

Index

- A-Priori Algorithm, 224, 225, 231
- Accessible page, 199
- Accumulated sum, 507
- Accuracy impurity, 502
- Action, 44
- Active learning, 464
- Ad-hoc query, 146
- Adjacency matrix, 375
- Adomavicius, G., 352
- Advertising, 16, 128, 216, 293
- Adwords, 302
- Affiliation-graph model, 383
- Afrati, F.N., 77, 78, 420
- Agglomerative clustering, *see* Hierarchical clustering
- Aggregation, 34, 38
- Agrawal, R., 250
- Alexandrov, A., 78
- All-pairs problem, 64, 68
- Alon, N., 174
- Alon-Matias-Szegedy Algorithm, 158
- Amplification, 113
- Analytic query, 60
- AND-construction, 113
- Anderson, C., 352, 353
- Andoni, A., 141
- ANF, *see* Approximate neighborhood function
- ANF Algorithm, 414
- Apache, 24, 25, 78, 79
- Approximate neighborhood function, 414
- Arc, 400
- Archive, 144
- Ask, 204
- Association rule, 217, 219
- Associativity, 27, 45
- Attribute, 33
- Auction, 305
- Austern, M.H., 79
- Authority, 204
- Average, 156
- B-tree, 292
- Babcock, B., 174, 292
- Babu, S., 174
- Backstrom, L., 420
- Bad point, 483
- Bag, 40, 85
- Balance Algorithm, 305
- Balazinska, M., 78
- Band, 100
- Bandwidth, 22
- Basket, *see* Market basket, 214, 216, 217, 246
- Batch gradient descent, 489
- Batch learning, 463
- Bayes net, 4
- BDMO Algorithm, 283
- Beer and diapers, 218
- Bell, R., 353
- Bellkor's Pragmatic Chaos, 322
- Bergmann, R., 78
- Berkhin, P., 212
- Berrar, D.P., 455
- Betweenness, 363
- BFR Algorithm, 266, 269
- BFS, *see* Breadth-first search
- Bi-clique, 369
- Bickson, D., 79
- Bid, 303, 305, 312, 313
- BigTable, 77
- Bik, A.J.C., 79

- Binary Classification, 458
- Biomarker, 217
- Bipartite graph, 299, 359, 369, 370
- BIRCH Algorithm, 292
- Birrell, A., 79
- Bitmap, 232, 233
- Block, 11, 21, 192
- Blockeel, H., 515
- Blocking property, 43, 49
- Blog, 200
- Bloom filter, 152, 230
- Bloom, B.H., 174
- Blum, A., 515
- Bohannon, P., 78
- Boldi, P., 420
- Bonferroni correction, 5
- Bonferroni's principle, 4, 5
- Bookmark, 198
- Boral, H., 421
- Borkar, V., 77, 78
- Bottou, L., 515
- Bradley, P.S., 292
- Breadth-first search, 363
- Brick-and-mortar retailer, 216, 320, 321
- Brin, S., 212
- Broad matching, 305
- Broder, A.Z., 18, 141, 212
- Bu, Y., 78
- Bucket, 9, 149, 164, 168, 230, 283
- Budget, 304, 311
- Budiu, M., 79
- Bulk-synchronous system, 51
- Burges, C.J.C., 515
- Burrows, M., 78

