

Index

- A-Priori Algorithm, 216, 217, 223
- Accessible page, 193
- Accumulated sum, 489
- Accuracy impurity, 484
- Action, 42
- Activation function, 505, 506
- Activation map, 523
- Active learning, 447
- Ad-hoc query, 141
- Adjacency matrix, 359
- Adomavicius, G., 339
- Advertising, 17, 124, 209, 282
- Adwords, 291
- Affiliation-graph model, 367
- Afrati, F.N., 75, 408
- Agglomerative clustering, *see* Hierarchical clustering
- Aggregation, 32, 36
- Agrawal, R., 241
- Alexandrov, A., 75
- Algorithm, 1
- All-pairs problem, 63, 66
- Alon, N., 167
- Alon–Matias–Szegedy Algorithm, 152
- Amplification, 109
- Analytic query, 58
- AND-construction, 109
- Anderson, C., 339
- Andoni, A., 136
- ANF, *see* Approximate neighborhood function
- ANF Algorithm, 402
- Apache, 23, 76
- Approximate neighborhood function, 402
- Arbib, M., 543
- Arc, 388
- Archive, 139
- Ask, 198
- Association rule, 210, 211
- Associativity, 26, 44
- Attribute, 31
- Auction, 294
- Augmentation of data, 541
- Austern, M.H., 76
- Authority, 197, 198
- Average, 150
- B-tree, 280
- Babcock, B., 167, 280
- Babu, S., 167
- Backpropagation, 514–516
- Backpropagation through time, 534
- Backstrom, L., 408
- Bad point, 466
- Bag, 39, 82
- Balance Algorithm, 293
- Balazinska, M., 75
- Band, 97
- Bandwidth, 21
- Basket, *see* Market basket, 206, 208, 209, 237
- Batch gradient descent, 472
- Batch learning, 447
- Bayes net, 5
- BDMO Algorithm, 272
- Beer and diapers, 211
- Bell, R., 339
- Bellkor’s Pragmatic Chaos, 310
- Bengio, Y., 543
- Berges, C.J.C., 543
- Bergmann, R., 75
- Berkhin, P., 205
- Berrar, D.P., 440
- Betweenness, 348
- BFR Algorithm, 256, 259
- BFS, *see* Breadth-first search
- Bi-clique, 354
- Bias vector, 504
- Bickson, D., 76
- Bid, 292, 294, 300, 301
- Big data, 1
- BigTable, 75
- Bik, A.J.C., 76
- Binary classification, 442
- Biomarker, 210
- Bipartite graph, 288, 344, 354, 355
- BIRCH Algorithm, 280
- Birrell, A., 76
- Bitmap, 224, 225

- Block, 12, 20, 185, 186
- Blockeel, H., 497
- Blocking property, 41, 47
- Blog, 193
- Bloom filter, 146, 222
- Bloom, B.H., 168
- Blum, A., 497
- Bohannon, P., 76
- Boldi, P., 408
- Bonferroni correction, 6
- Bonferroni's principle, 5, 6
- Bookmark, 191
- Boral, H., 409
- Borkar, V., 75
- Boser, B., 543
- Bottou, L., 497
- BPTT, *see* Backpropagation through time
- Bradley, P.S., 280
- Breadth-first search, 348
- Breiman, L., 19
- Brick-and-mortar retailer, 209, 307, 309
- Brin, S., 205
- Broad matching, 294
- Broder, A.Z., 19, 136, 205
- Bu, Y., 75
- Bucket, 10, 144, 158, 162, 222, 272
- Budget, 292, 299
- Budiu, M., 76
- Bulk-synchronous system, 50
- Burges, C.J.C., 497
- Burrows, M., 76
- Buxketing, 533
- Caffe, 505, 543
- Candidate itemset, 219, 232
- Candidate pair, 78, 97, 223, 225
- Carey, M., 75
- Categorical feature, 442, 487, 493
- Cell state, 536
- Centroid, 245, 248, 253, 256, 260
- Chabbert, M., 339
- Chain rule, 516
- Chandra, T., 76
- Chang, F., 76
- Channel, 522
- Characteristic matrix, 86
- Charikar, M., 136
- Charikar, M.S., 136
- Chaudhuri, S., 136
- Checkpoint, 51
- Chen, M.-S., 241
- Child, 348
- Cholera, 4
- Chowdhury, M., 76
- Chronicle data model, 167
- Chunk, 23, 232, 259
- CineMatch, 336
- Class, 442
