Synopses for Massive Data: Samples, Histograms, Wavelets, Sketches

Synopses for Massive Data: Samples, Histograms, Wavelets, Sketches

Graham Cormode

AT&T Labs — Research
USA
graham@research.att.com

Minos Garofalakis

Technical University of Crete
Greece
minos@acm.org

Peter J. Haas

IBM Almaden Research Center
USA
peterh@almaden.ibm.com

Chris Jermaine

Rice University
USA
cmj4@cs.rice.edu

the essence of knowledge

Boston – Delft

Foundations and Trends® in Databases

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is G. Cormode, M. Garofalakis, P. J. Haas, and C. Jermaine, Synopses for Massive Data: Samples, Histograms, Wavelets, Sketches, Foundation and Trends® in Databases, vol 4, nos 1–3, pp 1–294, 2011

ISBN: 978-1-60198-516-3

Foundations and Trends® in Databases

Volume 4 Issues 1–3, 2011

Editorial Scope

Foundations and Trends® in Databases covers a breadth of topics relating to the management of large volumes of data. The journal targets the full scope of issues in data management, from theoretical foundations, to languages and modeling, to algorithms, system architecture, and applications. The list of topics below illustrates some of the intended coverage, though it is by no means exhaustive:

- Data Models and Query Languages
- Query Processing and Optimization
- Storage, Access Methods, and Indexing
- Transaction Management, Concurrency Control and Recovery
- Deductive Databases
- Parallel and Distributed Database Systems
- Database Design and Tuning
- Metadata Management
- Object Management
- Trigger Processing and Active Databases
- Data Mining and OLAP
- Approximate and Interactive Query Processing
- Data Warehousing
- Adaptive Query Processing
- Data Stream Management
- Search and Query Integration
- XML and Semi-Structured Data
- Web Services and Middleware
- Data Integration and Exchange
- Private and Secure Data Management
- Peer-to-Peer, Sensornet and Mobile Data Management
- Scientific and Spatial Data Management
- Data Brokering and Publish/Subscribe
- Data Cleaning and Information Extraction
- Probabilistic Data Management

Information for Librarians

Foundations and Trends® in Databases, 2011, Volume 4, 4 issues. ISSN paper version 1931-7883. ISSN online version 1931-7891. Also available as a combined paper and online subscription.

Foundations and Trends® in
Databases
Vol. 4, Nos. 1–3 (2011) 1–294

DOI: 10.1561/1900000004

Synopses for Massive Data: Samples, Histograms, Wavelets, Sketches

Graham Cormode[1], Minos Garofalakis[2], Peter J. Haas[3], and Chris Jermaine[4]

[1] *AT&T Labs — Research, 180 Park Avenue, Florham Park, NJ 07932, USA, graham@research.att.com*
[2] *Technical University of Crete, University Campus — Kounoupidiana, Chania, 73100, Greece, minos@acm.org*
[3] *IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95120-6099, USA, peterh@almaden.ibm.com*
[4] *Rice University, 6100 Main Street, Houston, TX 77005, USA, cmj4@cs.rice.edu*

Abstract

Methods for Approximate Query Processing (AQP) are essential for dealing with massive data. They are often the only means of providing interactive response times when exploring massive datasets, and are also needed to handle high speed data streams. These methods proceed by computing a lossy, compact synopsis of the data, and then executing the query of interest against the synopsis rather than the entire dataset. We describe basic principles and recent developments in AQP. We focus on four key synopses: random samples, histograms, wavelets, and sketches. We consider issues such as accuracy, space and time efficiency, optimality, practicality, range of applicability, error bounds on query answers, and incremental maintenance. We also discuss the tradeoffs between the different synopsis types.

Contents

1

Introduction

A synopsis of a massive dataset captures vital properties of the original data while typically occupying much less space. For example, suppose that our data consists of a large numeric time series. A simple summary allows us to compute the statistical variance of this series: we maintain the sum of all the values, the sum of the squares of the values, and the number of observations. Then the average is given by the ratio of the sum to the count, and the variance is ratio of the sum of squares to the count, less the square of the average. An important property of this synopsis is that we can build it efficiently. Indeed, we can find the three summary values in a single pass through the data.

However, we may need to know more about the data than merely its variance: how many different values have been seen? How many times has the series exceeded a given threshold? What was the behavior in a given time period? To answer such queries, our three-value summary does not suffice, and synopses appropriate to each type of query are needed. In general, these synopses will not be as simple or easy to compute as the synopsis for variance. Indeed, for many of these questions, there is no synopsis that can provide the exact answer, as is the case for variance. The reason is that for some classes of queries, the query answers collectively describe the data in full, and so any synopsis would effectively have to store the entire dataset.

To overcome this problem, we must relax our requirements. In many cases, the key objective is not obtaining the exact answer to a query, but rather receiving an accurate estimate of the answer. For example, in many settings, receiving an answer that is within 0.1% of the true result is adequate for our needs; it might suffice to know that the true answer is roughly $5 million without knowing that the exact answer is $5,001,482.76. Thus we can tolerate *approximation*, and there are many synopses that provide approximate answers. This small relaxation can make a big difference. Although for some queries it is impossible to provide a small synopsis that provides exact answers, there are many synopses that provide a very accurate approximation for these queries while using very little space.

1.1 The Need for Synopses

The use of synopses is essential for managing the massive data that arises in modern information management scenarios. When handling large datasets, from gigabytes to petabytes in size, it is often impractical to operate on them in full. Instead, it is much more convenient to build a synopsis, and then use this synopsis to analyze the data. This approach captures a variety of use-cases:

- A search engine collects logs of every search made, amounting to billions of queries every day. It would be too slow, and energy-intensive, to look for trends and patterns on the full data. Instead, it is preferable to use a synopsis that is guaranteed to preserve most of the as-yet undiscovered patterns in the data.
- A team of analysts for a retail chain would like to study the impact of different promotions and pricing strategies on sales of different items. It is not cost-effective to give each analyst the resources needed to study the national sales data in full, but by working with synopses of the data, each analyst can perform their explorations on their own laptops.

- A large cellphone provider wants to track the health of its network by studying statistics of calls made in different regions, on hardware from different manufacturers, under different levels of contention, and so on. The volume of information is too large to retain in a database, but instead the provider can build a synopsis of the data as it is observed live, and then use the synopsis off-line for further analysis.

These examples expose a variety of settings. The full data may reside in a traditional data warehouse, where it is indexed and accessible, but is too costly to work on in full. In other cases, the data is stored as flat-files in a distributed file system; or it may never be stored in full, but be accessible only as it is observed in a streaming fashion. Sometimes synopsis construction is a one-time process, and sometimes we need to update the synopsis as the base data is modified or as accuracy requirements change. In all cases though, being able to construct a high quality synopsis enables much faster and more scalable data analysis.

From the 1990s through today, there has been an increasing demand for systems to query more and more data at ever faster speeds. Enterprise data requirements have been estimated [173] to grow at 60% per year through at least 2011, reaching 1,800 exabytes. On the other hand, users — weaned on Internet browsers, sophisticated analytics and simulation software with advanced GUIs, and computer games — have come to expect real-time or near-real-time answers to their queries. Indeed, it has been increasingly realized that extracting knowledge from data is usually an interactive process, with a user issuing a query, seeing the result, and using the result to formulate the next query, in an iterative fashion. Of course, parallel processing techniques can also help address these problems, but may not suffice on their own. Many queries, for example, are not embarrassingly parallel. Moreover, methods based purely on parallelism can be expensive. Indeed, under evolving models for cloud computing, specifically "platform as a service" fee models, users will pay costs that directly reflect the computing resources that they use. In this setting, use of Approximate Query Processing (AQP) techniques can lead to significant cost savings. Similarly, recent work [15] has pointed out that approximate processing techniques can lead to energy savings and greener computing. Thus AQP techniques are essential for providing, in a cost-effective manner, interactive response times for exploratory queries over massive data.

Exacerbating the pressures on data management systems is the increasing need to query streaming data, such as real time financial data or sensor feeds. Here the flood of high speed data can easily overwhelm the often limited CPU and memory capacities of a stream processor unless AQP methods are used. Moreover, for purposes of network monitoring and many other applications, approximate answers suffice when trying to detect general patterns in the data, such as a denial-of-service attack. AQP techniques are thus well suited to streaming and network applications.

1.2 Survey Overview

In this survey, we describe basic principles and recent developments in building approximate synopses (i.e., lossy, compressed representations) of massive data. Such synopses enable AQP, in which the user's query is executed against the synopsis instead of the original data. We focus on the four main families of synopses: random samples, histograms, wavelets, and sketches.

A *random sample* comprises a "representative" subset of the data values of interest, obtained via a stochastic mechanism. Samples can be quick to obtain, and can be used to approximately answer a wide range of queries.

A *histogram* summarizes a dataset by grouping the data values into subsets, or "buckets," and then, for each bucket, computing a small set of summary statistics that can be used to approximately reconstruct the data in the bucket. Histograms have been extensively studied and have been incorporated into the query optimizers of virtually all commercial relational DBMSs.