- Candidate itemset, 227, 240
- Candidate pair, 81, 100, 231, 234
- Carey, M., 77, 78
- Categorical feature, 458, 505, 511
- Centroid, 255, 258, 264, 267, 271
- Chabbert, M., 353
- Chandra, T., 78
- Chang, F., 78
- Characteristic matrix, 89
- Charikar, M.S., 141
- Chaudhuri, S., 141
- Checkpoint, 52
- Chen, M.-S., 250
- Child, 363
- Cholera, 3
- Chowdhury, M., 79
- Chronicle data model, 173
- Chunk, 24, 240, 270
- CineMatch, 349
- Class, 458
- Classifier, 330, 457
- Click stream, 145
- Click-through rate, 297, 303
- Clique, 369
- Cloud computing, 15
- Cluster computing, 21, 22
- Cluster tree, 278, 279
- Clusters, 77
- Clustering, 3, 16, 253, 337, 355, 361, 457
- Clustroid, 258, 264
- Collaboration network, 358
- Collaborative filtering, 4, 17, 84, 293, 319, 333, 359
- Colossus, 24
- Column-orthonormal matrix, 437
- Combiner, 27, 189, 191
- Communication cost, 22, 54, 397
- Community, 17, 355, 366, 369, 394
- Commutativity, 27, 45
- Competitive ratio, 16, 298, 301, 306
- Complete graph, 369, 370
- Compressed set, 270
- Compute node, 21, 22
- Computer game, 327
- Computing cloud, *see* Cloud computing
- Concept, 438
- Concept space, 443
- Confidence, 218, 219
- Content-based recommendation, 319, 324
- Convergence, 469
- Convexity, 511
- Cooper, B.F., 78

- Coordinates, 254
- Cortes, C., 515
- Cosine distance, 107, 117, 325, 330, 444
- Counting ones, 162, 283
- Covering an output, 66
- Craig's List, 294
- Craswell, N., 317
- Credit, 364
- Cristianini, N., 515
- Cross-Validation, 463
- Crowdsourcing, 464
- CUR-decomposition, 423, 446
- CURE Algorithm, 274, 278
- Currey, J., 79
- Curse of dimensionality, 256, 280, 497, 512
- Cut, 374
- Cyclic permutation, 99
- Cylinder, 12
- Czajkowski, G., 79

- DAG, *see* Directed acyclic graph
- Darts, 152
- Das Sarma, A., 78
- Das, T., 79
- Dasgupta, A., 421
- Data mining, 1
- Data stream, 16, 244, 282, 296, 478
- Data-stream-management system, 144
- Database, 16
- Datar, M., 142, 174, 292
- Datar-Gionis-Indyk-Motwani Algorithm, 163
- Dave, A., 79
- De Raedt, L., 515
- Dead end, 179, 182, 183, 205
- Dean, J., 78
- Decaying window, 169, 246
- Decision forest, 509, 512
- Decision tree, 17, 330, 458, 461, 462, 499, 512
- Deerwester, S., 454
- Degree, 371, 394
- Degree matrix, 375
- Dehnert, J.C., 79
- del.icio.us, 326, 359
- Deletion, 108
- Dense matrix, 31, 446
- Density, 263, 265
- Depth-first search, 411
- Determinant, 425
- DeWitt, D.J., 78
- DFS, *see* Distributed file system
- Diagonal matrix, 437
- Diameter, 263, 265, 401
- Diapers and beer, 216
- Difference, 34, 37, 41
- Dimension table, 60
- Dimensionality reduction, 17, 340, 423, 497
- Directed acyclic graph, 363
- Directed graph, 400
- Discard set, 270
- Disk, 11, 221, 255, 278
- Disk block, *see* Block
- Display ad, 294, 295
- Distance measure, 105, 253, 361
- Distinct elements, 154, 157
- Distributed file system, 21, 24, 214, 221
- DMOZ, *see* Open directory
- Document, 82, 86, 217, 254, 313, 325, 326, 460
- Document frequency, *see* Inverse document frequency
- Domain, 202
- Dot product, 107
- Drineas, P., 454
- Dryad, 77
- DryadLINQ, 77
- Dual construction, 360
- Dubitzky, W., 455
- Dumais, S.T., 454
- Dup-elim task, 49

- e , 12
- Edit distance, 108, 110
- Eigenpair, 425
- Eigenvalue, 179, 376, 424, 434, 435