- Classification loss, 511
- Classifier, 318, 441
- Click stream, 140
- Click-through rate, 286, 292
- Clique, 353
- Cloud computing, 16
- Cluster computing, 20, 21
- Cluster tree, 267, 268
- Clustera, 74
- Clustering, 4, 17, 243, 325, 340, 345, 441
- Clustroid, 247, 254
- CNN, *see* Convolutional neural network
- Collaboration network, 343
- Collaborative filtering, 5, 17, 81, 282, 307, 321, 343
- Colossus, 23
- Column-orthonormal matrix, 423
- Combiner, 26, 183, 185
- Communication cost, 21, 52, 386
- Community, 17, 340, 351, 353, 383
- Commutativity, 26, 44
- Competitive ratio, 17, 287, 290, 294
- Complete graph, 354, 355
- Compressed set, 260
- Compute graph, 514
- Compute node, 20, 21
- Computer game, 314
- Concept, 425
- Concept space, 430
- Conductance, 380
- Confidence, 210, 211
- Content-based recommendation, 307, 311
- Convergence, 452
- Convexity, 493
- Convolutional layer, 523, 525
- Convolutional neural network, 18, 498, 502, 522
- Cooper, B.F., 76
- Coordinates, 243
- Cortes, C., 497, 543
- Cosine distance, 103, 113, 313, 317, 430
- Counting ones, 157, 272
- Covering an output, 64
- Craig's List, 282
- Craswell, N., 305
- Credit, 349
- Cristianini, N., 497
- Cross entropy, 508, 512
- Cross-Validation, 447
- Crowdsourcing, 448
- CUR-decomposition, 410, 432
- CURE Algorithm, 264, 267
- Currey, J., 76
- Curse of dimensionality, 245, 269, 479, 494
- Cut, 358, 359

- Cyclic permutation, 96
- Cylinder, 13
- Czajkowski, G., 76
- DAG, *see* Directed acyclic graph
- Darts, 147
- Das Sarma, A., 75
- Das, T., 76
- Dasgupta, A., 409
- Data mining, 1
- Data science, 1
- Data stream, 16, 235, 271, 285, 460
- Data-stream-management system, 138
- Database, 16
- Datar, M., 136, 167, 168, 280
- Datar–Gionis–Indyk–Motwani Algorithm, 158
- Dave, A., 76
- De Raedt, L., 497
- Dead end, 173, 176, 177, 199
- Dean, J., 76
- Decaying window, 163, 237
- Decision forest, 491, 494
- Decision tree, 18, 318, 441, 445, 446, 482, 494
- Deep learning, 3, 18, 498
- Deerwester, S., 440
- Degree, 356, 383
- Degree matrix, 360
- Dehnert, J.C., 76
- del.icio.us, 314, 344
- Deletion, 104
- Denker, J.S., 543
- Dense matrix, 30, 432
- Density, 253, 255
- Density of edges, 380
- Depth-first search, 400
- Determinant, 412
- Development set, 446
- DeWitt, D.J., 76
- DFS, *see* Distributed file system
- Diagonal matrix, 423
- Diameter, 252, 255, 390
- Diapers and beer, 209
- Difference, 32, 35, 39
- Dimension table, 58
- Dimensionality reduction, 18, 327, 410, 480
- Directed acyclic graph, 348
- Directed graph, 388
- Discard set, 259
- Disk, 12, 214, 245, 267
- Disk block, *see* Block
- Display ad, 282, 284
- Distance measure, 101, 243, 346, 512
- Distinct elements, 148, 152
- Distributed file system, 20, 22, 207, 213
- DMOZ, *see* Open directory
- Document, 79, 83, 209, 244, 301, 312, 314, 444
- Document frequency, *see* Inverse document frequency
- Domain, 196
- Dot product, 103, 518
- Drineas, P., 440
- Dropout, 540
- Dryad, 74
- DryadLINQ, 74
- Dual construction, 345
- Dubitzky, W., 440
- Dumais, S.T., 440
- Dup-elim task, 48
- Dying ReLU, 509
- e*, 13
- Edit distance, 104, 106
- Eigenpair, 411
- Eigenvalue, 173, 361, 410, 421
- Eigenvector, 173, 361, 410, 415, 420
- ELU, 510
