

Index

- A-Priori Algorithm, 224, 225, 231
- Accessible page, 199
- Accumulated sum, 513
- Accuracy impurity, 508
- Action, 44
- Activation function, 531
- Activation map, 549
- Active learning, 470
- Ad-hoc query, 146
- Adjacency matrix, 375
- Adomavicius, G., 352
- Advertising, 17, 128, 216, 293
- Adwords, 302
- Affiliation-graph model, 383
- Afrati, F.N., 77, 78, 426
- Agglomerative clustering, *see* Hierarchical clustering
- Aggregation, 34, 38
- Agrawal, R., 250
- Alexandrov, A., 78
- Algorithm, 1
- 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, 419
- Apache, 24, 25, 78, 79
- Approximate neighborhood function, 419
- Arbib, M., 569
- Arc, 405
- Archive, 144
- Ask, 204
- Association rule, 217, 219
- Associativity, 27, 45
- Attribute, 33
- Auction, 305
- Augmentation of data, 567
- Austern, M.H., 79
- Authority, 204
- Average, 156
- B-tree, 292
- Babcock, B., 174, 292
- Babu, S., 174
- Backpropagation, 539
- Backpropagation through time, 561
- Backstrom, L., 426
- Bad point, 489
- Bag, 40, 85
- Balance Algorithm, 305
- Balazinska, M., 78
- Band, 100
- Bandwidth, 22
- Basket, *see* Market basket, 214, 216, 217, 246
- Batch gradient descent, 495
- Batch learning, 469
- Bayes net, 5
- BDMO Algorithm, 283
- Beer and diapers, 218
- Bell, R., 353
- Bellkor's Pragmatic Chaos, 322
- Bengio, Y., 569
- Berges, C.J.C., 570
- Bergmann, R., 78
- Berkhin, P., 212

- Berrar, D.P., 461
 Betweenness, 363
 BFR Algorithm, 266, 269
 BFS, *see* Breadth-first search
 Bi-clique, 370
 Bias vector, 530
 Bickson, D., 79
 Bid, 303, 305, 312, 313
 Big data, 1
 BigTable, 77
 Bik, A.J.C., 79
 Binary Classification, 464
 Biomarker, 217
 Bipartite graph, 299, 359, 369, 370
 BIRCH Algorithm, 292
 Birrell, A., 79
 Bitmap, 232, 233
 Block, 13, 21, 192
 Blockeel, H., 521
 Blocking property, 43, 49
 Blog, 200
 Bloom filter, 152, 230
 Bloom, B.H., 174
 Blum, A., 521
 Bohannon, P., 78
 Boldi, P., 426
 Bonferroni correction, 6
 Bonferroni's principle, 5, 6
 Bookmark, 198
 Boral, H., 427
 Borkar, V., 77, 78
 Boser, B., 570
 Bottou, L., 521
 BPTT, *see* Backpropagation through time
 Bradley, P.S., 292
 Breadth-first search, 363
 Breiman, L., 20
 Brick-and-mortar retailer, 216, 320, 321
 Brin, S., 212
 Broad matching, 305
 Broder, A.Z., 20, 141, 212
 Bu, Y., 78
 Bucket, 10, 149, 164, 168, 230, 283
 Budget, 304, 311
 Budiu, M., 79
 Bulk-synchronous system, 51
 Burges, C.J.C., 521
 Burrows, M., 78
 Buxketing, 559
 Caffè, 530, 569
 Candidate itemset, 227, 240
 Candidate pair, 81, 100, 231, 234
 Carey, M., 77, 78
 Categorical feature, 464, 511, 517
 Cell state, 562
 Centroid, 255, 258, 264, 267, 271
 Chabbert, M., 353
 chain rule, 542
 Chandra, T., 78
 Chang, F., 78
 Channel, 548
 Characteristic matrix, 89
 Charikar, M.S., 141
 Chaudhuri, S., 141
 Checkpoint, 52
 Chen, M.-S., 250
 Child, 364
 Cholera, 4
 Chowdhury, M., 79
 Chronicle data model, 173
 Chunk, 24, 240, 270
 CineMatch, 349
 Class, 464
 Classification loss, 537
 Classifier, 330, 463
 Click stream, 145
 Click-through rate, 297, 303
 Clique, 369
 Cloud computing, 17
 Cluster computing, 21, 22
 Cluster tree, 278, 279
 Clusters, 77
 Clustering, 4, 17, 253, 337, 355, 361, 463
 Clustroid, 258, 264
 CNN, *see* Convolutional neural network
 Collaboration network, 358