- Eigenvector, 179, 376, 424, 429, 434
- Email, 358
- Energy, 442
- Ensemble, 331
- Ensemble methods, 510
- Entity resolution, 122, 123
- Entropy impurity, 502
- Equijoin, 34
- Erlingsson, I., 79
- Ernst, M., 78
- Ethernet, 21, 22
- Euclidean distance, 105, 119, 496
- Euclidean space, 105, 109, 254, 255, 258, 274
- Ewen, S., 78
- Exponentially decaying window, *see* Decaying window
- Extrapolation, 495

- Facebook, 17, 198, 356
- Fact table, 60
- Failure, 23, 30, 43, 51
- Faloutsos, C., 421, 455
- False negative, 100, 111, 239
- False positive, 100, 111, 152, 239
- Family of functions, 112
- Fang, M., 250
- Fayyad, U.M., 292
- Feature, 278, 324–326
- Feature selection, 464
- Feature vector, 458, 511
- Fetterly, D., 79
- Fikes, A., 78
- File, 23, 24, 221, 239
- Filter, 44
- Filtering, 151
- Fingerprint, 125
- First-price auction, 305
- Fixedpoint, 114, 204
- Flajolet, P., 174
- Flajolet-Martin Algorithm, 155, 413
- Flatmap, 44
- Flink, 77, 78
- Flow graph, 42
- Fortunato, S., 420
- Fotakis, D., 420
- Franklin, M.J., 79
- French, J.C., 292
- Frequent bucket, 231, 233
- Frequent itemset, 4, 214, 224, 226, 370, 457
- Frequent pairs, 225
- Frequent-items table, 226
- Freund, Y., 515
- Freytag, J.-C., 78
- Friends, 356
- Friends relation, 59
- Frieze, A.M., 141
- Frobenius norm, 427, 441
- Furnas, G.W., 454

- Gaber, M.M., 18
- Ganti, V., 141, 292
- Garcia-Molina, H., 18, 212, 250, 292, 421
- Garofalakis, M., 174
- Gaussian elimination, 180
- Gehrke, J., 174, 292
- Generalization, 463
- Generated subgraph, 369
- Genre, 324, 336, 350
- GFS, *see* Google file system
- Ghemawat, S., 78
- Gibbons, P.B., 174, 421
- GINI impurity, 502
- Gionis, A., 142, 174
- Giraph, 77, 78
- Girvan, M., 421
- Girvan-Newman Algorithm, 363
- Global minimum, 342
- GN Algorithm, *see* Girvan-Newman Algorithm
- Gobioff, H., 78
- Golub, G.H., 454
- Gonzalez, J., 79
- Google, 176, 187, 302
- Google file system, 24
- Google+, 356
- Gradient descent, 17, 48, 348, 386, 486
- Granzow, M., 455

- Graph, 51, 64, 355, 356, 393, 400
- GraphLab, 77
- GraphX, 79
- Greedy algorithm, 296, 297, 300, 304
- GRGPF Algorithm, 278
- GroupByKey, 45
- Grouping, 26, 34, 38
- Grouping attribute, 34
- Groupon, 359
- Grover, R., 78
- Gruber, R.E., 78
- Guestrin, C., 79
- Guha, S., 292
- Gunda, P.K., 79
- Gyongyi, Z., 212

- Hadoop, 25, 79
- Hadoop distributed file system, 24
- HaLoop, 51, 77
- Hamming distance, 74, 109, 117
- Harris, M., 350
- Harshman, R., 454
- Hash function, 9, 87, 92, 100, 149, 152, 155
- Hash join, 37
- Hash key, 9, 312
- Hash table, 9, 10, 12, 223, 230, 233, 234, 312, 314, 394
- Haveliwala, T.H., 212
- HDFS, *see* Hadoop distributed file system
- Head, 407
- Heavy hitter, 394
- Heise, A., 78
- Hellerstein, J.M., 79
- Henzinger, M., 142
- Hierarchical clustering, 255, 257, 275, 338, 361
- Hinge loss, 485
- HITS, 204
- Hive, 77, 79
- Hoger, M., 78
- Hopcroft, J.E., 411
- Horn, H., 79
- Howe, B., 78