- Email, 343
- Energy, 428
- Ensemble, 318
- Ensemble methods, 492
- Entity resolution, 118
- Entropy, 484, 512
- Equijoin, 32
- Erlingsson, I., 76
- Ernst, M., 75
- Ethernet, 20, 21
- Euclidean distance, 102, 115, 479
- Euclidean space, 102, 105, 243, 245, 248, 264
- Ewen, S., 75
- Exclusive-or, 503
- Explainability, 3
- Exploding derivative, 506
- Exploding gradient, 535
- Exponential linear unit, *see* ELU
- Exponentially decaying window, *see* Decaying window
- Extrapolation, 477
- Facebook, 17, 191, 341
- Fact table, 58
- Failure, 22, 27, 41, 49
- Faloutsos, C., 409, 440
- False negative, 97, 107, 231
- False positive, 97, 107, 147, 231
- Family of functions, 108
- Fang, M., 241
- Fayyad, U.M., 280
- Feature, 267, 312, 313
- Feature selection, 447
- Feature vector, 442, 493
- Feedforward network, 505
- Fetterly, D., 76

Fikes, A., 76	Google, 170, 181, 291
File, 22, 213, 230	Google file system, 23
Filter, 43, 523	Google+, 341
Filtering, 145	GPU, <i>see</i> Graphics processing unit, 530
Fingerprint, 121	Gradient descent, 18, 47, 335, 370, 468, 501, 504, 506, 514
First-price auction, 294	Granzow, M., 440
Fixedpoint, 111, 197	Graph, 50, 62, 340, 341, 383, 389
Flajolet, P., 168	Graphics processing unit, 505
Flajolet–Martin Algorithm, 149, 401	GraphLab, 75
Flatmap, 43	GraphX, 76
Flattening, 521	Greedy algorithm, 285, 286, 289, 293
Flink, 75, 76	GRGPF Algorithm, 267
Flow graph, 40	GroupByKey, 44
Forget gate, 537	Grouping, 25, 32, 36
Fortunato, S., 408	Grouping attribute, 32
Fotakis, D., 408	Groupon, 344
Franklin, M.J., 76	Grover, R., 75
French, J.C., 280	GRU, <i>see</i> Gated recurrent unit
Frequent bucket, 223, 224	Gruber, R.E., 76
Frequent itemset, 5, 207, 216, 219, 355, 441	Guestrin, C., 76
Frequent pairs, 217	Guha, S., 281
Frequent-items table, 218	Gunda, P.K., 76
Freund, Y., 497	Gyongyi, Z., 205
Freytag, J.-C., 75	Hadamard product, 518, 536
Friends, 341	Hadoop, 23, 76
Friends relation, 57	Hadoop distributed file system, 23
Frieze, A.M., 136	HaLoop, 50, 75
Frobenius norm, 414, 428	Hamming distance, 72, 105, 112
Fukushima, K., 543	Hanazawa, T., 543
Fully connected layer, 501	Harris, M., 336
Furnas, G.W., 440	Harshman, R., 440
Gaber, M.M., 19	Hash function, 10, 84, 89, 97, 144, 146, 149
Ganti, V., 136, 280	Hash join, 35
Garcia-Molina, H., 19, 205, 241, 280, 409	Hash key, 10, 300
Garofalakis, M., 168	Hash table, 10–12, 215, 222, 225, 226, 300, 302, 384
Gate, 536	Haveliwala, T.H., 205
Gated recurrent unit, 538	HDFS, <i>see</i> Hadoop distributed file system
Gaussian elimination, 174	Head, 396
Gehrke, J., 168, 280	Heavy hitter, 383
Generalization, 447	Heise, A., 75
Generated subgraph, 354	Hellerstein, J.M., 76
Genre, 312, 324, 336	Henderson, D., 543
GFS, <i>see</i> Google file system	Henzinger, M., 137
Ghemawat, S., 76	Hidden layer, 500
Gibbons, P.B., 168, 409	Hidden state, 532
GINI impurity, 484	Hierarchical clustering, 245, 247, 264, 325, 346
Gionis, A., 137, 168	Hinge function, 468
Giraph, 75, 76	Hinge loss, 468
Girvan, M., 408	Hinton, G.E., 543
Girvan–Newman Algorithm, 348	HITS, 197
Global minimum, 329	Hive, 75, 76
GN Algorithm, <i>see</i> Girvan–Newman Algorithm	Hochreiter, S., 543
Gobioff, H., 76	Hoger, M., 75
Golub, G.H., 440	
Gonzalez, J., 76	

