

---

# Index

Page references followed by b denote boxes; those followed by f denote figures; those followed by t denote tables.

## A

- Abbeel, Pieter, 80
- ABI BioScope, 182
- ABNER, 291t
- Abstraction, 7, 11
- ABySS, 171, 171t
- Accuracy, 27, 54–55, 55f
- Accurate mass and time tag (AMT) approach, 193–194
- Actin cytoskeleton pathway, regulation of, 235
- Active contour algorithm, 97–98, 98f
- Additive color mixing, 85
- Affymetrix, 112, 113, 180
- Agglomerative clustering, 77
- Agilent, 110–113, 180, 181
- AILUN, 265
- Akt, 242f, 243–244
- Algorithm(s)
  - active contour, 97–98, 98f
  - CHIP-seq, 173–177
  - Cocke-Kasami-Younger (CKY), 290
  - computational image analysis, 101–103
  - conditioned random fields, 298, 298t
  - constant-time, 12–13
  - cubic, 13
  - described, 11–12
  - Earley, 290
  - ease of implementation, 15
  - expectation–maximization (EM), 119, 212, 214
  - experimental time, 13
  - feature selection, 65–66
  - greedy, 72
  - image registration, 101–103
    - feature-based algorithm, 102–103
    - intensity-based registration, 101–102
    - mutual information theoretic technique, 102
  - k*-nearest neighbors, 67–68, 298t
  - level set, 98–99, 99f–100f
  - linear time, 13
  - machine learning, 62–78
    - classification task, 49, 52f
    - classifier, 49–50
    - clustering, 51, 52f
    - data, 62–67
    - features selection, 52
    - probabilistic models, 72–76, 73f
    - regression task, 50, 52f, 57
    - semisupervised learning, 51, 52f
    - supervised learning, 50, 52–53, 52f, 67–72
    - terminology, 49–53
    - training phase, 50
    - unsupervised learning, 51, 52f, 53, 76–78, 76f
  - maximum entropy, 298, 298t
  - naive Bayes, 68–69, 298, 298t
  - parallelizability, 14–15
  - partitioning, 71–72, 118–119
  - performance evaluation, 53–57, 55f
  - quadratic, 13
  - random forests, 298, 298t
  - running time analysis, 12–13
  - running time of, 65
  - SEQUEST, 191–192, 192b
  - space complexity, 13–14
  - support vector machines (SVM), 298, 298t
- Alignment
  - anchor, 164–165
  - indel, 163–164
  - paired-end, 164–165, 165f
  - partners, 164–165
  - programs for, 166–167, 166t
  - split-read, 168–169, 168f
- Alleles
  - biallelic, 126
  - defined, 126
  - Hardy–Weinberg equilibrium, 137
  - major, 126
  - maternal, 126
  - minor, 126
  - paternal, 126

## 306 Index

- Allelic odds ratio, in GWAS, 129–130
  - Alternative hypothesis, 26
  - Alzheimer's disease, 235
  - AMT (accurate mass and time tag) approach, 193–194
  - Anchor, alignment, 164–165
  - Annotation
    - biomedical text, 292
    - incomplete and inaccurate annotations, 233–234, 235f
  - ANOVA (analysis of variance), 30
  - Anscombe's quartet, 41, 42f, 43
  - Anxiety, bioinformatics, 1
  - Archon X Prize for Genomics, 184
  - Area under the curve, 56f, 57
  - Arginine, heavy, 198
  - ArrayWxpress, 261
  - Association
    - genome-wide association study (GWAS), 125–150
    - genotype–phenotype, 128–130, 128f
    - interpreting genetic association, 144–145
  - Association testing, 137–141
    - $\chi^2$  test of statistical independence, 139–140
    - improving statistical power, 141–144
      - alternate tests of association, 141–142
      - alternate types of variation, 142
      - genotype imputation, 142–144, 143f
      - meta-analysis, 144
  - Assortative mating, 137
  - Assumptions, 26
  - Autosomes, 126
  - Auxiliary space complexity, 14
- ### B
- Bagging, 72
  - BAM file format, 165
  - BANNER, 291t
  - Base-calling error probability, 159
  - Bayesian networks, 73–74, 73f, 245–259
    - in action, 255–256
    - chain rule, 247
    - dynamic Bayesian networks (DBNs), 244–245
    - joint probability distribution, 245–248
    - learning signaling pathway structure from flow cytometry data, 256
    - Markov assumptions, 247
    - model properties, 251–255
      - causality, 254–255
      - dependencies and independencies in graph structure, 251–253
    - model semantics, 245–246
    - notation, 246–249
    - structure learning, 249–251
      - model averaging, 251
      - scoring, 249–250, 250b
      - searching the space of possible graph structures, 250–251
  - Bayes' Rule, 20–21, 249
  - Bayes' theorem, 68
  - Benjamini–Hochberg false discovery rate, 36, 121
  - Biallelic, 126
  - Bias, described, 23
  - Binary images, 93
  - Binning
    - 1D, 209–210
    - 2D, 210–211, 211f
  - Bioconductor, 113, 265, 266
  - Biomedical images, 87–90
    - computed tomography (CT), 88, 89f
    - magnetic resonance imaging (MRI), 89, 89f
    - microscope images, 87–88, 88f
    - positron emission tomography (PET), 90
  - Biomedical text, 285–301
    - annotation, 292
    - applications, 286–287
      - biosurveillance, 287
      - consumer health informatics, 287
      - data integration, 287
      - document classification, 286–287
      - drug discovery, 287
      - health services delivery, 287
    - clinical document classification example, 298–300, 299t–300t
    - goal of mining, 285–286
    - growth in biomedical literature, 285–301
    - machine learning, 297–300, 298t–300t
    - named entity recognition, 290–291, 291t
    - ontologies in biomedicine, 295–297, 296f
    - preprocessing raw text, 288–290, 288f–289f
      - chunking and parsing, 288f–289f, 289–290
      - part of speech tagging, 288–289, 288f
      - stemming, 290
      - stop word removal, 290
      - tokenization, 288, 288f
    - processing, 300–301
    - standard terminologies, 292–295, 294t–295t
  - BioPortal, 297
  - Biosurveillance, 287
  - Bio Tagger-GM, 291t
  - Bishop, Christopher M., 79
  - Bits in images, 86b
  - BLAST, 158–159, 166t, 167
  - BLAT (BLAST-like alignment tool), 166t, 167
  - Blocks, 109
  - Bonferroni procedure, 35, 121, 145
  - Boolean feature, 64–65
  - Boolean networks, 242–244, 242f, 245
  - Boosting, 72
  - Bootstrapping, 38
  - Bowtie, 166, 166t
  - Box-Cox transformation, 207
  - Brain images, analysis of, 96–99, 97f–98f
    - active contour algorithm, 97–98, 98f
    - k*-means, 97, 97f
    - level set algorithm, 98–99, 99f–100f