- Hsieh, W.C., 78
- Hub, 204
- Hueske, F., 78
- Hyperlink-induced topic search, *see* HITS
- Hyperplane, 480
- Hyracks, 77

- Identical documents, 130
- Identity matrix, 425
- IDF, *see* Inverse document frequency
- Image, 145, 325, 326
- IMDB, *see* Internet Movie Database
- Imielinski, T., 250
- Immediate subset, 242
- Immorlica, N., 142
- Important page, 176
- Impression, 294
- Impurity, 501
- In-component, 181
- Inaccessible page, 199
- Independent rows or columns, 437
- Index, 10, 394
- Indyk, P., 141, 142, 174
- Initialize clusters, 267
- Input, 64, 458
- Insertion, 108
- Instance-based learning, 461
- Interest, 218
- Internet Movie Database, 324, 350
- Interpolation, 495
- Intersection, 34, 36, 40, 85
- Into Thin Air*, 323
- Inverse document frequency, *see* TF.IDF, 8
- Inverted index, 176, 294
- Ioannidis, Y.E., 421
- IP packet, 145
- Isard, M., 79
- Isolated component, 182
- Item, 214, 216, 217, 320, 336, 337
- Item profile, 324, 327
- Itemset, 214, 222, 224

- Jaccard distance, 104, 106, 112, 325, 498

- Jaccard similarity, 82, 91, 104, 199
 Jacobsen, H.-A., 78
 Jagadish, H.V., 174
 Jahrer, M., 353
 Jeh, G., 421
 Joachims, T., 515
 Join, *see* Natural join, 45, *see* Multi-way join, *see* Star join, 396
 Join task, 49

 K-means, 266
 K-partite graph, 359
 Kahan, W., 454
 Kalyanasundaram, B., 318
 Kamm, D., 350
 Kang, U., 421
 Kannan, R., 454
 Kao, O., 78
 Karlin, A., 298
 Kaushik, R., 141
 Kautz, W.H., 174
 Kernel function, 492, 496
 Key component, 149
 Key-value pair, 25, 27
 Keyword, 303, 331
 Kleinberg, J.M., 212
 Knuth, D.E., 18
 Koren, Y., 353
 Krioukov, A., 78
 Kumar, R., 18, 79, 212, 421
 Kumar, V., 18
 Kyrola, A., 79

 Label, 356, 458
 Lagrangean multipliers, 58
 Landauer, T.K., 454
 Lang, K.J., 421
 Laplacian matrix, 376
 Lazy evaluation, 46
 LCS, *see* Longest common subsequence
 Leaf, 364
 Learning-rate parameter, 466
 Leich, M., 78
 Leiser, N., 79
 Length, 158, 400

 Length indexing, 131
 Leser, U., 78
 Leskovec, J., 420–422
 Leung, S.-T., 78
 Li, P., 142
 Likelihood, 381
 Lin, S., 142
 Linden, G., 353
 Lineage, 47
 Linear equations, 180
 Linear separability, 465, 469
 Linear transitive closure, 405, 409
 Link, 33, 176, 190
 Link matrix of the Web, 205
 Link spam, 195, 199
 Littlestone, N., 515
 Livny, M., 292
 Local minimum, 342
 Locality, 356
 Locality-sensitive family, 116
 Locality-sensitive function, 112
 Locality-sensitive hashing, 81, 100, 111, 326, 498
 Log likelihood, 387
 Logarithm, 12
 Long tail, 216, 320, 321
 Longest common subsequence, 108
 Low, Y., 79
 Lower bound, 68
 Lower hyperplane, 481
 LSH, *see* Locality-sensitive hashing