- Collaborative filtering, 5, 18, 84, 293, 319, 333, 359
- Colossus, 24
- Column-orthonormal matrix, 443
- Combiner, 27, 189, 191
- Communication cost, 22, 54, 403
- Community, 18, 355, 366, 368, 399
- Commutativity, 27, 45
- Competitive ratio, 17, 298, 301, 306
- Complete graph, 369, 370
- Compressed set, 270
- Compute graph, 540
- Compute node, 21, 22
- Computer game, 327
- Concept, 444
- Concept space, 449
- Conductance, 397
- Confidence, 218, 219
- Content-based recommendation, 319, 324
- Convergence, 475
- Convexity, 517
- Convolutional layer, 549, 551
- Convolutional neural network, 18, 523, 527, 548
- Cooper, B.F., 78
- Coordinates, 254
- Cortes, C., 521, 570
- Cosine distance, 107, 117, 325, 330, 450
- Counting ones, 162, 283
- Covering an output, 66
- Craig's List, 294
- Craswell, N., 317
- Credit, 364
- Cristianini, N., 521
- Cross entropy, 533, 538
- Cross-Validation, 469
- Crowdsourcing, 470
- CUR-decomposition, 429, 452
- CURE Algorithm, 274, 278
- Currey, J., 79
- Curse of dimensionality, 256, 280, 503, 518
- Cut, 374, 375
- Cyclic permutation, 99
- Cylinder, 13
- Czajkowski, G., 79
- DAG, *see* Directed acyclic graph
- Darts, 152
- Das Sarma, A., 78
- Das, T., 79
- Dasgupta, A., 426
- Data mining, 1
- Data science, 1
- Data stream, 17, 244, 282, 296, 484
- Data-stream-management system, 144
- Database, 17
- Datar, M., 142, 174, 292
- Datar-Gionis-Indyk-Motwani Algorithm, 163
- Dave, A., 79
- De Raedt, L., 521
- Dead end, 179, 182, 183, 205
- Dean, J., 78
- Decaying window, 169, 246
- Decision forest, 515, 518
- Decision tree, 18, 330, 464, 467, 468, 505, 518
- Deep learning, 3, 18, 523
- Deerwester, S., 460
- Degree, 371, 400
- Degree matrix, 376
- Dehnert, J.C., 79
- del.icio.us, 326, 359
- Deletion, 108
- Denker, J.S., 570
- Dense matrix, 31, 452
- Density, 263, 265
- Density of edges, 396
- Depth-first search, 417
- Determinant, 431
- Development set, 468
- DeWitt, D.J., 78
- DFS, *see* Distributed file system
- Diagonal matrix, 443
- Diameter, 263, 265, 406
- Diapers and beer, 216
- Difference, 34, 37, 41