Wavelet-based synopses were originally developed in the context of image and signal processing. The dataset is viewed as a set of M elements in a vector — that is, as a function defined on the set $\{0,1,2,\ldots,M-1\}$ — and the wavelet transform of this function is found as a weighted sum of wavelet "basis functions." The weights, or coefficients, can then be "thresholded," for example, by eliminating coefficients that are close to zero in magnitude. The remaining small set of coefficients serves as the synopsis. Wavelets are good at capturing features of the dataset at various scales.

Sketch summaries are particularly well suited to streaming data. Linear sketches, for example, view a numerical dataset as a vector or matrix, and multiply the data by a fixed matrix. Such sketches are massively parallelizable. They can accommodate streams of transactions in which data is both inserted and removed. Sketches have also been used successfully to estimate the answer to COUNT DISTINCT queries, a notoriously hard problem.

Many questions arise when evaluating or using synopses.

- What is the class of queries that can be approximately answered?
- What is the approximation accuracy for a synopsis of a given size?
- What are the space and time requirements for constructing a synopsis of a given size, as well as the time required to approximately answer the query?
- How should one choose synopsis parameters such as the number of histogram buckets or the wavelet thresholding value? Is there an optimal, that is, most accurate, synopsis of a given size?
- When using a synopsis to approximately answer a query, is it possible to obtain error bounds on the approximate query answer?
- Can the synopsis be incrementally maintained in an efficient manner?
- Which type of synopsis is best for a given problem?

We explore these issues in subsequent chapters.

1.3 Outline

It is possible to read the discussion of each type of synopsis in isolation, to understand a particular summarization approach. We have tried to use common notation and terminology across all chapters, in order to facilitate comparison of the different synopses. In more detail, the topics covered by the different chapters are given below.

1.3.1 Sampling

Random samples are perhaps the most fundamental synopses for AQP, and the most widely implemented. The simplicity of the idea — executing the desired query against a small representative subset of the data — belies centuries of research across many fields, with decades of effort in the database community alone. Many different methods of extracting and maintaining samples of data have been proposed, along with multiple ways to build an estimator for a given query. This chapter introduces the mathematical foundations for sampling, in terms of accuracy and precision, and discusses the key sampling schemes: Bernoulli sampling, stratified sampling, and simple random sampling with and without replacement.

For simple queries, such as basic SUM and AVERAGE queries, it is straightforward to build unbiased estimators from samples. The more general case — an arbitrary SQL query with nested subqueries — is more daunting, but can sometimes be solved quite naturally in a procedural way.

For small tables, drawing a sample can be done straightforwardly. For larger relations, which may not fit conveniently in memory, or may not even be stored on disk in full, more advanced techniques are needed to make the sampling process scalable. For disk-resident data, sampling methods that operate at the granularity of a block rather than a tuple may be preferred. Existing indices can also be leveraged to help the sampling. For large streams of data, considerable effort has been put into maintaining a uniform sample as new items arrive or existing items are deleted. Finally, "online aggregation" algorithms enhance interactive exploration of massive datasets by exploiting the fact that an imprecise sampling-based estimate of a query result can be incrementally improved simply by collecting more samples.

1.3.2 Histograms

The histogram is a fundamental object for summarizing the frequency distribution of an attribute or combination of attributes. The most basic histograms are based on a fixed division of the domain (equi-width), or using quantiles (equi-depth), and simply keep statistics on the number of items from the input which fall in each such bucket. But many more complex methods have been designed, which aim to provide the most accurate summary possible within a limited space budget. Schemes differ in how the buckets are chosen, what statistics are stored, how estimates are extracted, and what classes of query are supported. They are quantified based on the space and time requirements used to build them, and the resulting accuracy guarantees that they provide.

The one-dimensional case is at the heart of histogram construction, since higher dimensions are typically handled via extensions of one-dimensional ideas. Beyond equi-width and equi-depth, end biased and high biased, maxdiff and other generalizations have been proposed. For a variety of approximation-error metrics, dynamic programming (DP) methods can be used to find histograms — notably the "v-optimal histograms" — that minimize the error, subject to an upper bound on the allowable histogram size. Approximate methods can be used when the quadratic cost of DP is not practical. Many other constructions, both optimal and heuristic, are described, such as lattice histograms, STHoles, and maxdiff histograms. The extension of these methods to higher dimensions adds complexity. Even the two-dimensional case presents challenges in how to define the space of possible bucketings. The cost of these methods also rises exponentially with the dimensionality of the data, inspiring new approaches that combine sets of low-dimensional histograms with high-level statistical models.

Histograms most naturally answer range–sum queries — for example, "compute total sales between July and September for adults from age 25 through 40" — and their variations. They can also be used to approximate more general classes of queries, such as aggregations over joins. Various negative theoretical and empirical results indicate that one should not expect histograms to give accurate answers to arbitrary queries. Nevertheless, due to their conceptual simplicity, histograms can be effectively used for a broad variety of estimation tasks, including set-valued queries, real-valued data, and aggregate queries over predicates more complex than simple ranges.

1.3.3 Wavelets

The wavelet synopsis is conceptually close to the histogram summary. The central difference is that, whereas histograms primarily produce buckets that are subsets of the original data-attribute domain, wavelet representations transform the data and seek to represent the most significant features in a wavelet (i.e., "frequency") domain, and can capture combinations of high and low frequency information. The most widely discussed wavelet transformation is the Haar-wavelet transform (HWT), which can, in general, be constructed in time linear in the size of the underlying data array. Picking the B largest HWT coefficients results in a synopsis that provides the optimal L_2 (sum-squared) error for the reconstructed data. Extending from one-dimensional to multi-dimensional data, as with histograms, provides more definitional challenges. There are multiple plausible choices here, as well as algorithmic challenges in efficiently building the wavelet decomposition.

The core AQP task for wavelet summaries is to estimate the answer to range sums. More general SPJ (select, project, join) queries can also be directly applied on relation summaries, to generate a summary of the resulting relation. This is made possible through an appropriately-defined AQP algebra that operates entirely in the domain of wavelet coefficients.

Recent research into wavelet representations has focused on error guarantees beyond L_2. These include L_1 (sum of errors) or L_∞ (maximum error), as well as relative-error versions of these measures. A fundamental choice here is whether to restrict the possible coefficient values to those arising under the basic wavelet transform, or to allow other (unrestricted) coefficient values, specifically chosen to reduce the target error metric. The construction of such (restricted or unrestricted) wavelet synopses optimized for non-L_2 error metrics is a challenging problem.

1.3.4 Sketches

Sketch techniques have undergone extensive development over the past few years. They are especially appropriate for streaming data, in which large quantities of data flow by and the sketch summary must continually be updated quickly and compactly. Sketches, as presented here, are designed so that the update caused by each new piece of data is largely independent of the current state of the summary. This design choice makes them faster to process, and also easy to parallelize.

"Frequency based sketches" are concerned with summarizing the observed frequency distribution of a dataset. From these sketches, accurate estimations of individual frequencies can be extracted. This leads to algorithms for finding approximate "heavy hitters" — items that account for a large fraction of the frequency mass — and quantiles such as the median and its generalizations. The same sketches can also be used to estimate the sizes of (equi)joins between relations, self-join sizes, and range queries. Such sketch summaries can be used as primitives within more complex mining operations, and to extract wavelet and histogram representations of streaming data.

A different style of sketch construction leads to sketches for "distinct-value" queries that count the number of distinct values in a given multiset. As mentioned above, using a sample to estimate the answer to a COUNT DISTINCT query may give highly inaccurate results. In contract, sketching methods that make a pass over the entire dataset can provide guaranteed accuracy. Once built, these sketches estimate not only the cardinality of a given attribute or combination of attributes, but also the cardinality of various operations performed on them, such as set operations (union and difference), and selections based on arbitrary predicates.

In the final chapter, we compare the different synopsis methods. We also discuss the use of AQP within research systems, and discuss challenges and future directions.

In our discussion, we often use terminology and examples that arise in classical database systems, such as SQL queries over relational databases. These artifacts partially reflect the original context of the results that we survey, and provide a convenient vocabulary for the various data and access models that are relevant to AQP. We emphasize that the techniques discussed here can be applied much more generally. Indeed, one of the key motivations behind this survey is the hope that these techniques — and their extensions — will become a fundamental component of tomorrow's information management systems.

References

[1] A. Abouzeid, K. Bajda-Pawlikowski, D. J. Abadi, A. Rasin, and A. Silberschatz, "HadoopDB: An architectural hybrid of MapReduce and DBMS technologies for analytical workloads," *PVLDB*, vol. 2, no. 1, pp. 922–933, 2009.

[2] S. Acharya, P. B. Gibbons, and V. Poosala, "Aqua: A fast decision support system using approximate query answers," in *International Conference on Very Large Data Bases*, 1999.