 Ma, J., 79
 Machine learning, 2, 17, 330, 457
 Maggioni, M., 454
 Maghoul, F., 18, 212
 Mahalanobis distance, 273
 Mahoney, M.W., 421, 454
 Main memory, 221, 222, 230, 255
 Malewicz, G., 79
 Malik, J., 421
 Manber, U., 142
 Manhattan distance, 106
 Manning, C.P., 18
 Many-many matching, 126

- Many-many relationship, 64, 214
- Many-one matching, 126
- Map, 44
- Map task, 25, 27
- Map worker, 28, 30
- Mapping schema, 65
- MapReduce, 15, 21, 25, 30, 189, 191, 241, 287, 396, 403, 477
- Margin, 479
- Market basket, 4, 16, 213, 214, 221
- Markl, V., 78
- Markov process, 179, 182, 390
- Martin, G.N., 174
- Master controller, 25, 27, 28
- Master node, 24
- Matching, 299
- Matias, Y., 174
- Matrix, 31, *see* Transition matrix of the Web, *see* Stochastic matrix, *see* Substochastic matrix, 189, 204, *see* Utility matrix, 340, *see* Adjacency matrix, *see* Degree matrix, *see* Laplacian matrix, *see* Symmetric matrix
- Matrix multiplication, 38, 39, 48, 69
- Matrix of distances, 435
- Matthew effect, 14
- Maximal itemset, 224
- Maximal matching, 299
- Maximum-likelihood estimation, 381
- McAuley, J., 422
- McCauley, M., 79
- Mean, *see* Average
- Mechanical Turk, 464
- Median, 156
- Mehta, A., 318
- Melnik, S., 421
- Merging clusters, 258, 261, 272, 276, 281, 285
- Merton, P., 18
- Miller, G.L., 421
- Minhashing, 82, 90, 103, 106, 113, 326
- Minibatch gradient descent, 490
- Miniclust, 270
- Minsky, M., 516
- Minutiae, 125
- Mirrokn, V.S., 142
- Mirror page, 83
- Mitzenmacher, M., 141
- ML, *see* Machine learning
- MLE, *see* Maximum-likelihood estimation
- Model, 381
- Moments, 157
- Monotonicity, 224
- Montavon, G., 515
- Moore-Penrose pseudoinverse, 447
- Most-common elements, 169
- Motwani, R., 142, 174, 250, 292
- Mueller, K.-R., 515
- Multiclass classification, 458, 473
- Multidimensional index, 497
- Multihash Algorithm, 234
- Multiplication, 31, *see* Matrix multiplication, 189, 204
- Multiset, *see* Bag
- Multistage Algorithm, 232
- Multiway join, 56, 397
- Mumick, I.S., 174
- Mutation, 111
- Name node, *see* Master node
- Natural join, 34, 37, 38, 55
- Naughton, J.F., 78
- Naumann, F., 78
- Navathe, S.B., 250
- Near-neighbor search, *see* Locality-sensitive hashing
- Nearest neighbor, 17, 458, 462, 491, 512
- Negative border, 242
- Negative example, 466
- Neighbor, 390
- Neighborhood, 400, 413
- Neighborhood profile, 400
- Netflix challenge, 2, 322, 349, 510
- Network, *see* Social network
- Neural net, 461
- Newman, M.E.J., 421
- Newspaper articles, 128, 313, 322

- Node, 502
- Node pruning, 508
- Non-Euclidean distance, 264, *see* Co-sine distance, *see* Edit distance, *see* Hamming distance, *see* Jac-card distance
- Non-Euclidean space, 278, 280
- Norm, 105, 106
- Normal distribution, 269
- Normalization, 333, 335, 346
- Normalized cut, 375
- NP-complete problem, 369
- Numerical feature, 458, 503, 511