- Hopcroft, J.E., 400
 Hopfield, J.J., 543
 Horn, H., 76
 Howard, R.E., 543
 Howe, B., 75
 Hsieh, W.C., 76
 Hub, 197, 198
 Hubbard, W., 543
 Huber loss, 511
 Hueske, F., 75
 Hyperbolic tangent, 507, 535
 Hyperlink-induced topic search, *see* HITS
 Hyperparameter, 510
 Hyperplane, 463
 Hyracks, 74
 Identical documents, 126
 Identity matrix, 412
 IDF, *see* Inverse document frequency
 Image, 140, 312, 313
 ImageNet, 522, 527, 543
 IMDB, *see* Internet Movie Database
 Imielinski, T., 241
 Immediate subset, 233
 Immorlica, N., 136
 Important page, 170
 Impression, 283
 Impurity, 484
 In-component, 174
 Inaccessible page, 193
 Independent rows or columns, 423
 Index, 11, 383
 Indyk, P., 136, 137, 168
 Information integration, 6
 Initialize clusters, 257
 Input, 62, 442
 Input gate, 537
 Input layer, 500
 Insertion, 104
 Instance-based learning, 445
 Interest, 211
 Internet Movie Database, 312, 336
 Interpolation, 477
 Intersection, 32, 35, 39, 82
Into Thin Air, 310
 Inverse document frequency, *see* TF.IDF, 9
 Inverted index, 170, 283
 Ioannidis, Y.E., 408
 IP packet, 140
 Isard, M., 76
 Isolated component, 175
 Item, 206, 208, 209, 308, 323, 325
 Item profile, 312, 314
 Itemset, 206, 214, 216
 Jaccard distance, 101, 103, 108, 313, 480
 Jaccard similarity, 78, 79, 88, 101, 192
 Jacobian, 516
 Jacobsen, H.-A., 76
 Jagadish, H.V., 168
 Jahrer, M., 339
 Jeh, G., 409
 Joachims, T., 497
 Join, *see* Natural join, 44, *see* Multiway join,
 see Star join, 385
 Join task, 48
 K-means, 256
 K-partite graph, 344
 Kahan, W., 440
 Kalyanasundaram, B., 306
 Kamm, D., 336
 Kang, U., 409
 Kannan, R., 440
 Kao, O., 75
 Karlin, A., 287
 Kaushik, R., 136
 Kautz, W.H., 168
 Kernel, 526
 Kernel function, 474, 478
 Key component, 144
 Key-value pair, 24, 26
 Keyword, 291, 318
 KL-divergence, 512
 Kleinberg, J.M., 205
 Knuth, D.E., 19
 Koren, Y., 339
 Krioukov, A., 76
 Krizhevsky, A., 543
 Kumar, R., 19, 76, 205, 409
 Kumar, V., 19
 Kyrola, A., 76
 Label, 341, 442
 Lagrangean multipliers, 56
 Landauer, T.K., 440
 Lang, K.J., 409, 543
 Laplacian matrix, 360
 Layer, 498
 Lazy evaluation, 45
 LCS, *see* Longest common subsequence
 Leaf, 349
 Leaky ReLU, 509
 Learning rate, 450, 520
 LeCun, Y., 543
 Leich, M., 75
 Leiser, N., 76
 Length, 152, 389
 Length indexing, 127
 Leser, U., 75
 Leskovec, J., 408, 409
 Leung, S.-T., 76
 Li, P., 137
 Likelihood, 366
 Lin, S., 137
 Linden, G., 339

- Lineage, 45
- Linear equations, 174
- Linear separability, 449, 452
- Linear transitive closure, 393, 398
- Link, 31, 170, 183
- Link matrix of the Web, 199
- Link spam, 189, 193
- Littlestone, N., 497
- Livny, M., 281
- Local minimum, 329
- Locality, 340, 341
- Locality-sensitive family, 112
- Locality-sensitive function, 108
- Locality-sensitive hashing, 1, 78, 97, 107, 314, 480
- Log likelihood, 371
- Logarithm, 13
- Logistic sigmoid, 506
- Long short-term memory network, 18, 498, 535
- Long tail, 209, 307, 309
- Longest common subsequence, 104
- Low, Y., 76
- Lower bound, 66
- Lower hyperplane, 464
- LSH, *see* Locality-sensitive hashing
- LSTM, *see* Long short-term memory network
- Ma, J., 76
- Machine learning, 2, 18, 318, 441
- Maggioni, M., 440
- Maghoul, F., 19, 205