[3] S. Acharya, P. B. Gibbons, V. Poosala, and S. Ramaswamy, "The Aqua approximate query answering system," in *ACM SIGMOD International Conference on Management of Data*, 1999.

[4] S. Acharya, P. B. Gibbons, V. Poosala, and S. Ramaswamy, "Join synopses for approximate query answering," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 275–286, New York, NY, USA, 1999.

[5] C. C. Aggarwal, "On biased reservoir sampling in the presence of stream evolution," in *Proceedings of the International Conference on Very Large Data Bases*, pp. 607–618, 2006.

[6] N. Alon, P. Gibbons, Y. Matias, and M. Szegedy, "Tracking join and self-join sizes in limited storage," in *ACM Principles of Database Systems*, 1999.

[7] N. Alon, Y. Matias, and M. Szegedy, "The space complexity of approximating the frequency moments," in *ACM Symposium on Theory of Computing*, 1996.

[8] A. Andoni, R. Krauthgamer, and K. Onak, "Streaming algorithms from precision sampling," *CoRR*, p. abs/1011.1263, 2010.

[9] G. Antoshenkov, "Random sampling from pseudo-ranked B+ trees," in *Proceedings of the International Conference on Very Large Data Bases*, pp. 375–382, 1992.

[10] P. M. Aoki, "Algorithms for index-assisted selectivity estimation," in *Proceedings of the International Conference on Data Engineering*, p. 258, 1999.
[11] R. Avnur, J. M. Hellerstein, B. Lo, C. Olston, B. Raman, V. Raman, T. Roth, and K. Wylie, "CONTROL: Continuous output and navigation technology with refinement on-line," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 567–569, 1998.
[12] B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom, "Models and issues in data stream systems," in *ACM Principles of Database Systems*, 2002.
[13] B. Babcock, S. Chaudhuri, and G. Das, "Dynamic sample selection for approximate query processing," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 539–550, New York, NY, USA, 2003.
[14] B. Babcock, M. Datar, and R. Motwani, "Sampling from a moving window over streaming data," in *SODA*, pp. 633–634, 2002.
[15] W. Baek and T. Chilimbi, "Green: A framework for supporting energy-conscious programming using controlled approximation," in *Proceedings of PLDI*, pp. 198–209, 2010.
[16] L. Baltrunas, A. Mazeika, and M. H. Böhlen, "Multi-dimensional histograms with tight bounds for the error," in *Proceedings of IDEAS*, pp. 105–112, 2006.
[17] Z. Bar-Yossef, T. Jayram, R. Kumar, D. Sivakumar, and L. Trevisian, "Counting distinct elements in a data stream," in *Proceedings of RANDOM 2002*, 2002.
[18] R. Bellman, "On the approximation of curves by line segments using dynamic programming," *Communications of ACM*, vol. 4, no. 6, p. 284, 1961.
[19] K. S. Beyer, P. J. Haas, B. Reinwald, Y. Sismanis, and R. Gemulla, "On synopses for distinct-value estimation under multiset operations," in *ACM SIGMOD International Conference on Management of Data*, 2007.
[20] S. Bhattacharrya, A. Madeira, S. Muthukrishnan, and T. Ye, "How to scalably skip past streams," in *Scalable Stream Processing Systems (SSPS) Workshop with ICDE 2007*, 2007.
[21] L. Bhuvanagiri, S. Ganguly, D. Kesh, and C. Saha, "Simpler algorithm for estimating frequency moments of data streams," in *ACM-SIAM Symposium on Discrete Algorithms*, 2006.
[22] P. Billingsley, *Probability and Measure*. Wiley, third ed., 1999.
[23] B. Blohsfeld, D. Korus, and B. Seeger, "A comparison of selectivity estimators for range queries on metric attributes," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 239–250, 1999.
[24] B. Bloom, "Space/time trade-offs in hash coding with allowable errors," *Communications of the ACM*, vol. 13, no. 7, pp. 422–426, July 1970.
[25] A. Broder, M. Charikar, A. Frieze, and M. Mitzenmacher, "Min-wise independent permutations," in *ACM Symposium on Theory of Computing*, 1998.
[26] A. Z. Broder and M. Mitzenmacher, "Network applications of bloom filters: A survey," *Internet Mathematics*, vol. 1, no. 4, 2003.
[27] P. G. Brown and P. J. Haas, "Techniques for warehousing of sample data," in *Proceedings of the International Conference on Data Engineering*, p. 6, Washington, DC, USA, 2006.

[28] B. Bru, "The estimates of Laplace. an example: Research concerning the population of a large empire, 1785–1812," in *Journal de la Société de statistique de Paris*, vol. 129, pp. 6–45, 1988.

[29] N. Bruno and S. Chaudhuri, "Exploiting statistics on query expressions for optimization," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 263–274, 2002.

[30] N. Bruno, S. Chaudhuri, and L. Gravano, "STHoles: A multidimensional workload-aware histogram," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 211–222, 2001.

[31] T. Bu, J. Cao, A. Chen, and P. P. C. Lee, "A fast and compact method for unveiling significant patterns in high speed networks," in *IEEE INFOCOMM*, 2007.

[32] F. Buccafurri, F. Furfaro, G. Lax, and D. Saccà, "Binary-tree histograms with tree indices," in *Proceedings of Database and Expert Systems Applications*, pp. 861–870, 2002.

[33] F. Buccafurri, F. Furfaro, and D. Saccà, "Estimating range queries using aggregate data with integrity constraints: A probabilistic approach," in *Proceedings of the International Conference on Database Theory*, pp. 390–404, 2001.

[34] F. Buccafurri, F. Furfaro, D. Saccà, and C. Sirangelo, "A quad-tree based multiresolution approach for two-dimensional summary data," in *Proceedings of the International Conference on Scientific and Statistical Database Management*, pp. 127–137, 2003.

[35] F. Buccafurri and G. Lax, "Fast range query estimation by n-level tree histograms," *Data Knowledge in Engineering*, vol. 51, no. 2, pp. 257–275, 2004.

[36] F. Buccafurri and G. Lax, "Reducing data stream sliding windows by cyclic tree-like histograms," in *PKDD*, pp. 75–86, 2004.

[37] F. Buccafurri, G. Lax, D. Saccà, L. Pontieri, and D. Rosaci, "Enhancing histograms by tree-like bucket indices," *VLDB Journal*, vol. 17, no. 5, pp. 1041–1061, 2008.

[38] C. Buragohain, N. Shrivastava, and S. Suri, "Space efficient streaming algorithms for the maximum error histogram," in *Proceedings of the International Conference on Data Engineering*, pp. 1026–1035, 2007.

[39] J. L. Carter and M. N. Wegman, "Universal classes of hash functions," *Journal of Computer and System Sciences*, vol. 18, no. 2, pp. 143–154, 1979.

[40] A. Chakrabarti, G. Cormode, and A. McGregor, "A near-optimal algorithm for computing the entropy of a stream," in *ACM-SIAM Symposium on Discrete Algorithms*, 2007.

[41] K. Chakrabarti, M. Garofalakis, R. Rastogi, and K. Shim, "Approximate query processing using wavelets," in *Proceedings of the International Conference on Very Large Data Bases*, pp. 111–122, Cairo, Egypt, September 2000.

[42] K. Chakrabarti, M. N. Garofalakis, R. Rastogi, and K. Shim, "Approximate query processing using wavelets," *The VLDB Journal*, vol. 10, no. 2–3, pp. 199–223, September 2001. (Best of VLDB'2000 Special Issue).

[43] D. Chamberlin, *A Complete Guide to DB2 Universal Database*. Morgan Kaufmann, 1998.

[44] M. Charikar, S. Chaudhuri, R. Motwani, and V. R. Narasayya, "Towards estimation error guarantees for distinct values," in *Proceedings of ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pp. 268–279, 2000.

[45] M. Charikar, K. Chen, and M. Farach-Colton, "Finding frequent items in data streams," in *International Colloquium on Automata, Languages and Programming*, 2002.

[46] M. Charikar, K. Chen, and M. Farach-Colton, "Finding frequent items in data streams," *Theoretical Computer Science*, vol. 312, no. 1, pp. 3–15, 2004.

[47] S. Chaudhuri, G. Das, and V. R. Narasayya, "A robust, optimization-based approach for approximate answering of aggregate queries," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 295–306, 2001.

[48] S. Chaudhuri, G. Das, and V. R. Narasayya, "Optimized stratified sampling for approximate query processing," *ACM Transactions on Database Systems*, vol. 32, no. 2, p. 9, 2007.

[49] S. Chaudhuri, G. Das, and U. Srivastava, "Effective use of block-level sampling in statistics estimation," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 287–298, 2004.

[50] S. Chaudhuri, R. Motwani, and V. Narasayya, "On random sampling over joins," *SIGMOD Record*, vol. 28, no. 2, pp. 263–274, 1999.

[51] S. Chaudhuri, R. Motwani, and V. R. Narasayya, "Random sampling for histogram construction: How much is enough?," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 436–447, 1998.