- O’Callaghan, L., 292
- Off-line algorithm, 296
- Olston, C., 79
- Omicinski, E., 250
- On-line advertising, *see* Advertising
- On-line algorithm, 16, 296
- On-line learning, 463
- On-line retailer, 216, 294, 320, 321
- Onose, N., 78
- Open directory, 196, 464
- OR-construction, 114
- Orr, G.B., 515
- Orthogonal vectors, 256, 428
- Orthonormal matrix, 437, 442
- Orthonormal vectors, 429, 432
- Out-component, 181
- Out-degree, 410
- Outlier, 255
- Output, 64, 458
- Overfitting, 331, 348, 461, 462, 475, 500, 508, 512
- Overlapping Communities, 381
- Overture, 303
- Owen, A.B., 142
- Own pages, 200

- Paepcke, A., 142
- Page, L., 175, 212
- PageRank, 3, 16, 31, 32, 49, 175, 177, 189
- Pairs, *see* Frequent pairs

- Palmer, C.R., 421
- Pan, J.-Y., 421
- Papert, S., 516
- Parallelism, 507
- Parent, 363
- Park, J.S., 250
- Partition, 373
- Pass, 222, 225, 233, 238
- Path, 400
- Paulson, E., 78
- PCA, *see* Principal-component analysis
- PCY Algorithm, 230, 233, 234
- Pedersen, J., 212
- Perceptron, 17, 457, 461, 465, 511
- Perfect matching, 299
- Permutation, 90, 99
- Peters, M., 78
- PIG, 77
- Pigeonhole principle, 369
- Piotte, M., 353
- Pivotal condensation, 425
- Plagiarism, 83, 217
- Pnuts, 77
- Point, 253, 283
- Point assignment, 255, 266, 362
- Polyzotis, A., 77
- Position indexing, 133, 135
- Positive example, 466
- Positive integer, 168
- Powell, A.L., 292
- Power iteration, 425, 426
- Power law, 13
- Predicate, 330
- Prefix indexing, 132, 133, 135
- Pregel, 51, 77
- Principal eigenvector, 179, 425
- Principal-component analysis, 423, 430
- Priority queue, 261
- Priors, 383
- Privacy, 296
- Probe string, 133
- Profile, *see* Item profile, *see* User profile
- Projection, 34, 36

- Pruhs, K.R., 318
 Pseudoinverse, *see* Moore-Penrose pseudoinverse
 Puz, N., 78

 Quadratic programming, 486
 Query, 146, 165, 287
 Query example, 492
 Quinlan, J.R., 516

 R-tree, 292
 Rack, 22
 Radius, 263, 265, 400
 Raghavan, P., 18, 212, 421
 Rahm, E., 421
 Rajagopalan, S., 18, 212, 421
 Ramakrishnan, R., 78, 292
 Ramsey, W., 317
 Random hyperplanes, 117, 326
 Random surfer, 176, 177, 182, 196, 389
 Randomization, 238
 Rank, 436
 Rarest-first order, 313
 Rastogi, R., 174, 292
 Rating, 320, 323
 RDD, *see* Resilient distributed dataset
 Reachability, 402, 403, 409
 Recommendation system, 16, 319
 Recursion, 49
 Recursive doubling, 406, 409
 Reduce task, 25, 27
 Reduce worker, 28, 30
 Reducer, 27, 45
 Reducer size, 61, 67
 Reed, B., 79
 Regression, 458, 496, 512
 Regularization parameter, 484
 Reichsteiner, A., 455
 Reina, C., 292
 Relation, 33
 Relational algebra, 33
 Replication, 24
 Replication rate, 61, 68
 Representation, 278
 Representative point, 275
 Representative sample, 149
 Reservoir sampling, 174
 Resilient distributed dataset, 43
 Restart, 390
 Retained set, 270
 Revenue, 304
 Rheinlander, A., 78
 Ripple-carry adder, 168
 RMSE, *see* Root-mean-square error
 Robinson, E., 78
 Rocha, L.M., 455
 Root-mean-square error, 322, 341, 441
 Rosa, M., 420
 Rosenblatt, F., 516
 Rounding data, 335
 Row, *see* Tuple
 Row-orthonormal matrix, 442
 Rowsum, 278
 Royalty, J., 78

