

## **Basic techniques**

- Vector spaces
- Transformations: SVD, FFT, Wavelets, Distance metric
  - Linear Algebra and its applications, G. Strang,
  - [A brief guide to wavelet sources](#)
  - [Wavelets in Computer Graphics: A Primer](#) E.J. Stollnitz, T. D. DeRose, D.H. Salesin
- Hash tables, B-trees
  - Undergraduate database textbook
  - Appendix of textbook

## **Dimensionality reduction**

- C. Faloutsos and K.-I. Lin, "FastMap: A Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets", *Proc. ACM SIGMOD*, 1995, 163-174.
- G. Hjaltson and H. Samet, Properties of Embedding Methods for Similarity Searching in Metric Spaces, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5): 530-549 (May 2003).
- E. Bingham and H. Mannila, "Random projection in dimensionality reduction: Applications to image and text data, KDD 2001: 245-250.
- J. Bourgain, "On Lipschitz embedding of finite metric spaces in Hilbert space", *Israel J. of Math.*, 52, 1985, 46-52.
- N. Linial, E. London and Y. Rabinovich, "The Geometry of Graphs and some of its algorithmic applications", *Combinatorica*, 15, 1995, 215-245.
- D. Achlioptas, "Database-friendly random projections", PODS 2001.
- W.B. Johnson and J. Lindenstrauss, "Extensions of Lipschitz maps into Hilbert spaces", *Contemporary Mathematics*, 29:189-206.
- C.H. Papadimitriou, P. Raghvan, H. Tamaki, and S. Vempala, "Latent Semantic Indexing: A probabilistic analysis", PODS 1998: 159-168.
- P. Indyk and R. Motwani, "Towards removing the curse of dimensionality", STOC 1998: 604-613.
- J. Buhler and M. Tompa, "Finding motifs using random projections", *JCB*, 9(2), pp. 225-242, 2002

## **Index structures**

- The R\*-tree: An Efficient and Robust Access Method for Points and Rectangles, N. Beckmann, H.-P. Kriegel, R. Schneider, and B. Seeger, SIGMOD 1990: 322-331. Available from ACM Digital Library
- M-tree: An efficient access method for similarity search in metric spaces, P. Ciaccia, M. Patella, and P. Zezula, VLDB 1997.
- Optimal multi-step k-nearest neighbor search, T. Seidl and H. Kriegel, SIGMOD 1998: 154-165.
- [A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces](#), R. Weber and H.-J. Schek and S. Blott, VLDB 1998.
- [The pyramid technique: Towards breaking the curse of dimensionality](#), B.S. Berchtold, C. Bohm, and H.-P. Kriegel, SIGMOD 1998: 142-153.
- [R-Trees: A Dynamic Index Structure for Spatial Searching](#), A. Guttman, Proc. ACM SIGMOD International Conference on Management of Data (1984): 47-57.
- G. R. Hjaltason and H. Samet, [Distance browsing in spatial databases](#), *ACM Transactions on Database Systems* 24, 2 (June 1999), 265-318

- A cost model for nearest neighbor search in high-dimensional data space, S. Berchtold, C. Bohm, D. Keim, and H.-P. Kriegel, ACM Symp. on Principles of Database Systems (PODS), 1997.
- On the analysis of indexing schemes, J. M. Hellerstein, E. Koutsoupias, and C.H. Papadimitriou, PODS 1997, pages 249-256.
- Linear clustering of objects with multiple attributes, H.V. Jagadish, SIGMOD 1990: 332—342
- Dimensionality reduction for similarity searching in dynamic databases, K.V. Ravi Kanth, D. Agrawal, A. El Abbadi, and A.K. Singh, Computer Vision and Image Understanding, 75(1/2), July/August 1999, pp. 59-72.
- When is "Nearest Neighbor" meaningful?, K. S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, ICDT 1999, pp. 217—235. C. Lang and A. K. Singh, "Modeling high dimensional index structures using sampling", SIGMOD 2001.
- (X-tree) S. Berchtold, D. A. Keim, and H.-P. Kriegel. The X-tree: An Index Structure for High-Dimensional Data. In *Proc. of the 22nd International Conference on Very Large Data Bases (VLDB)*, pp. 28–39, Bombay, September 1996.
- (K-D-B tree) J. T. Robinson. *The kdb-tree: A search structure for large multi-dimensional dynamic indexes*. In Proc. ACM SIGMOD Int. Conf. on Management of Data, pages 10--18, 1981.
- (Hilbert R-tree) I. Kamel and C. Faloutsos. Hilbert R-tree: An Improved R-tree using Fractals. In *Proceedings of the Twentieth International Conference on Very Large Databases*, pp. 500–509, Santiago, Chile, 1994.
- (SS tree) D. A. White and R. Jain. Similarity Indexing with the SS-tree. In *Proc. of IEEE 12th International Conference on Data Engineering*, pp. 516–523, 1996. (SR-tree) N. Katayama and S. Satoh.
- The SR-tree: An Index Structure for High-Dimensional Nearest Neighbor Queries. In *Proc. ACM SIGMOD International Conference on Management of Data*, pp. 369–380, May 1997.
- (VAMSplit R-tree) D. A. White and R. Jain. Similarity Indexing: Algorithms and Performance. In SPIE: Storage and Retrieval for Image and Video Databases IV , pages 62{75, 1996. Also available on the WWW at url <http://vision.ucsd.edu/papers/sindexalg>.
- (geometric hashing) H. Wolfson and I. Rigoutsos, Geometric hashing: an overview", IEEE Computational Science and Engineering, 1997.