- Mahalanobis distance, 262
- Mahoney, M.W., 409, 440
- Main memory, 214, 222, 245
- Malewicz, G., 76
- Malik, J., 409
- Manber, U., 137
- Manhattan distance, 102
- Manning, C.P., 19
- Many-many matching, 121
- Many-many relationship, 62
- Many-one matching, 121
- Many-many relationship, 206
- Map, 43
- Map task, 23, 24, 26
- Map worker, 26, 28
- Mapping schema, 64
- MapReduce, 20, 23, 29, 183, 184, 232, 276, 385, 392, 461
- Margin, 462
- Market basket, 5, 17, 206, 213
- Markl, V., 75
- Markov process, 173, 175, 374
- Martin, G.N., 168
- Master controller, 24–26
- Master node, 23
- Matching, 288
- Matias, Y., 167
- Matrix, 29, *see* Transition matrix of the Web, *see* Stochastic matrix, *see* Substochastic matrix, 183, 198, *see* Utility matrix, 327, *see* Adjacency matrix, *see* Degree matrix, *see* Laplacian matrix, *see* Symmetric matrix
- Matrix multiplication, 37, 38, 47, 67
- Matrix of distances, 421
- Matthew effect, 15
- Max pooling, 527
- Maximal itemset, 216
- Maximal matching, 288
- Maximum-likelihood estimation, 366
- McAuley, J., 409
- McCauley, M., 76
- Mean, *see* Average
- Mean squared error, 511
- Mechanical Turk, 448
- Median, 150
- Mehta, A., 306
- Melnik, S., 409
- Merging clusters, 248, 252, 262, 266, 270, 274
- Merton, P., 19
- Miller, G.L., 409
- Minhashing, 78, 87, 100, 103, 109, 314
- Minibatch gradient descent, 473
- Minicluster, 260
- Minsky, M., 497
- Minutiae, 121
- Mirroknai, V.S., 136
- Mirror page, 80
- Mitzenmacher, M., 136
- ML, *see* Machine learning
- MLE, *see* Maximum-likelihood estimation
- MNIST dataset, 521
- Model, 366
- Modeling, 1
- Moments, 151
- Monotonicity, 216
- Montavon, G., 497
- Moore–Penrose pseudoinverse, 433
- Most-common elements, 163
- Motwani, R., 137, 167, 168, 241, 280
- Mueller, K.-R., 497
- Multiclass classification, 442, 456
- Multidimensional index, 479
- Multihash Algorithm, 226
- Multiplication, 29, *see* Matrix multiplication, 183, 198
- Multiset, *see* Bag
- Multistage Algorithm, 224
- Multiway join, 54, 386
- Mumick, I.S., 168
- Mutation, 107

- Name node, *see* Master node
- Natural join, 32, 35, 37, 53
- Naughton, J.F., 76
- Naumann, F., 75
- Navathe, S.B., 241
- Near-neighbor search, *see* Locality-sensitive hashing
- Nearest neighbor, 18, 441, 445, 474, 494
- Negative border, 233
- Negative example, 449
- Neighbor, 374
- Neighborhood, 389, 401
- Neighborhood profile, 389
- Netflix challenge, 2, 310, 336, 492
- Network, *see* Social network
- Neural net, 18, 445, 498
- Neuron, 502
- Newman, M.E.J., 408
- Newspaper articles, 123, 301, 311
- Node, 484, 498
- Node pruning, 490
- Non-Euclidean distance, 253, *see* Cosine distance, *see* Edit distance, *see* Hamming distance, *see* Jaccard distance
- Non-Euclidean space, 267, 269
- Norm, 102
- Norm penalty, 539
- Normal distribution, 259
- Normalization, 320, 322, 333
- Normalized cut, 359
- NP-complete problem, 353
- Numerical feature, 442, 485, 493
- O’Callaghan, L., 280
- Off-line algorithm, 285
- Olston, C., 76
- Omicinski, E., 241
- On-line advertising, *see* Advertising
- On-line algorithm, 17, 285
- On-line learning, 447
- On-line retailer, 209, 283, 307, 309
- Onose, N., 75
- Open directory, 189, 448
- OR-construction, 110