 S-curve, 102, 111
 Saberi, A., 318
 Salihoglu, S., 78
 Sample, 238, 242, 245, 247, 267, 275, 279
 Sampling, 148, 162
 Savasere, A., 250
 Sax, M.J., 78
 SCC, *see* Strongly connected component, *see* Strongly connected component
 Schapire, R.E., 515
 Schelter, S., 78
 Schema, 33
 Schutze, H., 18
 Score, 123
 Search ad, 294
 Search engine, 187, 203
 Search query, 145, 176, 198, 294, 312
 Second-price auction, 305
 Secondary storage, *see* Disk
 Selection, 33, 35
 Seminaive evaluation, 404, 405, 407, 409
 Sensor, 145

- Sentiment analysis, 465
- Set, 89, 131, *see* Itemset
- Set difference, *see* Difference
- Shankar, S., 78
- Shawe-Taylor, J., 515
- Shenker, S., 79
- Shi, J., 421
- Shim, K., 292
- Shingle, 86, 103, 128
- Shivakumar, N., 250
- Shopping cart, 216
- Shortest paths, 52
- Siddharth, J., 142
- Signature, 82, 89, 91, 103
- Signature matrix, 92, 100
- Silberschatz, A., 174
- Silberstein, A., 78
- Similarity, 4, 15, 82, 213, 326, 334
- Similarity join, 62
- Simrank, 389
- Singleton, R.C., 174
- Singular value, 437, 441, 442
- Singular-value decomposition, 340, 423, 436, 446
- Six degrees of separation, 402
- Sketch, 119
- Skew, 28
- Sliding window, 146, 162, 169, 283
- Smart transitive closure, 407, 409
- Smith, B., 353
- SNAP, 420
- Social Graph, 356
- Social network, 17, 355, 356, 423
- SON Algorithm, 240
- Source, 400
- Space, 105, 253
- Spam, *see* Term spam, *see* Link spam, 358, 464
- Spam farm, 199, 202
- Spam mass, 202, 203
- Spark, 41, 43, 51, 77, 79
- Sparse matrix, 31, 90, 91, 189, 190, 320
- Spectral partitioning, 373
- Spider trap, 182, 185, 205
- Split, 46
- Splitting clusters, 281
- SQL, 22, 33, 77, 79
- Squares, 399
- Srikant, R., 250
- Srivastava, U., 78, 79
- Standard deviation, 271, 273
- Standing query, 146
- Stanford Network Analysis Platform, *see* SNAP
- Star join, 60
- Stata, R., 18, 212
- Statistical model, 1
- Status, 313
- Steinbach, M., 18
- Stochastic gradient descent, 348, 489
- Stochastic matrix, 179, 425
- Stoica, I., 79
- Stop clustering, 259, 263, 265
- Stop words, 8, 88, 128, 217, 325
- Stratosphere, 77
- Stream, *see* Data stream
- Strength of membership, 386
- String, 131
- Striping, 32, 189, 191
- Strong edge, 358
- Strongly connected component, 181, 411
- Strongly connected graph, 179, 401
- Substochastic matrix, 182
- Suffix length, 135
- Summarization, 3
- Summation, 168
- Sun, J., 455
- Supercomputer, 21
- Superimposed code, *see* Bloom filter, 173
- Supermarket, 216, 238
- Superstep, 52
- Supervised learning, 457, 459
- Support, 214, 239, 240, 242, 244
- Support vector, 480
- Support-vector machine, 17, 457, 462, 479, 511
- Supporting page, 200
- Suri, S., 421
- Surprise number, 158