## **Sampling, histograms, and sketches**

- Optimal Histograms with Quality Guarantees, H. V. Jagadish, N. Koudas, S. Muthukrishnan, V. Poosala, K. C. Sevcik, and T. Suel, VLDB 1998
- Surfing wavelets on streams: one-pass summaries for approximate aggregate queries, A.C. Gilbert, Y. Kotidis, S. Muthukrishnan, M.J. Strauss, VLDB 2001.
- New Sampling-Based Summary Statistics for Improving Approximate Query Answers, P.B. Gibbons and Y. Matias, ACM SIGMOD 1998.
- G.S. Manku, S. Rajagopalan, and B. G. Lindsay. "Approximate Medians and other Quantiles in One Pass and with Limited Memory". ACM SIGMOD 1998.
- N. Alon, Y. Matias, M. Szegedy. The space complexity of approximating the frequency moments. ACM STOC, 1996.
- P. B. Gibbons, Y. Matias, and V. Poosala. Fast Incremental Maintenance of Approximate Histograms, VLDB 1997
- P. Indyk. Stable Distributions, Pseudorandom Generators, Embeddings, and Data Stream Computation. IEEE FOCS, 2000
- W.B. Johnson, J. Lindenstrauss. Extensions of Lipschitz Mapping into Hilbert space. Contemporary Mathematics, 26, 1984.
- Y. Matias, J.S. Vitter, and M. Wang. "Wavelet-based Histograms for Selectivity Estimation". ACM SIGMOD 1998.

- Y. Matias, J.S. Vitter, and M. Wang. “Dynamic Maintenance of Wavelet-based Histograms”. VLDB 2000. V. Poosala, Y. Ioannidis, P. Haas, and E. Shekita. “Improved Histograms for Selectivity Estimation of Range Predicates”. ACM SIGMOD 1996.
- Improved histograms for selectivity estimation of range predicates. V. Poosala, Y. Ioannidis, P. Haas, and E. Shekita. *Proc. of ACM SIGMOD Conf*, pages 294–305, June 1996.
- Wavelet Synopses with Error Guarantees, M. Garofalakis and P.B. Gibbons. In *Proc. of the 1998 ACM SIGMOD Intl. Conf. on Management of Data*, June 1998. Approximation Algorithms for Array Partitioning Problems. T. Suel and S. Muthukrishnan. *Journal of Algorithms*.
- Random sampling from databases---a survey. F. Olken and D. Rotem. <http://pueblo.lbl.gov/~olken/sampling.html>.
- Aqua: A fast decision support system using approximate query answers. S. Acharya, P. B. Gibbons, and V. Poosala. In *Proc. 1999 Intl. Conf. on Very Large Data Bases*, pages 754–755, Sept. 1999.
- Overcoming limitations of sampling for aggregation queries. S. Chaudhuri, G. Das, M. Datar, R. Motwani, and V. Narasayya. In *Proc. 2001 Intl. Conf. on Data Engineering*, 2001.
- Density Biased Sampling: An Improved Method for Data Mining and Clustering. C.R. Palmer and C. Faloutsos. In *Proc. ACM SIGMOD conference*, 2000.
- Dynamic Sample Selection for Approximate Query Processing. B. Babcock, S. Chaudhuri, G. Das. In *Proc. ACM SIGMOD conference*, 2003: 539-550.
- New Sampling-Based Summary Statistics for Improving Approximate Query Answers. P.B. Gibbons and Y. Matias. In *Proc. ACM SIGMOD conference*, 1998.
- A Robust, optimization-based approach for approximate answering of aggregate queries, S. Chaudhuri, G. Das, and V. Narasayya, In *Proc. ACM SIGMOD conference*, 2001. Congressional samples for approximate answering of group-by queries, S. Acharya, P.B. Gibbons, and V. Poosala, In *Proc. ACM SIGMOD conference*, 2000
- S. Guha, N. Koudas, and K. Shim. “Data Streams and Histograms”. *ACM STOC 2001*.