- Dimension table, 60
- Dimensionality reduction, 18, 340, 429, 503
- Directed acyclic graph, 364
- Directed graph, 405
- Discard set, 270
- Disk, 13, 221, 255, 278
- Disk block, *see* Block
- Display ad, 294, 295
- Distance measure, 105, 253, 361, 538
- Distinct elements, 154, 157
- Distributed file system, 21, 24, 214, 221
- DMOZ, *see* Open directory
- Document, 82, 86, 217, 254, 313, 325, 326, 466
- Document frequency, *see* Inverse document frequency
- Domain, 202
- Dot product, 107, 544
- Drineas, P., 460
- Dropout, 566
- Dryad, 77
- DryadLINQ, 77
- Dual construction, 360
- Dubitzky, W., 461
- Dumais, S.T., 460
- Dup-elim task, 49
- Dying ReLU, 535
- e*, 13
- Edit distance, 108, 110
- Eigenpair, 431
- Eigenvalue, 179, 377, 430, 440, 441
- Eigenvector, 179, 377, 430, 435, 440
- ELU, 535
- Email, 358
- Energy, 448
- Ensemble, 331
- Ensemble methods, 516
- Entity resolution, 122, 123
- Entropy, 508, 537
- Equijoin, 34
- Erlingsson, I., 79
- Ernst, M., 78
- Ethernet, 21, 22
- Euclidean distance, 105, 119, 502
- Euclidean space, 105, 109, 254, 255, 258, 274
- Ewen, S., 78
- Exclusive-or, 529
- Explainability, 3
- Exploding derivative, 531
- Exploding gradient, 562
- Exponential linear unit, *see* ELU
- Exponentially decaying window, *see* Decaying window
- Extrapolation, 501
- Facebook, 18, 198, 356
- Fact table, 60
- Failure, 23, 30, 43, 51
- Faloutsos, C., 427, 461
- 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, 470
- Feature vector, 464, 517
- Feedforward network, 531
- Fetterly, D., 79
- Fikes, A., 78
- File, 23, 24, 221, 239
- Filter, 44, 549
- Filtering, 151
- Fingerprint, 125
- First-price auction, 305
- Fixedpoint, 114, 204
- Flajolet, P., 174
- Flajolet-Martin Algorithm, 155, 419
- Flatmap, 44
- Flattening, 547
- Flink, 77, 78
- Flow graph, 42
- Forget gate, 563
- Fortunato, S., 426
- Fotakis, D., 426
- Franklin, M.J., 79

- French, J.C., 292
Frequent bucket, 231, 233
Frequent itemset, 5, 214, 224, 226, 370, 463
Frequent pairs, 225
Frequent-items table, 226
Freund, Y., 521
Freytag, J.-C., 78
Friends, 356
Friends relation, 59
Frieze, A.M., 141
Frobenius norm, 433, 447
Fukushima, K., 569
Fully connected layer, 527
Furnas, G.W., 460
- Gaber, M.M., 20
Ganti, V., 141, 292
Garcia-Molina, H., 20, 212, 250, 292, 427
Garofalakis, M., 174
Gate, 562
Gated recurrent unit, 564
Gaussian elimination, 180
Gehrke, J., 174, 292
Generalization, 469
Generated subgraph, 370
Genre, 324, 336, 350
GFS, *see* Google file system
Ghemawat, S., 78
Gibbons, P.B., 174, 427
GINI impurity, 508
Gionis, A., 142, 174
Giraph, 77, 78
Girvan, M., 426
Girvan-Newman Algorithm, 363
Global minimum, 342
GN Algorithm, *see* Girvan-Newman Algorithm
Gobioff, H., 78
Golub, G.H., 460
Gonzalez, J., 79
Google, 176, 187, 302
Google file system, 24
Google+, 356
GPU, *see* Graphics processing unit, 556
Gradient descent, 18, 48, 348, 386, 492, 526, 529, 531, 539
Granzow, M., 461
Graph, 51, 64, 356, 399, 406
Graphics processing unit, 530
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
GRU, *see* Gated recurrent unit
Gruber, R.E., 78
Guestrin, C., 79
Guha, S., 292
Gunda, P.K., 79
Gyongyi, Z., 212
- Hadamard product, 544, 563
Hadoop, 25, 79
Hadoop distributed file system, 24
HaLoop, 51, 77
Hamming distance, 74, 109, 117
Hanazawa, T., 570
Harris, M., 350
Harshman, R., 460
Hash function, 10, 87, 92, 100, 149, 152, 155
Hash join, 37
Hash key, 10, 312
Hash table, 10, 12, 13, 223, 230, 233, 234, 312, 314, 400
Haveliwala, T.H., 212
HDFS, *see* Hadoop distributed file system
Head, 413
Heavy hitter, 400
Heise, A., 78
Hellerstein, J.M., 79
Henderson, D., 570
Henzinger, M., 142