- Orr, G.B., 497
- Orthogonal vectors, 246, 415
- Orthonormal matrix, 423, 428
- Orthonormal vectors, 415, 419
- Out-component, 174
- Out-degree, 399
- Outlier, 245
- Output, 62, 442
- Output gate, 537
- Output layer, 500
- Overfitting, 318, 334, 445, 446, 459, 482, 490, 494, 538
- Overlapping Communities, 365
- Overture, 291
- Owen, A.B., 137
- Own pages, 193
- Paepcke, A., 137
- Page, L., 169, 205
- PageRank, 4, 16, 29, 31, 47, 169, 171, 183
- Pairs, *see* Frequent pairs
- Palmer, C.R., 409
- Pan, J.-Y., 409
- Papert, S., 497
- Parallelism, 489
- Parametric ReLU, 510
- Parent, 348
- Park, J.S., 241
- Partition, 358
- Pass, 214, 217, 224, 229
- Path, 389
- Paulson, E., 76
- PCA, *see* Principal-component analysis
- PCY Algorithm, 222, 224, 226
- Pearson, K., 440
- Pedersen, J., 205
- Perceptron, 18, 441, 445, 449, 493, 498, 499
- Perfect matching, 288
- Permutation, 87, 96
- Peters, M., 75
- Phishing, 1
- PIG, 75
- Pigeonhole principle, 354
- Piotte, M., 339
- Pivotal condensation, 412
- Plagiarism, 80, 210
- Pnuts, 75
- Point, 243, 272
- Point assignment, 245, 256, 347
- Polyzotis, A., 75
- Pooling function, 527
- Pooling layer, 501, 527
- Position indexing, 129, 130
- Positive example, 449
- Positive integer, 162
- Powell, A.L., 280
- Power iteration, 412, 413
- Power law, 14
- Predicate, 318
- Prefix indexing, 127, 129, 130
- Pregel, 50, 75
- Principal eigenvector, 173, 411
- Principal-component analysis, 410, 417
- Priority queue, 251
- Priors, 367
- Privacy, 284
- Probe string, 128
- Profile, *see* Item profile, *see* User profile
- Projection, 32, 34
- Pruhs, K.R., 306

- Pseudoinverse, *see* Moore–Penrose pseudoinverse
- Puz, N., 76
- PyTorch, 505, 543
- Quadratic programming, 468
- Query, 140, 159, 276
- Query example, 474
- Quinlan, J.R., 497
- R-tree, 280
- Rack, 21
- Radius, 252, 255, 389
- Raghavan, P., 19, 205, 409
- Rahm, E., 409
- Rajagopalan, S., 19, 205, 409
- Ramakrishnan, R., 76, 280, 281
- Ramsey, W., 305
- Random hyperplanes, 113, 314
- Random interconnection layer, 501
- Random surfer, 170, 171, 176, 190, 373
- Randomization, 230
- Rank, 423
- Rarest-first order, 302
- Rastogi, R., 168, 281
- Rating, 308, 311
- RDD, *see* Resilient distributed dataset
- Reachability, 391, 392, 398
- Recommendation system, 17, 307
- Rectified linear unit, *see* ReLU
- Recurrent neural network, 18, 498, 531
- Recursion, 47
- Recursive doubling, 394, 398
- Reduce task, 24, 25
- Reduce worker, 26, 28
- Reducer, 26, 43
- Reducer size, 59, 65
- Reed, B., 76
- Regression, 442, 478, 494
- Regression loss, 510
- Regularization, 538
- Regularization parameter, 467
- Reichsteiner, A., 440
- Reina, C., 280
- Relation, 31
- Relational algebra, 31, 32
- ReLU, 509, 535
- Replication, 23
- Replication rate, 59, 66
- Representation, 268
- Representative point, 265
- Representative sample, 143
- Reservoir sampling, 167
- Residual PageRank, 377
- Resilient distributed dataset, 42
- Response, 523
- Restart, 374
- Retained set, 260
- Revenue, 292
- Rheinlander, A., 75
- Ripple-carry adder, 162
- RMSE, *see* Root-mean-square error
- RNN, *see* Recurrent neural network
- Robinson, E., 76
- Rocha, L.M., 440
- Root-mean-square error, 310, 328, 428, 511
- Rosa, M., 408