## Contemporary datasets

- (text) [Matrices, Vector Spaces, and Information Retrieval](#), M.W. Berry, Z. Drmac, E.R. Jessup, SIAM Review, 41(2), 1999, 335-362.
- (image) Efficient and effective querying by image content, C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic, and W. Equitz, *Journal of Intelligent Information Systems*, 3:231-262, 1994.
- (stream) Data-streams and histograms, S. Guha, N. Koudas and K. Shim, *ACM Symposium on Theory of Computing*, pages 471-475, 2001.
- (time-series) matching in time-series databases, C. Faloutsos and M. Ranganathan, and Y. Manolopoulos, *SIGMOD 1994*: pages 419-429.
- (spatio-temporal) •G. Kollios, D. Gunopulos, and V. J. Tsotras. On Indexing Mobile Objects. In *Proc. of the ACM Symp. on Principles of Database Systems, PODS*, pages 261–272, June 1999.
- (top-merge) Fagin et al., [Optimal Aggregation Algorithms for Middleware](#), PODS 2001.
- (internet) A.N. Langville and C.D. Meyer, A survey of eigenvector based methods for web information retrieval, 47(1), pp. 135-162, 2005
- (text) Indexing by latent semantic analysis, S. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, and R.A. Harshman, *Journal of the Society for Information Science*, 41(6), 1990, 391-407. <http://citeseer.nj.nec.com/deerwester90indexing.html>

- (audio) Content-based classification, search, and retrieval of audio, E. Wold, T. Blum, D. Keislar, and J. Wheaton, *IEEE Multimedia*, 3(3):27--36, 1996, <http://www.musclefish.com>
- (shape) Geometric-Similarity Retrieval in Large Image Bases, I. Fudos, L. Palios, E. Pitoura, *ICDE 2002*.
- (stream) Maintaining Stream Statistics over Sliding Windows, M. Datar, A. Gionis, P. Indyk and R. Motwani, *ACM Symposium on Discrete Algorithms*, 2002.
- (stream) Approximating a Data Stream for Querying and Estimation: Algorithms and Performance Evaluation, Sudipto Guha, Nick Koudas, *ICDE 2002*.
- (time-series) Similarity Search in Sequence Databases, R. Agrawal, C. Faloutsos, and A. Swami, *FODO conference*, Evanston, Illinois, Oct. 13-15, 1993, <http://citeseer.nj.nec.com/agrawal93efficient.html>
- (time-series) Variable Length Queries for Time Series Data, T. Kahveci and A.K. Singh, *International Conference on Data Engineering (ICDE) 2001*: 273-282, <http://citeseer.nj.nec.com/323328.html>
- (time-series) Fast Time-Series Searching with Scaling and Shifting, K. Chu and M.H. Wong, *PODS 1999*: 237-248.
- (image) Efficient Retrieval for Browsing Large Image Databases, D. Wu, D. Agrawal, A. El Abbadi, A. Singh, and T. R. Smith, <http://citeseer.nj.nec.com/wu96efficient.html>
- (spatio-temporal) Y. Tao and D. Papadias. MV3R-Tree: A Spatio-Temporal Access Method for Timestamp and Interval Queries. In *Proc. of the Intl. Conf. on Very Large Data Bases, VLDB*, pages 431–440, Sept. 2001.
- (spatio-temporal) Y. Tao and D. Papadias. Efficient Historical R-trees. In *Proc. of the Intl. Conf. on Scientific and Statistical Database Management, SSDBM*, pages 223–232, July 2001.
- (spatio-temporal) M. Hadjieleftheriou, G. Kollios, V. J. Tsotras, and D. Gunopulos. Efficient Indexing of Spatiotemporal Objects. In *Proc. of the Intl. Conf. on Extending Database Technology, EDBT*, pages 251–268, Czech Republic, Mar. 2002.
- (spatio-temporal) B. Becker, S. Gschwind, T. Ohler, B. Seeger, and P. Widmayer. An Asymptotically Optimal Multiversion B-Tree. *VLDB Journal*, 5(4):264–275, 1996.
- (spatio-temporal) Y. Theodoridis, M. Vazirgiannis, and T. Sellis. Spatio-Temporal Indexing for Large Multimedia Applications. In *Proc. of the IEEE Conference on Multimedia Computing and Systems, ICMCS*, June 1996.
- (stream) Graham Cormode, Flip Korn, S. Muthukrishnan, Divesh Srivastava: Diamond in the Rough: Finding Hierarchical Heavy Hitters in Multi-Dimensional Data, *SIGMOD 2004*.
- (top-merge) K. Chang and S.-W. Hwang. [Minimal Probing: Supporting Expensive Predicates for Top-k Queries](#), *SIGMOD 2002*.