- Hidden layer, 525
- Hidden state, 558
- Hierarchical clustering, 255, 257, 275, 338, 361
- Hinge loss, 491
- Hinton, G.E., 569, 570
- HITS, 204
- Hive, 77, 79
- Hochreiter, S., 569
- Hoger, M., 78
- Hopcroft, J.E., 417
- Hopfield, J.J., 569
- Horn, H., 79
- Howard, R.E., 570
- Howe, B., 78
- Hsieh, W.C., 78
- Hub, 204
- Hubbard, W., 570
- Huber loss, 536
- Hueske, F., 78
- Hyperbolic tangent, 532, 562
- Hyperlink-induced topic search, *see* HITS
- Hyperparameter, 535
- Hyperplane, 486
- Hyracks, 77

- Identical documents, 130
- Identity matrix, 431
- IDF, *see* Inverse document frequency
- Image, 145, 325, 326
- ImageNet, 548, 553, 569
- IMDB, *see* Internet Movie Database
- Imielinski, T., 250
- Immediate subset, 242
- Immorlica, N., 142
- Important page, 176
- Impression, 294
- Impurity, 507
- In-component, 181
- Inaccessible page, 199
- Independent rows or columns, 443
- Index, 11, 400
- Indyk, P., 141, 142, 174
- Information integration, 6
- Initialize clusters, 267

- Input, 64, 464
- Input gate, 563
- Input layer, 525
- Insertion, 108
- Instance-based learning, 467
- Interest, 218
- Internet Movie Database, 324, 350
- Interpolation, 501
- Intersection, 34, 36, 40, 85
- Into Thin Air*, 323
- Inverse document frequency, 9, *see* TF.IDF
- Inverted index, 176, 294
- Ioannidis, Y.E., 426
- 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, 504
- Jaccard similarity, 82, 91, 104, 199
- Jacobian, 541
- Jacobsen, H.-A., 78
- Jagadish, H.V., 174
- Jahrer, M., 353
- Jeh, G., 426
- Joachims, T., 521
- Join, *see* Natural join, 45, *see* Multi-way join, *see* Star join, 401
- Join task, 49

- K-means, 266
- K-partite graph, 359
- Kahan, W., 460
- Kalyanasundaram, B., 318
- Kamm, D., 350
- Kang, U., 427
- Kannan, R., 460
- Kao, O., 78
- Karlin, A., 298
- Kaushik, R., 141
- Kautz, W.H., 174
- Kernel, 552

- Kernel function, 498, 502
- Key component, 149
- Key-value pair, 25, 27
- Keyword, 303, 331
- KL-divergence, 538
- Kleinberg, J.M., 212
- Knuth, D.E., 20
- Koren, Y., 353
- Krioukov, A., 78
- Krizhevsky, A., 569
- Kumar, R., 20, 79, 212, 426
- Kumar, V., 20
- Kyrola, A., 79

- Label, 356, 464
- Lagrangian multipliers, 58
- Landauer, T.K., 460
- Lang, K.J., 426, 570
- Laplacian matrix, 376
- Layer, 523
- Lazy evaluation, 46
- LCS, *see* Longest common subsequence
- Leaf, 365
- Leaky ReLU, 535
- Learning rate, 472, 545
- LeCun, Y., 569, 570
- Leich, M., 78
- Leiser, N, 79
- Length, 158, 405
- Length indexing, 131
- Leser, U., 78
- Leskovec, J., 426, 427
- Leung, S.-T., 78
- Li, P., 142
- Likelihood, 381
- Lin, S., 142
- Linden, G., 353
- Lineage, 47
- Linear equations, 180
- Linear separability, 471, 475
- Linear transitive closure, 410, 415
- Link, 33, 176, 190
- Link matrix of the Web, 205
- Link spam, 195, 199
- Littlestone, N., 521