[52] K. K. Chen, "Influence query optimization with optimization profiles and statistical views in DB2 9: Optimal query performance in DB2 9 for Linux, UNIX, and Windows," Available at `www.ibm.com/developerworks/db2/library/techarticle/dm-0612chen`, 2006.

[53] W. Cheney and W. Light, *A Course in Approximation Theory.* Brooks/Cole, Pacific Grove, CA, 2000.

[54] E. Cohen, G. Cormode, and N. G. Duffield, "Structure-aware sampling on data streams," in *SIGMETRICS*, pp. 197–208, 2011.

[55] E. Cohen, N. Grossaug, and H. Kaplan, "Processing top-k queries from samples," in *Proceedings of CoNext*, p. 7, 2006.

[56] E. Cohen and H. Kaplan, "Summarizing data using bottom-k sketches," in *ACM Conference on Principles of Distributed Computing (PODC)*, 2007.

[57] S. Cohen and Y. Matias, "Spectral bloom filters," in *ACM SIGMOD International Conference on Management of Data*, 2003.

[58] T. Condie, N. Conway, P. Alvaro, J. Hellerstein, K. Elmeleegy, and R. Sears, "MapReduce online," in *NSDI*, 2010.

[59] J. Considine, M. Hadjieleftheriou, F. Li, J. W. Byers, and G. Kollios, "Robust approximate aggregation in sensor data management systems," *ACM Transactions on Database Systems*, vol. 34, no. 1, April 2009.

[60] J. Considine, F. Li, G. Kollios, and J. Byers, "Approximate aggregation techniques for sensor databases," in *IEEE International Conference on Data Engineering*, 2004.

[61] G. Cormode, M. Datar, P. Indyk, and S. Muthukrishnan, "Comparing data streams using Hamming norms," in *International Conference on Very Large Data Bases*, 2002.

[62] G. Cormode, A. Deligiannakis, M. N. Garofalakis, and A. McGregor, "Probabilistic histograms for probabilistic data," *PVLDB*, vol. 2, no. 1, pp. 526–537, 2009.

[63] G. Cormode and M. Garofalakis, "Sketching streams through the net: Distributed approximate query tracking," in *International Conference on Very Large Data Bases*, 2005.

[64] G. Cormode, M. Garofalakis, and D. Sacharidis, "Fast approximate wavelet tracking on streams," in *Proceedings of the International Conference on Extending Database Technology*, Munich, Germany, March 2006.

[65] G. Cormode and M. Hadjieleftheriou, "Finding frequent items in data streams," in *International Conference on Very Large Data Bases*, 2008.

[66] G. Cormode, P. Indyk, N. Koudas, and S. Muthukrishnan, "Fast mining of tabular data via approximate distance computations," in *IEEE International Conference on Data Engineering*, 2002.

[67] G. Cormode, F. Korn, S. Muthukrishnan, T. Johnson, O. Spatscheck, and D. Srivastava, "Holistic UDAFs at streaming speeds," in *ACM SIGMOD International Conference on Management of Data*, pp. 35–46, 2004.

[68] G. Cormode, F. Korn, S. M. Muthukrishnan, and D. Srivastava, "Space- and time-efficient deterministic algorithms for biased quantiles over data streams," in *Proceedings of ACM Principles of Database Systems*, pp. 263–272, 2006.

[69] G. Cormode and S. Muthukrishnan, "What's new: Finding significant differences in network data streams," in *Proceedings of IEEE Infocom*, 2004.

[70] G. Cormode and S. Muthukrishnan, "An improved data stream summary: The count-min sketch and its applications," *Journal of Algorithms*, vol. 55, no. 1, pp. 58–75, 2005.

[71] G. Cormode and S. Muthukrishnan, "Summarizing and mining skewed data streams," in *SIAM Conference on Data Mining*, 2005.

[72] G. Cormode, S. Muthukrishnan, and I. Rozenbaum, "Summarizing and mining inverse distributions on data streams via dynamic inverse sampling," in *International Conference on Very Large Data Bases*, 2005.

[73] G. Cormode, S. Muthukrishnan, K. Yi, and Q. Zhang, "Optimal sampling from distributed streams," in *Proceedings of ACM Principles of Database Systems*, pp. 77–86, 2010.

[74] C. Cranor, T. Johnson, O. Spatscheck, and V. Shkapenyuk, "Gigascope: A stream database for network applications," in *ACM SIGMOD International Conference on Management of Data*, 2003.

[75] N. N. Dalvi, C. Ré, and D. Suciu, "Probabilistic databases: Diamonds in the dirt," *Communications of the ACM*, vol. 52, no. 7, pp. 86–94, 2009.

[76] A. Das, J. Gehrke, and M. Riedewald, "Approximation techniques for spatial data," in *ACM SIGMOD International Conference on Management of Data*, 2004.

[77] M. Datar, A. Gionis, P. Indyk, and R. Motwani, "Maintaining stream statistics over sliding windows," in *ACM-SIAM Symposium on Discrete Algorithms*, 2002.

[78] I. Daubechies, *Ten Lectures on Wavelets*. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM), 1992.

[79] A. Deligiannakis, M. Garofalakis, and N. Roussopoulos, "An approximation scheme for probabilistic wavelet synopses," in *Proceedings of the International Conference on Scientific and Statistical Database Management*, Santa Barbara, California, June 2005.

[80] A. Deligiannakis, M. Garofalakis, and N. Roussopoulos, "Extended wavelets for multiple measures," *ACM Transactions on Database Systems*, vol. 32, no. 2, June 2007.

[81] A. Deligiannakis and N. Roussopoulos, "Extended wavelets for multiple measures," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, San Diego, California, June 2003.

[82] F. Deng and D. Rafiei, "New estimation algorithms for streaming data: Count-min can do more," `http://www.cs.ualberta.ca/~fandeng/paper/cmm.pdf`, 2007.

[83] A. Deshpande, M. N. Garofalakis, and R. Rastogi, "Independence is good: Dependency-based histogram synopses for high-dimensional data," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 199–210, 2001.

[84] R. A. DeVore, "Nonlinear approximation," *Acta Numerica*, vol. 7, pp. 51–150, 1998.

[85] L. Devroye and G. Lugosi, *Combinatorial Methods in Density Estimation*. Springer, 2001.

[86] A. Dobra, "Histograms revisited: When are histograms the best approximation method for aggregates over joins?," in *Proceedings of ACM Principles of Database Systems*, pp. 228–237, 2005.

[87] A. Dobra, M. Garofalakis, J. E. Gehrke, and R. Rastogi, "Processing complex aggregate queries over data streams," in *ACM SIGMOD International Conference on Management of Data*, 2002.

[88] A. Dobra and F. Rusu, "Statistical analysis of sketch estimators," *ACM Transactions on Database Systems*, vol. 33, no. 3, 2008.

[89] D. Donjerkovic, Y. E. Ioannidis, and R. Ramakrishnan, "Dynamic histograms: Capturing evolving data sets," in *Proceedings of the International Conference on Data Engineering*, p. 86, 2000.

[90] D. Donjerkovic and R. Ramakrishnan, "Probabilistic optimization of top N queries," in *Proceedings of the International Conference on Very Large Data Bases*, pp. 411–422, 1999.

[91] N. Duffield, C. Lund, and M. Thorup, "Estimating flow distributions from sampled flow statistics," in *ACM SIGCOMM*, 2003.

[92] M. Durand and P. Flajolet, "Loglog counting of large cardinalities," in *European Symposium on Algorithms (ESA)*, 2003.

[93] C. Estan and G. Varghese, "New directions in traffic measurement and accounting," in *ACM SIGCOMM*, 2002.

[94] C. T. Fan, M. E. Muller, and I. Rezucha, "Development of sampling plans by using sequential (item by item) selection techniques and digital computers," *Journal of the American Statistical Association*, pp. 387–402, 1962.

[95] G. Fishman, *Monte Carlo: Concepts, Algorithms and Applications*. Springer, 1996.

[96] P. Flajolet, "On adaptive sampling," *Computing*, vol. 43, no. 4, 1990.

[97] P. Flajolet and G. N. Martin, "Probabilistic counting algorithms for database applications," *Journal of Computer and System Sciences*, vol. 31, pp. 182–209, 1985.

[98] G. Frahling, P. Indyk, and C. Sohler, "Sampling in dynamic data streams and applications," in *Symposium on Computational Geometry*, June 2005.

[99] D. Fuchs, Z. He, and B. S. Lee, "Compressed histograms with arbitrary bucket layouts for selectivity estimation," *Information on Science*, vol. 177, no. 3, pp. 680–702, 2007.

[100] S. Ganguly, "Counting distinct items over update streams," *Theoretical Computer Science*, vol. 378, no. 3, pp. 211–222, 2007.

[101] S. Ganguly, M. Garofalakis, and R. Rastogi, "Processing set expressions over continuous update streams," in *ACM SIGMOD International Conference on Management of Data*, 2003.