## VLDB 2007:

- *A Bayesian Method for Guessing the Extreme Values in a Data Set*  
Mingxi Wu, Chris Jermaine
- *Efficient Computation of Reverse Skyline Queries*  
Evangelos Dellis, Bernhard Seeger
- *Efficient Processing of Top-k Dominating Queries on Multi-Dimensional Data*  
Man Lung Yiu, Nikos Mamoulis
- *Extending Q-Grams to Estimate Selectivity of String Matching with Low Edit Distance*  
Hongrae Lee, Raymond Ng, Kyuseok Shim
- *Fast nGram-Based String Search Over Data Encoded Using Algebraic Signatures*  
Witold Litwin, Riad Mokadem, Philippe Rigaux, Thomas Schwarz

- *Graph Indexing: Tree + Delta  $\geq$  Graph*  
Peixiang Zhao, Jeffrey Xu Yu, Philip Yu
- *Measuring the Structural Similarity of Semistructured Documents Using Entropy*  
Sven Helmer
- *Mining Approximate Top-K Subspace Anomalies in Multi-Dimensional Time-Series Data*  
Xiaolei Li, Jiawei Han
- *Modeling and Querying Vague Spatial Objects Using Shapelets*  
Daniel Zinn, Jim Bosch, Michael Gertz
- *On Efficient Spatial Matching*  
Raymond Chi-Wing Wong, Yufei Tao, Ada Wai-Chee Fu, Xiaokui Xiao
- *Peer-to-Peer Similarity Search in Metric Spaces*  
Christos Doukeridis, Akrivi Vlachou, Yannis Kotidis, Michalis Vazirgiannis
- *Probabilistic Skylines on Uncertain Data*  
Jian Pei, Bin Jiang, Xuemin Lin, Yidong Yuan
- *Ranked Subsequence Matching in Time-Series Databases*  
Wook-Shin Han, Jinsoo Lee, Yang-Sae Moon, Haifeng Jiang
- *Towards Graph Containment Search and Indexing*  
Chen Chen, Xifeng Yan, Philip Yu, Jiawei Han, Dong-Qing Zhang, Xiaohui Gu
- *VGRAM: Improving Performance of Approximate Queries on String Collections Using Variable-Length Grams*  
Chen Li, Bin Wang, Xiaochun Yang
- *Multi-Probe LSH: Efficient Indexing for High-Dimensional Similarity Search*  
Qin Lv, William Josephson, Zhe Wang, Moses Charikar, Kai Li
- *On Dominating Your Neighborhood Profitably*  
Cuiping Li, Anthony Tung, Wen Jin, Martin Ester
- *On Efficient Spatial Matching*  
Raymond Chi-Wing Wong, Yufei Tao, Ada Wai-Chee Fu, Xiaokui Xiao
- *Indexable PLA for Efficient Similarity Search*  
Qiuxia Chen, Lei Chen, Xiang Lian, Yunhao Liu, Jeffrey Xu Yu
- *MIST: Distributed Indexing and Querying in Sensor Networks using Statistical Models*  
Arnab Bhattacharya, Anand Meka, Ambuj Singh

#### **SIGMOD 2007:**

- *Fast Data Stream Algorithms Using Associative Memories*  
Nagender Bandi, Ahmed Metwally, Divyakant Agrawal, Amr El Abbadi (UC Santa Barbara)
- *Effective Variation Management for Pseudo Periodical Streams*  
Lv-an Tang, Bin Cui, Hongyan Li, Gaoshan Miao, Dongqing Yang, Xinbiao Zhou (Peking University)
- *Efficient Algorithms for Evaluating XPath over Streams*  
Gang Gou, Rada Chirkova (North Carolina State University)
- *Sketching Probabilistic Data Streams*

Graham Cormode (AT&T Labs, Research), Minos Garofalakis (Yahoo! Research & UC Berkeley)

- In-Network Execution of Monitoring Queries in Sensor Networks

Xiaoyan Yang (National University of Singapore), Hock Beng Lim (National University of Singapore) M. Tamer Ozsu (University of Waterloo), Kian Lee Tan (National University of Singapore)