- Livny, M., 292
- Local minimum, 342
- Locality, 356
- Locality-sensitive family, 116
- Locality-sensitive function, 112
- Locality-sensitive hashing, 1, 81, 100, 111, 326, 504
- Log likelihood, 387
- Logarithm, 13
- Logistic sigmoid, 532
- Long short-term memory network, 18, 523, 562
- Long tail, 216, 320, 321
- Longest common subsequence, 108
- Low, Y., 79
- Lower bound, 68
- Lower hyperplane, 487
- LSH, *see* Locality-sensitive hashing
- LSTM, *see* Long short-term memory network

- Ma, J., 79
- Machine learning, 2, 18, 330, 463
- Maggioni, M., 460
- Maghoul, F., 20, 212
- Mahalanobis distance, 273
- Mahoney, M.W., 426, 460
- Main memory, 221, 222, 230, 255
- Malewicz, G, 79
- Malik, J., 427
- Manber, U., 142
- Manhattan distance, 106
- Manning, C.P., 20
- 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, 21, 25, 30, 189, 191, 241, 287, 401, 408, 483
- Margin, 485
- Market basket, 5, 17, 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, 441
- Matthew effect, 16
- Max pooling, 553
- Maximal itemset, 224
- Maximal matching, 299
- Maximum-likelihood estimation, 381
- McAuley, J., 427
- McCauley, M., 79
- Mean, *see* Average
- Mean squared error, 536
- Mechanical Turk, 470
- Median, 156
- Mehta, A., 318
- Melnik, S., 427
- Merging clusters, 258, 261, 272, 276, 281, 285
- Merton, P., 20
- Miller, G.L., 427
- Minhashing, 82, 90, 103, 106, 113, 326
- Minibatch gradient descent, 496
- Minicluster, 270
- Minsky, M., 522
- Minutiae, 125
- Mirroknii, V.S., 142
- Mirror page, 83
- Mitzenmacher, M., 141
- ML, *see* Machine learning
- MLE, *see* Maximum-likelihood estimation
- MNIST dataset, 547
- Model, 381
- Modeling, 1
- Moments, 157
- Monotonicity, 224
- Montavon, G., 521
- Moore-Penrose pseudoinverse, 453
- Most-common elements, 169
- Motwani, R., 142, 174, 250, 292
- Mueller, K.-R., 521
- Multiclass classification, 464, 479
- Multidimensional index, 503
- Multihash Algorithm, 234
- Multiplication, 31, *see* Matrix multiplication, 189, 204
- Multiset, *see* Bag
- Multistage Algorithm, 232
- Multiway join, 56, 402
- 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, 18, 464, 468, 497, 518
- Negative border, 242
- Negative example, 472
- Neighbor, 390
- Neighborhood, 406, 418
- Neighborhood profile, 406
- Netflix challenge, 2, 322, 349, 516
- Network, *see* Social network
- Neural net, 18, 467, 523, 524
- Neuron, 527
- Newman, M.E.J., 426
- Newspaper articles, 128, 313, 322
- Node, 508, 523
- Node pruning, 514
- Non-Euclidean distance, 264, *see* Cosine distance, *see* Edit distance, *see* Hamming distance, *see* Jaccard distance

- Non-Euclidean space, 278, 280
- Norm, 105, 106
- Norm penalty, 565
- Normal distribution, 269
- Normalization, 333, 335, 346
- Normalized cut, 375
- NP-complete problem, 369
- Numerical feature, 464, 509, 517