[102] S. Ganguly, M. Garofalakis, and R. Rastogi, "Processing data-stream join aggregates using skimmed sketches," in *International Conference on Extending Database Technology*, 2004.

[103] S. Ganguly, P. B. Gibbons, Y. Matias, and A. Silberschatz, "Bifocal sampling for skew-resistant join size estimation," *SIGMOD Record*, vol. 25, no. 2, pp. 271–281, 1996.

[104] S. Ganguly and A. Majumder, "CR-precis: A deterministic summary structure for update data streams," in *ESCAPE*, 2007.

[105] M. Garofalakis, J. Gehrke, and R. Rastogi, "Querying and mining data streams: You only get one look," in *ACM SIGMOD International Conference on Management of Data*, 2002.

[106] M. Garofalakis, J. Gehrke, and R. Rastogi, eds., *Data Stream Management: Processing High-Speed Data Streams*. Springer, 2011.

[107] M. Garofalakis and P. B. Gibbons, "Approximate query processing: Taming the terabytes," Tutorial in *International Conference on Very Large Data Bases*, Roma, Italy, September 2001.

[108] M. Garofalakis and P. B. Gibbons, "Wavelet synopses with error guarantees," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 476–487, Madison, Wisconsin, June 2002.

[109] M. Garofalakis and P. B. Gibbons, "Probabilistic wavelet synopses," *ACM Transactions on Database Systems*, vol. 29, no. 1, March 2004. (SIGMOD/PODS'2002 Special Issue).

[110] M. Garofalakis and A. Kumar, "Deterministic wavelet thresholding for maximum-error metrics," in *Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, Paris, France, June 2004.

[111] M. Garofalakis and A. Kumar, "Wavelet synopses for general error metrics," *ACM Transactions on Database Systems*, vol. 30, no. 4, December 2005. (SIGMOD/PODS'2004 Special Issue).

[112] R. Gemulla, "Sampling algorithms for evolving datasets," PhD Thesis, Technische Universität Dresden, Available at http://nbn-resolving.de/urn:nbn:de:bsz:14-ds-1224861856184-11644, 2009.

[113] R. Gemulla and W. Lehner, "Deferred maintenance of disk-based random samples," in *Proceedings of International Conference on Extending Database Technology*, Lecture Notes in Computer Science, pp. 423–441, 2006.

[114] R. Gemulla and W. Lehner, "Sampling time-based sliding windows in bounded space," in *SIGMOD Conference*, pp. 379–392, 2008.

[115] R. Gemulla, W. Lehner, and P. J. Haas, "A dip in the reservoir: Maintaining sample synopses of evolving datasets," in *Proceedings of the International Conference on Very Large Data Bases*, pp. 595–606, 2006.

[116] R. Gemulla, W. Lehner, and P. J. Haas, "Maintaining Bernoulli samples over evolving multisets," in *Proceedings of ACM Principles of Database Systems*, pp. 93–102, 2007.

[117] R. Gemulla, W. Lehner, and P. J. Haas, "Maintaining bounded-size sample synopses of evolving datasets," *VLDB Journal*, vol. 17, no. 2, pp. 173–202, 2008.

[118] J. E. Gentle, *Random Number Generation and Monte Carlo Methods*. Springer, second ed., 2003.

[119] L. Getoor, B. Taskar, and D. Koller, "Selectivity estimation using probabilistic models," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 461–472, 2001.

[120] C. Giannella and B. Sayrafi, "An information theoretic histogram for single dimensional selectivity estimation," in *Proceedings of ACM Conference on Applications on Computing*, pp. 676–677, 2005.

[121] P. Gibbons, "Distinct sampling for highly-accurate answers to distinct values queries and event reports," in *International Conference on Very Large Data Bases*, 2001.

[122] P. Gibbons and S. Tirthapura, "Estimating simple functions on the union of data streams," in *ACM Symposium on Parallel Algorithms and Architectures (SPAA)*, 2001.

[123] P. Gibbons and S. Tirthapura, "Distributed streams algorithms for sliding windows," in *ACM Symposium on Parallel Algorithms and Architectures (SPAA)*, 2002.

[124] P. B. Gibbons and Y. Matias, "New sampling-based summary statistics for improving approximate query answers," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 331–342, New York, NY, USA, 1998.

[125] P. B. Gibbons, Y. Matias, and V. Poosala, "Aqua project white paper," Technical report, Bell Laboratories, Murray Hill, NJ, 1997.

[126] P. B. Gibbons, Y. Matias, and V. Poosala, "Fast incremental maintenance of approximate histograms," *ACM Transactions on Database Systems*, vol. 27, no. 3, pp. 261–298, 2002.

[127] A. Gilbert, S. Guha, P. Indyk, Y. Kotidis, S. Muthukrishnan, and M. Strauss, "Fast, small-space algorithms for approximate histogram maintenance," in *ACM Symposium on Theory of Computing*, 2002.

[128] A. Gilbert, Y. Kotidis, S. Muthukrishnan, and M. Strauss, "Surfing wavelets on streams: One-pass summaries for approximate aggregate queries," in *International Conference on Very Large Data Bases*, 2001.

[129] A. C. Gilbert, Y. Kotidis, S. Muthukrishnan, and M. Strauss, "How to summarize the universe: Dynamic maintenance of quantiles," in *International Conference on Very Large Data Bases*, 2002.

[130] A. C. Gilbert, Y. Kotidis, S. Muthukrishnan, and M. J. Strauss, "One-pass wavelet decomposition of data streams," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 3, pp. 541–554, May 2003.

[131] M. Greenwald and S. Khanna, "Space-efficient online computation of quantile summaries," in *ACM SIGMOD International Conference on Management of Data*, 2001.

[132] S. Guha, "A note on wavelet optimization," (Manuscript available from: `http://www.cis.upenn.edu/~sudipto/note.html`.), September 2004.

[133] S. Guha, "On the space-time of optimal, approximate and streaming algorithms for synopsis construction problems," *VLDB Journal*, vol. 17, no. 6, pp. 1509–1535, 2008.

[134] S. Guha and B. Harb, "Wavelet synopsis for data streams: Minimizing non-euclidean error," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Chicago, Illinois, August 2005.

[135] S. Guha and B. Harb, "Approximation algorithms for wavelet transform coding of data streams," in *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms*, Miami, Florida, January 2006.

[136] S. Guha, P. Indyk, S. Muthukrishnan, and M. J. Strauss, "Histogramming data streams with fast per-item processing," in *Proceedings of the International Colloquium on Automata, Languages, and Programming*, Malaga, Spain, July 2002.

[137] S. Guha, C. Kim, and K. Shim, "XWAVE: Optimal and approximate extended wavelets for streaming data," in *Proceedings of the International Conference on Very Large Data Bases*, Toronto, Canada, September 2004.

[138] S. Guha, N. Koudas, and K. Shim, "Approximation and streaming algorithms for histogram construction problems," *ACM Transactions on Database Systems*, vol. 31, no. 1, pp. 396–438, 2006.

[139] S. Guha, N. Koudas, and D. Srivastava, "Fast algorithms for hierarchical range histogram construction," in *Proceedings of ACM Principles of Database Systems*, pp. 180–187, 2002.

[140] S. Guha and K. Shim, "A note on linear time algorithms for maximum error histograms," *IEEE Transactions on Knowledge Data Engineering*, vol. 19, no. 7, pp. 993–997, 2007.

[141] S. Guha, K. Shim, and J. Woo, "REHIST: Relative error histogram construction algorithms," in *Proceedings of the International Conference on Very Large Data Bases*, pp. 300–311, 2004.

[142] D. Gunopulos, G. Kollios, V. J. Tsotras, and C. Domeniconi, "Approximating multi-dimensional aggregate range queries over real attributes," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 463–474, 2000.

[143] A. P. Gurajada and J. Srivastava, "Equidepth partitioning of a data set based on finding its medians," in *Proceedings of Applications of Computing*, pp. 92–101, 1991.

[144] P. J. Haas, "Large-sample and deterministic confidence intervals for online aggregation," in *Proceedings of International Conference on Scientific and Statistical Database Management*, pp. 51–63, 1997.

[145] P. J. Haas, "The need for speed: Speeding up DB2 UDB using sampling," *IDUG Solutions Journal*, vol. 10, no. 2, pp. 32–34, 2003.

[146] P. J. Haas and J. M. Hellerstein, "Join algorithms for online aggregation," IBM Research Report RJ 10126, 1998.

[147] P. J. Haas and J. M. Hellerstein, "Ripple joins for online aggregation," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 287–298, 1999.

[148] P. J. Haas, I. F. Ilyas, G. M. Lohman, and V. Markl, "Discovering and exploiting statistical properties for query optimization in relational databases: A survey," *Statistical Analysis and Data Mining*, vol. 1, no. 4, pp. 223–250, 2009.

[149] P. J. Haas and C. König, "A bi-level bernoulli scheme for database sampling," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 275–286, 2004.

[150] P. J. Haas, Y. Liu, and L. Stokes, "An estimator of the number of species from quadrat sampling," *Biometrics*, vol. 62, pp. 135–141, 2006.