- An Efficient and Accurate Method for Evaluating Time Series Similarity

Michael Morse, Jignesh M. Patel (University of Michigan)

- Genome-scale Disk-based Suffix Tree Indexing

Benjarath Phoophakdee, Mohammed J. Zaki (Rensselaer Polytechnic Institute)

- Fast and Practical Indexing and Querying of Very Large Graphs

Silke Trißl, Ulf Leser (Humboldt-Universität zu Berlin)

- FG-Index: Towards Verification-Free Query Processing on Graph Databases

James Cheng, Yiping Ke, Wilfred Ng, An Lu (The Hong Kong University of Science and Technology)

### **ICDE 2007:**

- A General Cost Model for Dimensionality Reduction in High Dimensional Spaces

Xiang Lian, Lei Chen

- SD-Rtree: A Scalable Distributed Rtree

Cedric du Mouza, Philippe Rigaux, Witold Litwin

- On k-Nearest Neighbor Searching in Non-Ordered Discrete Spaces

Dashiell Kolbe, Qiang Zhu, Sakti Pramanik

- The Haar+ Tree: a Refined Synopsis Data Structure

Panagiotis Karras, Nikos Mamoulis

- DIKNN: An Itinerary-based KNN Query Processing Algorithm for Mobile Sensor Networks

Brandon Wu, Kun-Ta Chuang, Chung-Min Chen, Ming-Syan Chen

- Labeling network motifs in protein interactomes for protein function prediction

Jin Chen, Wynne Hsu, Mong Li Lee, See Kiong Ng

- Topology Search over Biological Databases

Lin Guo, Jayavel Shanmugasundaram, Golan Yona

- GString: A Novel Approach for Efficient Search in Graph Databases

Haoliang Jiang, Haixun Wang, Philip S. Yu, Shuigeng Zhou

- Efficient Evaluation of Imprecise Location-Dependent Queries

Jinchuan Chen, Reynold Cheng

- Representing and Querying Correlated Tuples in Probabilistic Databases

prithviraj sen, Amol Deshpande

- $10^{10^6}$  Worlds and Beyond: Efficient Representation and Processing of Incomplete Information

Lyublena Antova, Christoph Koch, Dan Olteanu

- Indexing Uncertain Categorical Data  
Sarvjeet Singh, Chris Mayfield, Sunil Prabhakar, Rahul Shah, Susanne Hambruch
- SpADe: On Shape-based Pattern Detection in Streaming Time Series  
Yueguo Chen, Mario A. Nascimento, Beng Chin Ooi, Anthony K. H. Tung
- Multi-source Skyline Query Processing in Road Networks  
Ke Deng, Xiaofang Zhou, Heng Tao Shen
- Continuous Evaluation of Monochromatic and Bichromatic Reverse Nearest Neighbors  
James Kang, Mohamed Mokbel, Shashi Shekhar, Tian Xia, Donghui Zhang
- Index-based Most Similar Trajectory Search  
Elias Frentzos, Kostas Gratsias, Yannis Theodoridis
- Evaluating Proximity Relations Under Uncertainty  
Zhengdao Xu, Arno Jacobsen
- Efficient Top-k Query Evaluation on Probabilistic Data  
Christopher Re, Nilesh Dalvi, Dan Suciu
- Top-k Query Processing in Uncertain Databases  
Mohamed Soliman, Ihab Ilyas, Kevin C. Chang
- APLA: Indexing Arbitrary Probability Distributions  
Vebjorn Ljosa, Ambuj K. Singh
- Graph Database Indexing Using Structured Graph Decomposition  
David Williams, Jun Huan, Wei Wang
- Space Efficient Streaming Algorithms for the Maximum Error Histogram  
Chiranjeeb Buragohain, Nisheeth Shrivastava, Subhash Suri
- Conquering the Divide: Continuous Clustering of Distributed Data Streams  
Graham Cormode, S. Muthukrishnan, Wei Zhuang
- Stream Monitoring under the Time Warping Distance  
Yasushi Sakurai, Christos Faloutsos, Masashi Yamamuro
- Efficient Evaluation of All-Nearest-Neighbors Queries  
Yun Chen, Jignesh M. Patel
- Pointwise-Dense Region Queries in Spatio-temporal Databases  
Jinfeng Ni, Chinya Ravishankar
- Top-k Spatial Preference Queries  
Man Lung Yiu, Xiangyuan Dai, Nikos Mamoulis, Michail Vaitis
- Similarity Match Over High Speed Time-Series Streams  
Xiang Lian, Lei Chen, Jeffrey Xu Yu, Guoren Wang, Ge Yu