- SVD, *see* Singular-value decomposition
 SVM, *see* Support-vector machine
 Swami, A., 250
 Symmetric matrix, 377, 424
 Szegedy, M., 174
- Tag, 326, 359
 Tail, 407
 Tail length, 155, 413
 Tan, P.-N., 18
 Target, 400
 Target page, 200
 Tarjan, R.E., 411
 Task, 23
 Taxation, 182, 185, 200, 205
 Taylor expansion, 12
 Taylor, M., 317
 Telephone call, 358
 Teleport set, 196, 197, 202, 390
 Teleportation, 186
 Tendril, 181
 Tensor, 48
 TensorFlow, 41, 79
 Term, 176
 Term frequency, *see* TF.IDF, 8
 Term spam, 176, 199
 Test set, 462, 469
 TF, *see* Term frequency
 TF.IDF, 7, 8, 325, 461
 Theobald, M., 142
 Thrashing, 191, 230
 Threshold, 102, 171, 214, 240, 244, 465, 471
 TIA, *see* Total Information Awareness
 Timestamp, 163, 284
 Toivonen's Algorithm, 242
 Toivonen, H., 250
 Tomkins, A., 18, 79, 212, 421
 Tong, H., 421
 Topic-sensitive PageRank, 195, 202
 Toscher, A., 353
 Total Information Awareness, 5
Touching the Void, 323
 Training example, 458
 Training rate, 469
- Training set, 457, 458, 464, 474
 Transaction, *see* Basket
 Transformation, 44
 Transition matrix, 391
 Transition matrix of the Web, 178, 189, 190, 192, 423
 Transitive closure, 49, 402
 Transitive reduction, 412
 Transpose, 205
 Transposition, 111
 Tree, 260, 278, 279, *see* Decision tree
 Triangle, 393
 Triangle inequality, 105
 Triangular matrix, 223, 232
 Tripartite graph, 359
 Triples method, 223, 232
 TrustRank, 202
 Trustworthy page, 202
 Tsourakakis, C.E., 421
 Tube, 182
 Tuple, 33
 Tuzhilin, A., 352
 Twitter, 17, 313, 356
 Tzoumas, K., 78
- Ullman, J.D., 18, 77–79, 250, 292, 420
 Undirected graph, *see* Graph
 Union, 34, 36, 40, 85
 Unit vector, 424, 429
 Universal set, 131
 Unsupervised learning, 457
 Upper hyperplane, 481
 User, 320, 336, 337
 User profile, 328
 Utility matrix, 320, 323, 340, 423
 UV-decomposition, 340, 350, 423, 490
- VA file, 497
 Valduriez, P., 421
 Validation set, 462
 Van Loan, C.F., 454
 Vapnik, V.N., 515
 Variable, 158
 Vassilvitskii, S., 421
 Vazirani, U., 318

- Vazirani, V., 318
Vector, 31, 105, 109, 179, 189, 204, 205, 254
Vernica, R., 78
Vigna, S., 420
Vitter, J., 174
Volume (of a set of nodes), 375
von Ahn, L., 327, 353
von Luxburg, U., 421
Voronoi diagram, 492
- Wall, M.E., 455
Wall-clock time, 55
Wallach, D.A., 78
Wang, J., 350
Wang, W., 142
Warneke, D., 78
Weak edge, 358
Weaver, D., 78
Web structure, 181
Weight, 465
Weiner, J., 18, 212
Whizbang Labs, 2
Widom, J., 18, 79, 174, 292, 421
Wikipedia, 358, 464
Window, *see* Sliding window, *see* Decaying window
Windows, 12
Winnow Algorithm, 469
Word, 217, 254, 325
Word count, 25, 44
Worker process, 28
Workflow, 42, 49, 54
Working store, 144
- Xiao, C., 142
Xie, Y., 455
- Yahoo, 303, 326
Yang, J., 421, 422
Yerneni, R., 78
York, J., 353
Yu, J.X., 142
Yu, P.S., 250
Yu, Y., 79
- Zaharia, M., 79
Zhang, C.H., 142
Zhang, H., 455
Zhang, T., 292
Zipf's law, 15, *see* Power law
Zoeter, O., 317