[151] P. J. Haas, J. F. Naughton, S. Seshadri, and L. Stokes, "Sampling-based estimation of the number of distinct values of an attribute," in *Proceedings of the International Conference on Very Large Data Bases*, pp. 311–322, 1995.

[152] P. J. Haas, J. F. Naughton, S. Seshadri, and A. N. Swami, "Fixed-precision estimation of join selectivity," in *Proceedings of ACM Principles of Database Systems*, pp. 190–201, 1993.

[153] P. J. Haas, J. F. Naughton, S. Seshadri, and A. N. Swami, "Selectivity and cost estimation for joins based on random sampling," *Journal of Computer and Systems Science*, vol. 52, no. 3, pp. 550–569, 1996.

[154] P. J. Haas, J. F. Naughton, and A. N. Swami, "On the relative cost of sampling for join selectivity estimation," in *Proceedings of ACM Principles of Database Systems*, pp. 14–24, 1994.

[155] P. J. Haas and L. Stokes, "Estimating the number of classes in a finite population," *Journal of American Statistical Association*, vol. 93, no. 444, pp. 1475–1487, 1998.

[156] P. J. Haas and A. N. Swami, "Sequential sampling procedures for query size estimation," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 341–350, New York, NY, USA, 1992.

[157] T. Hagerup and C. Rüb, "A guided tour of chernoff bounds," *Information on Processing Letters*, vol. 33, no. 6, pp. 305–308, 1990.

[158] P. Hall and C. Heyde, *Martingale Limit Theory and Its Application.* Academic Press, 1980.

[159] M. Hansen, "Some history and reminiscences on survey sampling," in *Statistical Science*, vol. 2, pp. 180–190, 1987.

[160] B. Harb, "Algorithms for linear and nonlinear approximation of large data," PhD Thesis, University of Pennsylvania, 2007.

[161] N. J. A. Harvey, J. Nelson, and K. Onak, "Sketching and streaming entropy via approximation theory," in *Proceedings of the IEEE Conference on Foundations of Computer Science*, 2008.

[162] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inferences, and Prediction.* Springer, 2001.

[163] Z. He, B. S. Lee, and X. S. Wang, "Proactive and reactive multi-dimensional histogram maintenance for selectivity estimation," *Journal of Systems and Software*, vol. 81, no. 3, pp. 414–430, 2008.

[164] J. M. Hellerstein, P. J. Haas, and H. J. Wang, "Online aggregation," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 171–182, 1997.

[165] M. Henzinger, "Algorithmic challenges in search engines," *Internet Mathematics*, vol. 1, no. 1, pp. 115–126, 2003.

[166] M. Henzinger, P. Raghavan, and S. Rajagopalan, "Computing on data streams," Technical Report SRC 1998-011, DEC Systems Research Centre, 1998.

[167] J. Hershberger, N. Shrivastava, S. Suri, and C. D. Tóth, "Adaptive spatial partitioning for multidimensional data streams," in *Proceedings of ISAAC*, pp. 522–533, 2004.

[168] C. Hidber, "Online association rule mining," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 145–156, 1999.

[169] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, no. 301, p. 1330, 1963.

[170] D. G. Horvitz and D. J. Thompson, "A generalization of sampling without replacement from a finite universe," *Journal of the American Statistical Association*, vol. 47, pp. 663–695, 1952.

[171] W.-C. Hou, G. Özsoyoglu, and B. K. Taneja, "Statistical estimators for relational algebra expressions," in *Proceedings of ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pp. 276–287, 1988.

[172] W.-C. Hou, G. Özsoyoglu, and B. K. Taneja, "Processing aggregate relational queries with hard time constraints," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 68–77, 1989.

[173] IDC, "The diverse and exploding digital universe," IDC White Paper, March 2008.

[174] P. Indyk, "Stable distributions, pseudorandom generators, embeddings and data stream computation," in *IEEE Conference on Foundations of Computer Science*, 2000.

[175] P. Indyk and D. Woodruff, "Tight lower bounds for the distinct elements problem," in *IEEE Conference on Foundations of Computer Science*, 2003.

[176] P. Indyk and D. P. Woodruff, "Optimal approximations of the frequency moments of data streams," in *ACM Symposium on Theory of Computing*, 2005.

[177] Y. E. Ioannidis, "Universality of serial histograms," in *Proceedings of the International Conference on Very Large Data Bases*, pp. 256–267, 1993.

[178] Y. E. Ioannidis, "Approximations in database systems," in *Proceedings of the International Conference on Database Theory*, pp. 16–30, 2003.

[179] Y. E. Ioannidis, "The history of histograms (abridged)," in *Proceedings of the International Conference on Very Large Data Bases*, pp. 19–30, 2003.

[180] Y. E. Ioannidis and S. Christodoulakis, "On the propagation of errors in the size of join results," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 268–277, 1991.

[181] Y. E. Ioannidis and S. Christodoulakis, "Optimal histograms for limiting worst-case error propagation in the size of join results," *ACM Transactions on Database Systems*, vol. 18, no. 4, 1993.

[182] Y. E. Ioannidis and V. Poosala, "Balancing histogram optimality and practicality for query result size estimation," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 233–244, 1995.

[183] Y. E. Ioannidis and V. Poosala, "Histogram-based approximation of set-valued query-answers," in *Proceedings of the International Conference on Very Large Data Bases*, pp. 174–185, 1999.

[184] H. V. Jagadish, H. Jin, B. C. Ooi, and K.-L. Tan, "Global optimization of histograms," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 223–234, 2001.

[185] H. V. Jagadish, N. Koudas, S. Muthukrishnan, V. Poosala, K. C. Sevcik, and T. Suel, "Optimal histograms with quality guarantees," in *Proceedings of the International Conference on Very Large Data Bases*, pp. 275–286, 1998.

[186] B. Jawerth and W. Sweldens, "An overview of wavelet based multiresolution analyses," *SIAM Review*, vol. 36, no. 3, pp. 377–412, 1994.

[187] T. S. Jayram, R. Kumar, and D. Sivakumar, "The one-way communication complexity of gap hamming distance," `http://www.madalgo.au.dk/img/SumSchoo2007_Lecture_20slides/Bibliography/p14_Jayram_07_Manusc_ghd.pdf`, 2007.

[188] C. Jermaine, S. Arumugam, A. Pol, and A. Dobra, "Scalable approximate query processing with the DBO engine," in *ACM SIGMOD International Conference on Management of Data*, 2007.

[189] C. Jermaine, A. Dobra, S. Arumugam, S. Joshi, and A. Pol, "The sort-merge-shrink join," *ACM Transactions on Database Systems*, vol. 31, no. 4, pp. 1382–1416, 2006.

[190] C. Jermaine, A. Dobra, A. Pol, and S. Joshi, "Online estimation for subset-based sql queries," in *Proceedings of the International Conference on Very Large Data Bases*, pp. 745–756, 2005.

[191] C. Jermaine, A. Pol, and S. Arumugam, "Online maintenance of very large random samples," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 299–310, 2004.

[192] C. Jin, W. Qian, C. Sha, J. X. Yu, and A. Zhou, "Dynamically maintaining frequent items over a data stream," in *Proceedings of the ACM Conference on Information and Knowledge Management*, 2003.

[193] R. Jin, L. Glimcher, C. Jermaine, and G. Agrawal, "New sampling-based estimators for OLAP queries," in *Proceedings of the International Conference on Data Engineering*, p. 18, Washington, DC, USA, 2006.

[194] W. Johnson and J. Lindenstrauss, “Extensions of Lipshitz mapping into Hilbert space,” *Contemporary Mathematics*, vol. 26, pp. 189–206, 1984.

[195] S. Joshi and C. Jermaine, “Sampling-based estimators for subset-based queries,” *VLDB Journal*, Accepted for Publication, 2008.

[196] H. Jowhari, M. Saglam, and G. Tardos, “Tight bounds for lp samplers, finding duplicates in streams, and related problems,” in *ACM Principles of Database Systems*, 2011.

[197] C.-C. Kanne and G. Moerkotte, “Histograms reloaded: The merits of bucket diversity,” in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 663–674, 2010.

[198] P. Karras, “Optimality and scalability in lattice histogram construction,” *PVLDB*, vol. 2, no. 1, pp. 670–681, 2009.

[199] P. Karras and N. Mamoulis, “Hierarchical synopses with optimal error guarantees,” *ACM Transactions on Database Systems*, vol. 33, no. 3, August 2008.

[200] P. Karras and N. Mamoulis, “Lattice histograms: A resilient synopsis structure,” in *Proceedings of the International Conference on Data Engineering*, pp. 247–256, 2008.

[201] P. Karras and N. Manoulis, “One-pass wavelet synopses for maximum-error metrics,” in *Proceedings of the International Conference on Very Large Data Bases*, Trondheim, Norway, September 2005.

