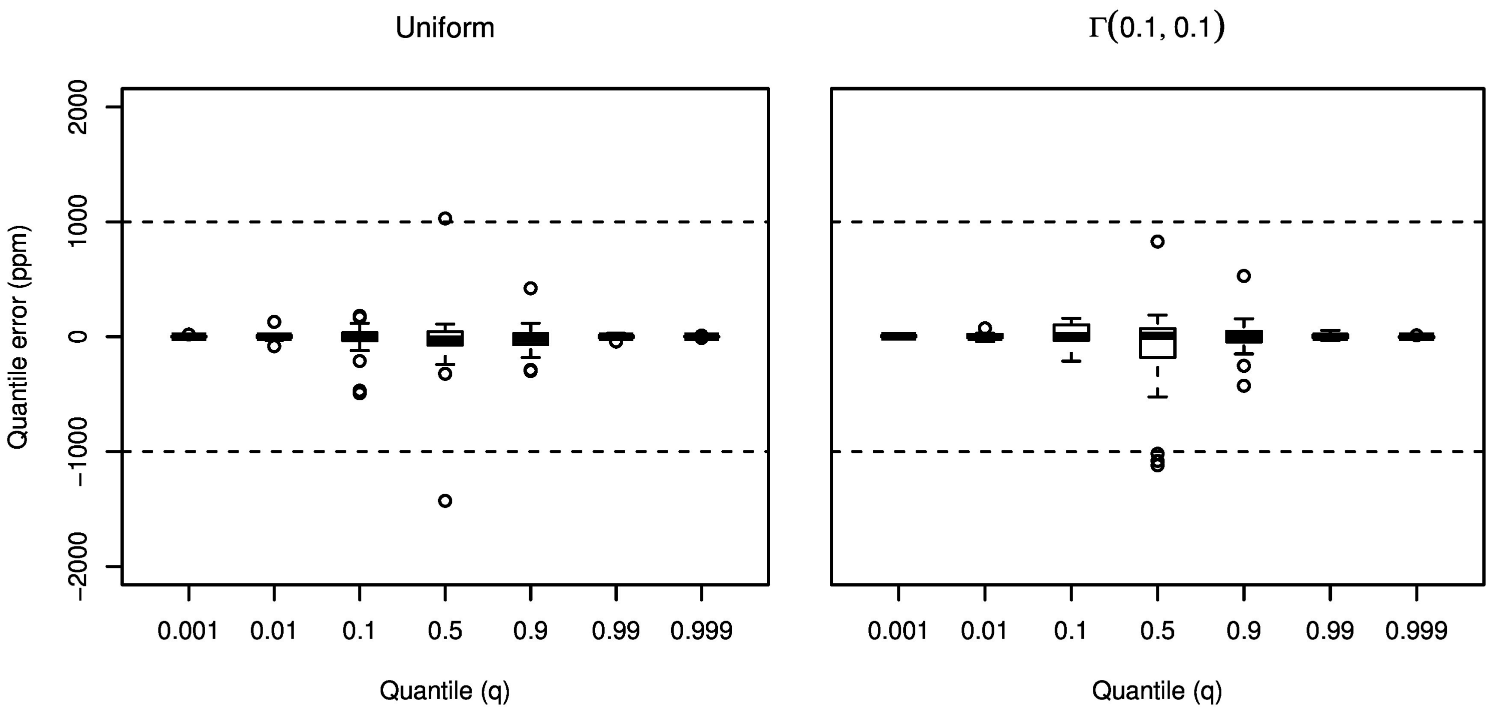


© 2014 MapR Technologies

Translation Between Quantile and Cluster



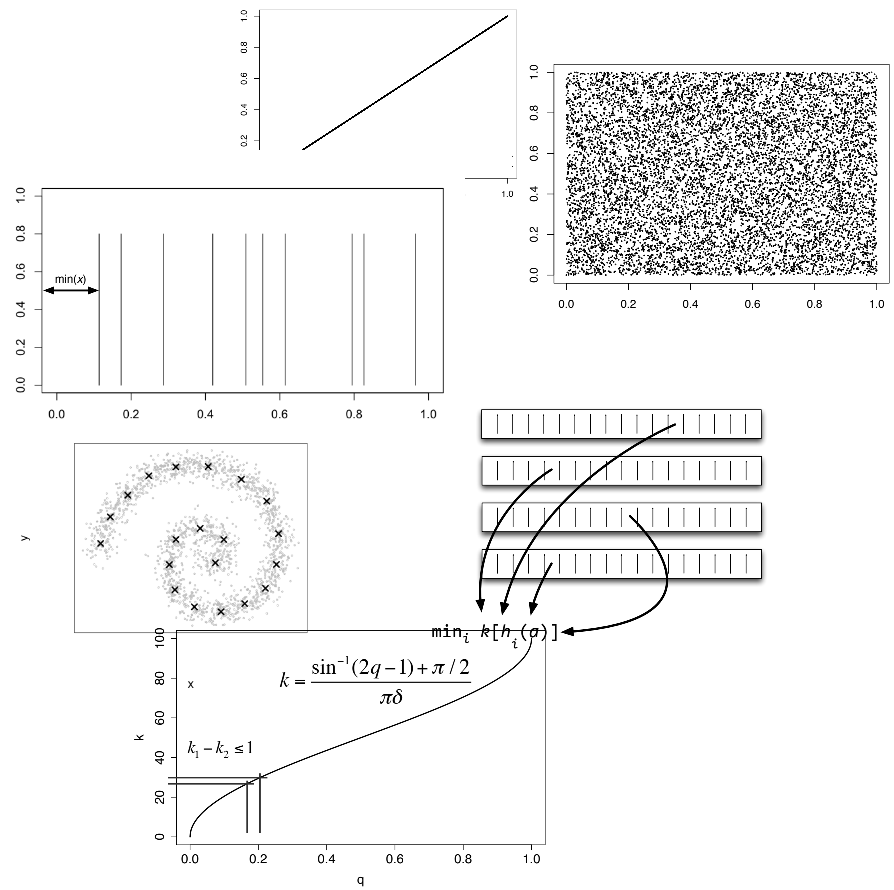
Actual Results



© 2014 MapR Technologies

Summary

- Hashing and sketching
- Hyper log log = count distinct
- Count min = count(s)
- Streaming k-means
- Quantiles via t-digest



MAPR

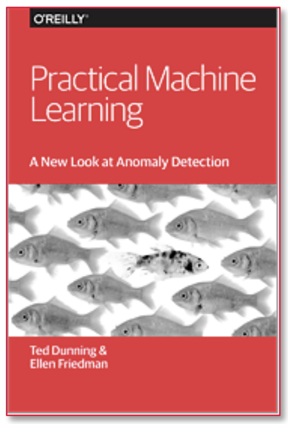
Free on-demand Hadoop training
leading to certification

Start becoming an expert now
mapr.com/training

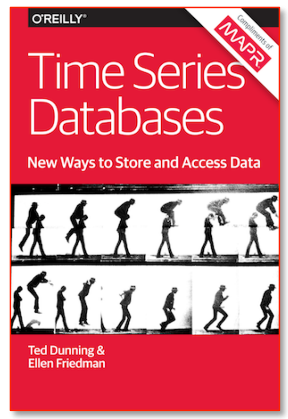
Short Books by Ted Dunning & Ellen Friedman