- O’Callaghan, L., 292
- Off-line algorithm, 296
- Olston, C., 79
- Omiecinski, E., 250
- On-line advertising, *see* Advertising
- On-line algorithm, 17, 296
- On-line learning, 469
- On-line retailer, 216, 294, 320, 321
- Onose, N., 78
- Open directory, 196, 470
- OR-construction, 114
- Orr, G.B., 521
- Orthogonal vectors, 256, 434
- Orthonormal matrix, 443, 448
- Orthonormal vectors, 435, 438
- Out-component, 181
- Out-degree, 416
- Outlier, 255
- Output, 64, 464
- Output gate, 563
- Output layer, 525
- Overfitting, 331, 348, 467, 468, 481, 506, 514, 518, 565
- Overlapping Communities, 381
- Overture, 303
- Owen, A.B., 142
- Own pages, 200

- Paepcke, A., 142
- Page, L., 175, 212
- PageRank, 4, 17, 31, 32, 49, 175, 177, 189
- Pairs, *see* Frequent pairs
- Palmer, C.R., 427
- Pan, J.-Y., 427
- Papert, S., 522

- Parallelism, 513
- Parametric ReLU, 535
- Parent, 364
- Park, J.S., 250
- Partition, 374
- Pass, 222, 225, 233, 238
- Path, 405
- Paulson, E., 78
- PCA, *see* Principal-component analysis
- PCY Algorithm, 230, 233, 234
- Pedersen, J., 212
- Perceptron, 18, 463, 467, 471, 517, 523, 524
- Perfect matching, 299
- Permutation, 90, 99
- Peters, M., 78
- Phishing, 2
- PIG, 77
- Pigeonhole principle, 369
- Piotte, M., 353
- Pivotal condensation, 431
- Plagiarism, 83, 217
- Pnuts, 77
- Point, 253, 283
- Point assignment, 255, 266, 362
- Polyzotis, A., 77
- Pooling function, 553
- Pooling layer, 527, 553
- Position indexing, 133, 135
- Positive example, 472
- Positive integer, 168
- Powell, A.L., 292
- Power iteration, 431, 432
- Power law, 14
- Predicate, 330
- Prefix indexing, 132, 133, 135
- Pregel, 51, 77
- Principal eigenvector, 179, 431
- Principal-component analysis, 429, 436
- 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
- PyTorch, 530, 570

- Quadratic programming, 492
- Query, 146, 165, 287
- Query example, 498
- Quinlan, J.R., 522

- R-tree, 292
- Rack, 22
- Radius, 263, 265, 406
- Raghavan, P., 20, 212, 426
- Rahm, E., 427
- Rajagopalan, S., 20, 212, 426
- Ramakrishnan, R., 78, 292
- Ramsey, W., 317
- Random hyperplanes, 117, 326
- Random interconnection layer, 527
- Random surfer, 176, 177, 182, 196, 390
- Randomization, 238
- Rank, 442
- Rarest-first order, 313
- Rastogi, R., 174, 292
- Rating, 320, 323
- RDD, *see* Resilient distributed dataset
- Reachability, 407, 408, 415
- Recommendation system, 18, 319
- Rectified linear unit, *see* ReLU
- Recurrent neural network, 18, 523, 557
- Recursion, 49
- Recursive doubling, 411, 415
- Reduce task, 25, 27
- Reduce worker, 28, 30
- Reducer, 27, 45
- Reducer size, 61, 67
- Reed, B., 79
- Regression, 464, 502, 518
- Regression loss, 536
- Regularization, 565
- Regularization parameter, 490
- Reichsteiner, A., 461
- Reina, C., 292
- Relation, 33
- Relational algebra, 33
- ReLU, 534, 562
- Replication, 24
- Replication rate, 61, 68
- Representation, 278
- Representative point, 275
- Representative sample, 149
- Reservoir sampling, 174
- Residual PageRank, 393
- Resilient distributed dataset, 43
- Response, 549
- Restart, 391
- Retained set, 270
- Revenue, 304
- Rheinlander, A., 78
- Ripple-carry adder, 168
- RMSE, *see* Root-mean-square error
- RNN, *see* Recurrent neural network
- Robinson, E., 78
- Rocha, L.M., 461
- Root-mean-square error, 322, 341, 447, 536
- Rosa, M., 426
- Rosenblatt, F., 522
- Rounding data, 335
- Row, *see* Tuple
- Row-orthonormal matrix, 448
- Rowsum, 278
- Royalty, J., 78
- Rumelhart, D.E., 570