[202] P. Karras and N. Manoulis, “The Haar$^+$ tree: A refined synopsis data structure,” in *Proceedings of the International Conference on Data Engineering*, Istanbul, Turkey, April 2007.

[203] P. Karras, D. Sacharidis, and N. Manoulis, “Exploiting duality in summarization with deterministic guarantees,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Jose, California, August 2007.

[204] R. Kaushik, J. F. Naughton, R. Ramakrishnan, and V. T. Chakaravarthy, “Synopses for query optimization: A space-complexity perspective,” *ACM Transactions on Database Systems*, vol. 30, no. 4, pp. 1102–1127, 2005.

[205] R. Kaushik and D. Suciu, “Consistent histograms in the presence of distinct value counts,” *PVLDB*, vol. 2, no. 1, pp. 850–861, 2009.

[206] D. Kempe, A. Dobra, and J. Gehrke, “Gossip-based computation of aggregate information,” in *Proceedings of the IEEE Conference on Foundations of Computer Science*, pp. 482–491, 2003.

[207] S. Khanna, S. Muthukrishnan, and S. Skiena, “Efficient array partitioning,” in *Proceedings of the International Colloquium on Automata, Languages and Programming*, pp. 616–626, 1997.

[208] A. C. König and G. Weikum, “Combining histograms and parametric curve fitting for feedback-driven query result-size estimation,” in *Proceedings of the International Conference on Very Large Data Bases*, pp. 423–434, 1999.

[209] F. Korn, T. Johnson, and H. V. Jagadish, “Range selectivity estimation for continuous attributes,” in *Proceedings of the International Conference on Scientific and Statistical Database Management*, pp. 244–253, 1999.

[210] E. Kushilevitz and N. Nisan, *Communication Complexity*. Cambridge University Press, 1997.

[211] Y.-K. Lai and G. T. Byrd, "High-throughput sketch update on a low-power stream processor," in *Proceedings of the ACM/IEEE Symposium on Architecture for Networking and Communications Systems*, 2006.

[212] M. R. Leadbetter, G. Lindgren, and H. Rootzen, *Extremes and Related Properties of Random Sequences and Processes: Springer Series in Statistics*. Springer, 1983.

[213] G. M. Lee, H. Liu, Y. Yoon, and Y. Zhang, "Improving sketch reconstruction accuracy using linear least squares method," in *Internet Measurement Conference (IMC)*, 2005.

[214] L. Lim, M. Wang, and J. S. Vitter, "SASH: A self-adaptive histogram set for dynamically changing workloads," in *Proceedings of the International Conference on Very Large Data Bases*, pp. 369–380, 2003.

[215] L. Lim, M. Wang, and J. S. Vitter, "CXHist: An on-line classification-based histogram for XML string selectivity estimation," in *Proceedings of the International Conference on Very Large Data Bases*, pp. 1187–1198, 2005.

[216] X. Lin and Q. Zhang, "Error minimization for approximate computation of range aggregate," in *Proceedings of the International Conference on Database Systems for Advanced Applications*, pp. 165–172, 2003.

[217] Y. Lu, A. Montanari, B. Prabhakar, S. Dharmapurikar, and A. Kabbani, "Counter braids: A novel counter architecture for per-flow measurement," in *SIGMETRICS*, 2008.

[218] G. Luo, C. J. Ellmann, P. J. Haas, and J. F. Naughton, "A scalable hash ripple join algorithm," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 252–262, New York, NY, USA, 2002.

[219] J. Luo, X. Zhou, Y. Zhang, H. T. Shen, and J. Li, "Selectivity estimation by batch-query based histogram and parametric method," in *Proceedings of Australasian Database Conference on ADC2007*, pp. 93–102, 2007.

[220] MADlib library for scalable analytics, `http://madlib.net`.

[221] S. Mallat, *A Wavelet Tour of Signal Processing*. Academic Press, Second ed., 1999.

[222] Y. Matias and D. Urieli, "Optimal workload-based weighted wavelet synopses," in *Proceedings of the International Conference on Database Theory*, Edinburgh, Scotland, January 2005.

[223] Y. Matias and D. Urieli, "Optimal workload-based weighted wavelet synopses," *Theoretical Computer Science*, vol. 371, no. 3, pp. 227–246, 2007.

[224] Y. Matias, J. S. Vitter, and M. Wang, "Wavelet-based histograms for selectivity estimation," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 448–459, Seattle, Washington, June 1998.

[225] Y. Matias, J. S. Vitter, and M. Wang, "Dynamic maintenance of wavelet-based histograms," in *Proceedings of the International Conference on Very Large Data Bases*, Cairo, Egypt, September 2000.

[226] A. Metwally, D. Agrawal, and A. E. Abbadi, "Why go logarithmic if we can go linear? Towards effective distinct counting of search traffic," in *International Conference on Extending Database Technology*, 2008.

[227] Microsoft. Microsoft StreamInsight. `http://msdn.microsoft.com/en-us/library/ee362541.aspx`, 2008.

[228] J. Misra and D. Gries, "Finding repeated elements," *Science of Computer Programming*, vol. 2, pp. 143–152, 1982.

[229] M. Mitzenmacher and E. Upfal, *Probability and Computing: Randomized Algorithms and Probabilistic Analysis.* CUP, 2005.

[230] G. Moerkotte, T. Neumann, and G. Steidl, "Preventing bad plans by bounding the impact of cardinality estimation errors," *PVLDB*, vol. 2, no. 1, pp. 982–993, 2009.

[231] M. Monemizadeh and D. P. Woodruff, "1-pass relative-error l_p-sampling with applications," in *ACM-SIAM Symposium on Discrete Algorithms*, 2010.

[232] R. Motwani and P. Raghavan, *Randomized Algorithms.* Cambridge University Press, 1995.

[233] J. I. Munro and M. S. Paterson, "Selection and sorting with limited storage," *Theoretical Computer Science*, vol. 12, pp. 315–323, 1980.

[234] M. Muralikrishna and D. J. DeWitt, "Equi-depth histograms for estimating selectivity factors for multi-dimensional queries," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 28–36, 1988.

[235] S. Muthukrishnan, "Subquadratic algorithms for workload-aware Haar wavelet synopses," in *Proceedings of the International Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS)*, Hyderabad, India, December 2005.

[236] S. Muthukrishnan, "Data streams: Algorithms and applications," *Foundations and Trends® in Theoretical Computer Science*, vol. 1, no. 2, pp. 117–236, 2005.

[237] S. Muthukrishnan, V. Poosala, and T. Suel, "On rectangular partitionings in two dimensions: Algorithms, complexity, and applications," in *Proceedings of the International Conference on Database Theory*, pp. 236–256, 1999.

[238] S. Muthukrishnan and M. Strauss, "Maintenance of multidimensional histograms," in *Proceedings of the International Conference on Foundations Software Technical and Theoretical Computer Science (FSTTCS)*, vol. 2914 of Lecture Notes in Computer Science, pp. 352–362, Springer, 2003.

[239] S. Muthukrishnan, M. Strauss, and X. Zheng, "Workload-optimal histograms on streams," in *Proceedings of ESA*, pp. 734–745, 2005.

[240] J. Neyman, "On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection," *Journal of the Royal Statistical Society*, vol. 97, pp. 558–625, 1934.

[241] A. O'Hagan and J. J. Forster, *Bayesian Inference.* Volume 2B of *Kendall's Advanced Theory of Statistics.* Arnold, second ed., 2004.

[242] F. Olken, "Random sampling from databases," Technical Report LBL-32883, Lawrence Berekeley National Laboratory, 1993.

[243] F. Olken and D. Rotem, "Simple random sampling from relational databases," in *Proceedings of the International Conference on Very Large Data Bases*, pp. 160–169, 1986.

[244] F. Olken and D. Rotem, "Random sampling from B+ trees," in *Proceedings of the International Conference on Very Large Data Bases*, pp. 269–277, 1989.

[245] F. Olken and D. Rotem, "Maintenance of materialized views of sampling queries," in *Proceedings of the International Conference on Data Engineering*, pp. 632–641, 1992.

[246] F. Olken, D. Rotem, and P. Xu, "Random sampling from hash files," *SIGMOD Record*, vol. 19, no. 2, pp. 375–386, 1990.

[247] C. Pang, Q. Zhang, D. Hansen, and A. Maeder, "Unrestricted wavelet synopses under maximum error bound," in *Proceedings of the International Conference on Extending Database Technology*, Saint Petersburg, Russia, March 2009.

[248] N. Pansare, V. Borkar, C. Jermaine, and T. Condie, "Online aggregation for large MapReduce jobs," *PVLDB*, vol. 5, 2011. (To appear).

[249] A. Pavan and S. Tirthapura, "Range-efficient counting of distinct elements in a massive data stream," *SIAM Journal on Computing*, vol. 37, no. 2, pp. 359–379, 2007.

[250] H. T. A. Pham and K. C. Sevcik, "Structure choices for two-dimensional histogram construction," in *Proceedings of CASCON*, pp. 13–27, 2004.