- Published by O'Reilly in 2014 and 2015
- For sale from Amazon or O'Reilly
- Free e-books currently available courtesy of MapR

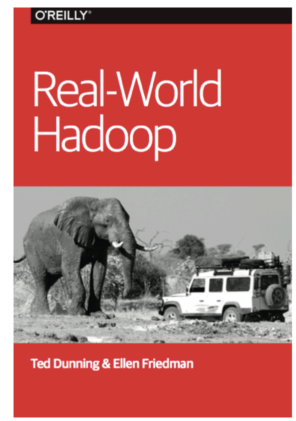
http://bit.ly/recommendation-ebook



http://bit.ly/ebook-anomaly



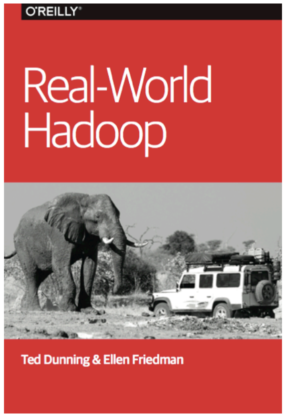
http://bit.ly/mapr-tsdb-ebook



http://bit.ly/ebook-real-world-hadoop

Real World Hadoop

by Ted Dunning and Ellen Friedman © Feb 2015 (published by O'Reilly)



Free copies at book signing today

Thank You!

Q&A

Engage with us!

@mapr	maprtech
mapr-technologies	MapR
tdunning@mapr.tech.com	maprtech

© 2014 MapR Technologies