- Rosenblatt, F., 497
- Rounding data, 322
- Row, *see* Tuple
- Row-orthonormal matrix, 428
- Rowsum, 267
- Royalty, J., 76
- Rumelhart, D.E., 543
- S-curve, 98, 107
- Saberi, A., 306
- Salihoglu, S., 75
- Sample, 230, 233, 236, 238, 257, 264, 268
- Sampling, 142, 156
- Saturation, 506
- Savasere, A., 241
- Sax, M.J., 75
- SCC, *see* Strongly connected component, *see* Strongly connected component
- Schapire, R.E., 497
- Schelter, S., 75
- Schema, 31
- Schmidhuber, J., 543
- Schutze, H., 19
- Score, 119
- Search ad, 283
- Search engine, 181, 196
- Search query, 140, 170, 191, 283, 300
- Second-price auction, 294
- Secondary storage, *see* Disk
- Selection, 32, 34
- Semi-naive evaluation, 393
- Seminaive evaluation, 394, 396, 397
- Sensor, 139
- Sentiment analysis, 448
- Set, 86, 126, *see* Itemset
- Set difference, *see* Difference
- Shankar, S., 76
- Shawe-Taylor, J., 497
- Shenker, S., 76
- Shi, J., 409
- Shikano, K., 543
- Shim, K., 281
- Shingle, 83, 100, 124
- Shivakumar, N., 241
- Shopping cart, 208
- Shortest paths, 50
- Siddharth, J., 137
- Sigmoid, 506

- Signature, 78, 86, 88, 100
- Signature matrix, 88, 97
- Silberschatz, A., 168
- Silberstein, A., 76
- Similarity, 5, 16, 79, 206, 314, 321
- Similarity join, 60
- Simonyan, K., 528
- Simrank, 373
- Singleton, R.C., 168
- Singular value, 423, 427, 428
- Singular-value decomposition, 327, 410, 423, 432
- Six degrees of separation, 391
- Sketch, 114
- Skew, 27
- Sliding window, 141, 157, 163, 272
- Smart transitive closure, 396, 398
- Smith, B., 339
- SNAP, 408
- Social Graph, 341
- Social network, 17, 340, 341, 410
- Softmax, 508
- SON Algorithm, 232
- Source, 389
- Source node, 373
- Space, 101, 102, 243
- Spam, *see* Term spam, *see* Link spam, 343, 447
- Spam farm, 193, 195
- Spam mass, 195, 196
- Spark, 40, 42, 50, 75, 76
- Sparse matrix, 30, 87, 88, 183, 308
- Spectral partitioning, 358
- Spider trap, 176, 179, 199
- Split, 44
- Splitting clusters, 270
- SQL, 20, 31, 75, 76
- Squared error, 510
- Squares, 387
- Srikant, R., 241
- Srivastava, U., 76
- Standard deviation, 260, 262
- Standing query, 140
- Stanford Network Analysis Platform, *see* SNAP
- Star join, 58
- Stata, R., 19, 205
- Statistical model, 2
- Status, 302
- Steinbach, M., 19
- Step function, 504, 506
- Stochastic gradient descent, 335, 472
- Stochastic matrix, 173, 411
- Stoica, I., 76
- Stop clustering, 249, 253, 255
- Stop words, 9, 85, 124, 209, 313
- Stratosphere, 75
- Stream, *see* Data stream
- Strength of membership, 371
- Stride, 525
- String, 126
- Striping, 30, 183, 185
- Strong edge, 343
- Strongly connected component, 174, 400
- Strongly connected graph, 173, 390
- Substochastic matrix, 176
- Suffix length, 130
- Summarization, 4
- Summation, 162
- Sun, J., 440
- Supercomputer, 20
- Superimposed code, *see* Bloom filter, 167
- Supermarket, 208, 229
- Superstep, 50
- Supervised learning, 441, 443
- Support, 207, 231, 233, 236
- Support vector, 463
- Support-vector machine, 18, 441, 445, 462, 493
- Supporting page, 193
- Suri, S., 409
- Surprise number, 152
- Sutskever, I., 543
- SVD, *see* Singular-value decomposition
- SVM, *see* Support-vector machine
- Swami, A., 241
- Symmetric matrix, 361, 410