[251] G. Piatetsky-Shapiro and C. Connell, "Accurate estimation of the number of tuples satisfying a condition," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 256–276, 1984.

[252] N. Polyzotis and M. N. Garofalakis, "XCluster synopses for structured XML content," in *Proceedings of the International Conference on Data Engineering*, p. 63, 2006.

[253] N. Polyzotis and M. N. Garofalakis, "XSKETCH synopses for XML data graphs," *ACM Transactions on Database Systems*, vol. 31, no. 3, pp. 1014–1063, 2006.

[254] V. Poosala and V. Ganti, "Fast approximate answers to aggregate queries on a data cube," in *Proceedings of the International Conference on Scientific and Statistical Database Management*, pp. 24–33, 1999.

[255] V. Poosala, V. Ganti, and Y. E. Ioannidis, "Approximate query answering using histograms," *IEEE Data Engineering Bulletins*, vol. 22, no. 4, pp. 5–14, 1999.

[256] V. Poosala and Y. E. Ioannidis, "Selectivity estimation without the attribute value independence assumption," in *Proceedings of the International Conference on Very Large Data Bases*, pp. 486–495, 1997.

[257] V. Poosala, Y. E. Ioannidis, P. J. Haas, and E. J. Shekita, "Improved histograms for selectivity estimation of range predicates," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 294–305, 1996.

[258] L. Qiao, D. Agrawal, and A. E. Abbadi, "RHist: Adaptive summarization over continuous data streams," in *Proceedings of the ACM Conference on Information and Knowledge Management*, pp. 469–476, 2002.

[259] V. Raman and J. M. Hellerstein, "Potter's Wheel: An interactive data cleaning system," in *Proceedings of the International Conference on Very Large Data Bases*, pp. 381–390, 2001.

[260] V. Raman, B. Raman, and J. M. Hellerstein, "Online dynamic reordering," *VLDB Journal*, vol. 9, no. 3, pp. 247–260, 2000.

[261] V. Raman, G. Swart, L. Qiao, F. Reiss, V. Dialani, D. Kossmann, I. Narang, and R. Sidle, "Constant-time query processing," in *Proceedings of the International Conference on Data Engineering*, pp. 60–69, 2008.

[262] R. L. Read, D. S. Fussell, and A. Silberschatz, "Computing bounds on aggregate operations over uncertain sets using histograms," in *Proceedings of*

Post-ILPS '94 Workshop on Uncertainty in Databases and Deductive Systems, pp. 107–117, 1994.
[263] F. Reiss, M. N. Garofalakis, and J. M. Hellerstein, "Compact histograms for hierarchical identifiers," in *Proceedings of the International Conference on Very Large Data Bases*, pp. 870–881, 2006.
[264] J. Rissanen, "Stochastic complexity and modeling," *Annals of Statistics*, vol. 14, no. 3, pp. 1080–1100, 1986.
[265] F. Rusu and A. Dobra, "Sketches for size of join estimation," *ACM Transactions on Database Systems*, vol. 33, no. 3, 2008.
[266] D. Sacharidis, A. Deligiannakis, and T. Sellis, "Hierarchically-compressed wavelet synopses," *The VLDB Journal*, vol. 18, no. 1, pp. 203–231, January 2009.
[267] A. D. Sarma, O. Benjelloun, A. Y. Halevy, S. U. Nabar, and J. Widom, "Representing uncertain data: Models, properties, and algorithms," *VLDB Journal*, vol. 18, no. 5, pp. 989–1019, 2009.
[268] C.-E. Sarndal, B. Swennson, and J. Wretman, *Model-Assisted Survey Sampling*. Springer, second ed., 1992.
[269] R. T. Schweller, Z. Li, Y. Chen, Y. Gao, A. Gupta, Y. Zhang, P. A. Dinda, M.-Y. Kao, and G. Memik, "Reversible sketches: Enabling monitoring and analysis over high-speed data streams," *IEEE Transactions on Networks*, vol. 15, no. 5, 2007.
[270] D. W. Scott, *Multivariate Density Estimation: Theory Practice, and Visualization*. Wiley, 1992.
[271] S. Seshadri, "Probabilistic methods in query processing," PhD Thesis, University of Wisconsin at Madison, Madison, WI, USA, 1992.
[272] J. Shao and D. Tu, *The Jackknife and Bootstrap*. Springer Series in Statistics, 1995.
[273] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986.
[274] J. Spiegel and N. Polyzotis, "TuG synopses for approximate query answering," *ACM Transactions on Database Systems*, vol. 34, no. 1, 2009.
[275] U. Srivastava, P. J. Haas, V. Markl, and N. Megiddo, "ISOMER: Consistent histogram construction using query feedback," in *Proceedings of the International Conference on Data Engineering*, 2006.
[276] E. J. Stollnitz, T. D. DeRose, and D. H. Salesin, *Wavelets for Computer Graphics — Theory and Applications*. San Francisco, CA: Morgan Kaufmann Publishers, 1996.
[277] M. Stonebraker, J. Becla, D. J. DeWitt, K.-T. Lim, D. Maier, O. Ratzesberger, and S. B. Zdonik, "Requirements for science data bases and scidb," in *Proceedings of CIDR*, 2009.
[278] N. Thaper, P. Indyk, S. Guha, and N. Koudas, "Dynamic multidimensional histograms," in *ACM SIGMOD International Conference on Management of Data*, 2002.
[279] D. Thomas, R. Bordawekar, C. C. Aggarwal, and P. S. Yu, "On efficient query processing of stream counts on the cell processor," in *IEEE International Conference on Data Engineering*, 2009.

[280] M. Thorup, "Even strongly universal hashing is pretty fast," in *ACM-SIAM Symposium on Discrete Algorithms*, 2000.

[281] M. Thorup and Y. Zhang, "Tabulation based 4-universal hashing with applications to second moment estimation," in *ACM-SIAM Symposium on Discrete Algorithms*, 2004.

[282] S. Venkataraman, D. X. Song, P. B. Gibbons, and A. Blum, "New streaming algorithms for fast detection of superspreaders," in *Network and Distributed System Security Symposium NDSS*, 2005.

[283] J. S. Vitter, "Random sampling with a reservoir," *ACM Transactions on Mathematical Software*, vol. 11, no. 1, pp. 37–57, 1985.

[284] J. S. Vitter and M. Wang, "Approximate computation of multidimensional aggregates of sparse data using wavelets," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Philadelphia, Pennsylvania, May 1999.

[285] M. P. Wand and M. C. Jones, *Kernel Smoothing*. Chapman and Hall, 1995.

[286] H. Wang and K. C. Sevcik, "Utilizing histogram information," in *Proceedings of CASCON*, p. 16, 2001.

[287] H. Wang and K. C. Sevcik, "A multi-dimensional histogram for selectivity estimation and fast approximate query answering," in *Proceedings of CASCON*, pp. 328–342, 2003.

[288] H. Wang and K. C. Sevcik, "Histograms based on the minimum description length principle," *VLDB Journal*, vol. 17, no. 3, 2008.

[289] K. Y. Whang, B. T. Vander-Zanden, and H. M. Taylor, "A linear-time probabilistic counting algorithm for database applications," *ACM Transactions on Database Systems*, vol. 15, no. 2, p. 208, 1990.

[290] J. R. Wieland and B. L. Nelson, "How simulation languages should report results: A modest proposal," in *Proceedings of Winter Simulation Conference*, pp. 709–715, 2009.

[291] R. Winter and K. Auerbach, "The big time: 1998 winter VLDB survey," *Database Programming and Design*, August 1998.

[292] D. Woodruff, "Optimal space lower bounds for all frequency moments," in *ACM-SIAM Symposium on Discrete Algorithms*, 2004.

[293] M. Wu and C. Jermaine, "A bayesian method for guessing the extreme values in a data set," in *Proceedings of the International Conference on Very Large Data Bases*, pp. 471–482, 2007.

[294] K. Yi, F. Li, M. Hadjieleftheriou, G. Kollios, and D. Srivastava, "Randomized synopses for query assurance on data streams," in *IEEE International Conference on Data Engineering*, 2008.

[295] Q. Zhang and X. Lin, "On linear-spline based histograms," in *Proceedings of the International Conference on Web-Age Information Management*, pp. 354–366, 2002.

[296] Q. Zhang and W. Wang, "A fast algorithm for approximate quantiles in high speed data streams," in *Proceedings of the International Conference on Scientific and Statistical Database Management*, p. 29, 2007.

[297] X. Zhang, M. L. King, and R. J. Hyndman, "Bandwidth selection for multivariate kernel density estimation using MCMC," in *Econometric Society 2004 Australasian Meetings*, 2004.

[298] Q. Zhu and P.-Å. Larson, "Building regression cost models for multidatabase systems," in *Proceedings of the International Conference on Parallel and Distributed Information Systems*, pp. 220–231, 1996.
[299] V. M. Zolotarev, *One Dimensional Stable Distributions*, volume 65 of *Translations of Mathematical Monographs*. American Mathematical Society, 1983.