- S-curve, 102, 111
- Saberi, A., 318
- Salihoglu, S., 78
- Sample, 238, 242, 245, 247, 267, 275, 279
- Sampling, 148, 162
- Saturation, 531
- Savasere, A., 250
- Sax, M.J., 78

- SCC, *see* Strongly connected component, *see* Strongly connected component
 Schapire, R.E., 521
 Schelter, S., 78
 Schema, 33
 Schmidhuber, J., 569
 Schutze, H., 20
 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, 409, 411, 413, 414
 Sensor, 145
 Sentiment analysis, 471
 Set, 89, 131, *see* Itemset
 Set difference, *see* Difference
 Shankar, S., 78
 Shawe-Taylor, J., 521
 Shenker, S., 79
 Shi, J., 427
 Shikano, K., 570
 Shim, K., 292
 Shingle, 86, 103, 128
 Shivakumar, N., 250
 Shopping cart, 216
 Shortest paths, 52
 Siddharth, J., 142
 Sigmoid, 532
 Signature, 82, 89, 91, 103
 Signature matrix, 92, 100
 Silberschatz, A., 174
 Silberstein, A., 78
 Similarity, 5, 17, 82, 213, 326, 334
 Similarity join, 62
 Simonyan, K., 554
 Simrank, 389
 Singleton, R.C., 174
 Singular value, 443, 447, 448
 Singular-value decomposition, 340, 429, 442, 452
 Six degrees of separation, 408
 Sketch, 119
 Skew, 28
 Sliding window, 146, 162, 169, 283
 Smart transitive closure, 413, 415
 Smith, B., 353
 SNAP, 426
 Social Graph, 356
 Social network, 18, 355, 356, 429
 Softmax, 533
 SON Algorithm, 240
 Source, 405
 Source node, 389
 Space, 105, 253
 Spam, *see* Term spam, *see* Link spam, 358, 470
 Spam farm, 199, 202
 Spam mass, 202, 203
 Spark, 41, 43, 51, 77, 79
 Sparse matrix, 31, 90, 91, 189, 190, 320
 Spectral partitioning, 374
 Spider trap, 182, 185, 205
 Split, 46
 Splitting clusters, 281
 SQL, 22, 33, 77, 79
 Squared error, 536
 Squares, 404
 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., 20, 212
 Statistical model, 2
 Status, 313
 Steinbach, M., 20
 Step function, 530, 531
 Stochastic gradient descent, 348, 495
 Stochastic matrix, 179, 431
 Stoica, I., 79
 Stop clustering, 259, 263, 265
 Stop words, 9, 88, 128, 217, 325
 Stratosphere, 77

- Stream, *see* Data stream
- Strength of membership, 387
- Stride, 551
- String, 131
- Striping, 32, 189, 191
- Strong edge, 358
- Strongly connected component, 181, 417
- Strongly connected graph, 179, 406
- Substochastic matrix, 182
- Suffix length, 135
- Summarization, 4
- Summation, 168
- Sun, J., 461
- Supercomputer, 21
- Superimposed code, *see* Bloom filter, 173
- Supermarket, 216, 238
- Superstep, 52
- Supervised learning, 463, 465
- Support, 214, 239, 240, 242, 244
- Support vector, 486
- Support-vector machine, 18, 463, 468, 485, 517
- Supporting page, 200
- Suri, S., 427
- Surprise number, 158
- Sutskever, I., 569
- SVD, *see* Singular-value decomposition
- SVM, *see* Support-vector machine
- Swami, A., 250
- Symmetric matrix, 377, 430
- Szegedy, M., 174