- Szegedy, M., 167
- Tag, 313, 344
- Tail, 396
- Tail length, 150, 401
- Tan, P.-N., 19
- Target, 389
- Target page, 193
- Tarjan, R.E., 400
- Task, 22
- Taxation, 176, 179, 193, 199
- Taylor expansion, 13
- Taylor, M., 305
- Telephone call, 343
- Teleport set, 190, 191, 196, 374, 375
- Teleportation, 180
- Tendrill, 174
- Tensor, 46, 520
- TensorFlow, 40, 76, 505, 513, 517, 543
- Term, 170
- Term frequency, *see* TF.IDF, 9
- Term spam, 170, 193
- Test set, 446, 452
- TF, *see* Term frequency
- TF.IDF, 8, 9, 313, 444
- Theobald, M., 137

-
- Thrashing, 185, 222
 - Threshold, 98, 165, 207, 231, 236, 449, 455
 - TIA, *see* Total Information Awareness
 - Timestamp, 158, 273
 - Toivonen's Algorithm, 233
 - Toivonen, H., 242
 - Tomkins, A., 19, 76, 205, 409
 - Tong, H., 409
 - Topic-sensitive PageRank, 189, 196
 - Toscher, A., 339
 - Total Information Awareness, 6
 - Touching the Void*, 310
 - Training example, 442
 - Training rate, 453
 - Training set, 441, 442, 448, 457
 - Transaction, *see* Basket
 - Transformation, 42
 - Transition matrix, 375
 - Transition matrix of the Web, 172, 183, 186, 410
 - Transitive closure, 47, 391
 - Transitive reduction, 400
 - Transpose, 199
 - Transposition, 107
 - Tree, 251, 267, 268, *see* Decision tree
 - Triangle, 382
 - Triangle inequality, 101
 - Triangular matrix, 215, 224
 - Tripartite graph, 345
 - Triples method, 215, 224
 - TrustRank, 195, 196
 - Trustworthy page, 196
 - Tsourakakis, C.E., 409
 - Tube, 175
 - Tuple, 31
 - Tuzhilin, A., 339
 - Twitter, 17, 301, 341
 - Tzoumas, K., 75
 - Ullman, J.D., 19, 75, 76, 241, 280, 408
 - Undirected graph, *see* Graph
 - Union, 32, 34, 39, 82
 - Unit vector, 411, 415
 - Universal set, 126
 - Unsupervised learning, 441
 - Upper hyperplane, 464
 - User, 308, 323, 325
 - User profile, 316
 - Utility matrix, 308, 311, 327, 410
 - UV-decomposition, 327, 336, 410, 472
 - VA file, 480
 - Valduriez, P., 409
 - Validation set, 446, 541
 - Van Loan, C.F., 440
 - Vanishing gradient, 535
 - Vapnik, V.N., 497
 - Variable, 153
 - Vassilvitskii, S., 409
 - Vazirani, U., 306
 - Vazirani, V., 306
 - Vector, 29, 102, 105, 172, 183, 198, 243
 - Vernica, R., 75
 - VGGnet, 527
 - Vigna, S., 408
 - Vitter, J., 168
 - Volume, 380
 - Volume (of a set of nodes), 359
 - von Ahn, L., 314, 339
 - von Luxburg, U., 409
 - Voronoi diagram, 475
 - Waibel, A., 543
 - Wall, M.E., 440
 - Wall-clock time, 54
 - Wallach, D.A., 76
 - Wang, J., 336
 - Wang, W., 137
 - Warneke, D., 75
 - Weak edge, 343
 - Weaver, D., 76
 - Web structure, 174
 - Weight, 449, 500
 - Weiner, J., 19, 205
 - Whizbang Labs, 2
 - Widom, J., 19, 76, 167, 280, 409
 - Wikipedia, 343, 448
 - Williams, R.J., 543
 - Window, *see* Sliding window, *see* Decaying window
 - Windows, 12
 - Winnnow algorithm, 453
 - Word, 209, 244, 313
 - Word count, 25, 43
 - Worker process, 26
 - Workflow, 40, 47, 52
 - Working store, 139
 - Xiao, C., 137
 - Xie, Y., 440
 - Yahoo, 291, 314
 - Yang, J., 409
 - Yerneni, R., 76
 - York, J., 339
 - Yu, J.X., 137
 - Yu, P.S., 241
 - Yu, Y., 76
 - Zaharia, M., 76
 - Zero padding, 525, 533
 - Zhang, C.H., 137
 - Zhang, H., 440
 - Zhang, T., 281
 - Zipf's law, 15, *see* Power law
 - Zoeter, O., 305
 - Zussman, A., 528