- Tag, 326, 359
- Tail, 413
- Tail length, 155, 419
- Tan, P.-N., 20
- Target, 405
- Target page, 200
- Tarjan, R.E., 417
- Task, 23
- Taxation, 182, 185, 200, 205
- Taylor expansion, 14
- Taylor, M., 317
- Telephone call, 358

- Teleport set, 196, 197, 202, 391
- Teleportation, 186
- Tendril, 181
- Tensor, 48, 546
- TensorFlow, 41, 79, 530, 538, 543, 570
- Term, 176
- Term frequency, 9, *see* TF.IDF
- Term spam, 176, 199
- Test set, 468, 475
- TF, *see* Term frequency
- TF.IDF, 9, 325, 467
- Theobald, M., 142
- Thrashing, 191, 230
- Threshold, 102, 171, 214, 240, 244, 471, 477
- TIA, *see* Total Information Awareness
- Timestamp, 163, 284
- Toivonen's Algorithm, 242
- Toivonen, H., 250
- Tomkins, A., 20, 79, 212, 426
- Tong, H., 427
- Topic-sensitive PageRank, 195, 202
- Toscher, A., 353
- Total Information Awareness, 6
- Touching the Void*, 323
- Training example, 464
- Training rate, 475
- Training set, 463, 464, 470, 480
- Transaction, *see* Basket
- Transformation, 44
- Transition matrix, 391
- Transition matrix of the Web, 178, 189, 190, 192, 429
- Transitive closure, 49, 407
- Transitive reduction, 418
- Transpose, 205
- Transposition, 111
- Tree, 260, 278, 279, *see* Decision tree
- Triangle, 399
- Triangle inequality, 105
- Triangular matrix, 223, 232
- Tripartite graph, 359
- Triples method, 223, 232
- TrustRank, 202
- Trustworthy page, 202

- Tsourakakis, C.E., 427
 Tube, 182
 Tuple, 33
 Tuzhilin, A., 352
 Twitter, 18, 313, 356
 Tzoumas, K., 78

 Ullman, J.D., 20, 77–79, 250, 292, 426
 Undirected graph, *see* Graph
 Union, 34, 36, 40, 85
 Unit vector, 430, 435
 Universal set, 131
 Unsupervised learning, 463
 Upper hyperplane, 487
 User, 320, 336, 337
 User profile, 328
 Utility matrix, 320, 323, 340, 429
 UV-decomposition, 340, 350, 429, 496

 VA file, 503
 Valduriez, P., 427
 Validation set, 468, 567
 Van Loan, C.F., 460
 Vanishing gradient, 561
 Vapnik, V.N., 521
 Variable, 158
 Vassilvitskii, S., 427
 Vazirani, U., 318
 Vazirani, V., 318
 Vector, 31, 105, 109, 179, 189, 204, 205, 254
 Vernica, R., 78
 VGGnet, 553
 Vigna, S., 426
 Vitter, J., 174
 Volume, 397
 Volume (of a set of nodes), 375
 von Ahn, L., 327, 353
 von Luxburg, U., 427
 Voronoi diagram, 498

 Waibel, A., 570
 Wall, M.E., 461
 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, 471, 526
 Weiner, J., 20, 212
 Whizbang Labs, 3
 Widom, J., 20, 79, 174, 292, 426
 Wikipedia, 358, 470
 Williams, R.J., 570
 Window, *see* Sliding window, *see* Decaying window
 Windows, 13
 Winnow Algorithm, 475
 Word, 217, 254, 325
 Word count, 25, 44
 Worker process, 28
 Workflow, 42, 49, 54
 Working store, 144

 Xiao, C., 142
 Xie, Y., 461

 Yahoo, 303, 326
 Yang, J., 427
 Yerneni, R., 78
 York, J., 353
 Yu, J.X., 142
 Yu, P.S., 250
 Yu, Y., 79

 Zaharia, M., 79
 Zero padding, 551, 559
 Zhang, C.H., 142
 Zhang, H., 461
 Zhang, T., 292
 Zipf's law, 15, *see* Power law
 Zoeter, O., 317
 Zussman, A., 554