- BreakDancer, 183  
Burroughs–Wheeler aligner (BWA), 166–167, 166t  
Burroughs–Wheeler aligner, Smith–Waterman alignment (BWA-SW), 166t, 167  
Byte, 8
- C**
- Candidate mapping locations, 161  
Canny, John F., 104b  
Canny edge detection, 104b–105b  
CASAVA, 182  
Cases, in GWAS, 128–129  
Categorical data  
  defined, 31  
  mixing categorical and continuous data, 33b–34b  
  statistical tests of, 31–33  
Causal interpretation, 254  
Causality, 254–255  
Cell images, 92–96, 92f–96f  
  *k*-means clustering, 95–96, 96f  
  Otsu’s method for image segmentation, 94–95, 94f–95f  
Censored data, 41–42  
Central dogma, 107, 187  
Central limit theorem, 29  
Central processing unit (CPU), 8, 8t  
CGH (comparative genomic hybridization), 180  
Chain rule, 247  
Chain-termination method, 155  
Channels, 84–85, 87f  
Chart parsing, 290  
ChIP-on-chip (chromatin immunoprecipitation on chip), 107, 172  
CHIP-seq, 172–179  
  advantages, 178  
  algorithms, 173–177  
  filtering, 176  
  identification of regions of enrichment, 175–176  
  ranking by significance, 176–177  
  signal shifting, 174–175, 175f  
  smoothed signal creation, 173–174, 173f  
  artifacts, 176, 177f  
  overview, 172–173  
  practical considerations, 177–178  
    confidence estimate, 177  
    performance, 177  
    usability, 178  
  software packages, 178, 179f  
 $\chi^2$  test of statistical independence, 32, 139–140  
Chromatin immunoprecipitation (ChIP), 172  
Chromatin immunoprecipitation on chip (ChIP-on-chip), 107, 172  
Chromosomes  
  autosomes, 126  
  sex, 126  
Chunking, 288f–289f, 289–290  
CIGAR, 166  
CisGenome, 178, 179f  
CKY (Cocke-Kasami-Younger) algorithm, 290  
Classification in a 2D feature space, 69–70, 69f  
Classification task, 49, 52f  
Classifier, 49–50  
  complexity of, 62  
  decision tree, 71–72, 71f  
  generalization, 58  
  linear, 70  
  overfitting, 62–64, 63f  
  performance evaluation and, 53–57, 55f  
  testing set, 58–61, 59f  
  training multiple, 72  
  training set, 58–61, 59f  
Closed-form expression, 142  
Cloud computing services, 9  
Clustering, 51, 52f  
  agglomerative, 77  
  divisive, 77  
  example of, 76f  
  flow cytometry, 211–215  
  hard clusters, 119  
  hierarchical, 76f, 77, 116–118, 116f–117f  
  *k*-means, 77–78  
    brain images, 97, 97f  
    cell images, 95–96, 96f  
    flow cytometry, 212–215  
    semisupervised clustering, 117–120, 119f  
  semisupervised clustering methods, 117–120, 119f  
  unsupervised clustering methods, 115–117, 116f–117f  
  unsupervised learning algorithms, 76–78  
Cluster plot, 136–137, 136f  
CMYK image, 85, 87f  
CNVer, 182  
CNVs (copy-number variants), 142, 180, 182  
CNV-seq, 182  
Cochran-Armitage trend test, 141  
Cocke-Kasami-Younger (CKY) algorithm, 290  
Color mixing  
  additive, 85  
  subtractive, 85  
Color space, 84–85  
Color space encoding, 157, 157f  
Comparative genomic hybridization (CGH), 180  
Comparative Toxicogenomics Database, 270  
Compensation, 206–207, 206f  
Complete Genomics, 183–184  
Complexity penalty, 250b  
Computational image analysis, 92–103  
  algorithms, 101–103  
  brain images, 96–99, 97f–98f  
  cell images, 92–96, 92f–96f  
  edge detection, 103, 104b–105b  
  image registration, 99–101  
Computed tomography (CT), 88, 89f

## 308 Index

- Computers
    - hardware components of, 8, 8t
    - limitations of, 12
    - overview of, 8–9
  - Computer science, introduction to, 7–15
    - algorithms, 11–12
    - computers, 8–9
    - ease of implementation, 15
    - parallelizability, 14–15
    - programs, 9–11
    - running time analysis, 12–13
    - space complexity analysis, 13–14
  - Conditional independency, 246, 247, 248, 252
  - Conditional probability, 18, 20, 68–69
  - Conditional probability distribution, 74, 246–248
  - Conditional probability table (CPT), 248, 252
  - Conditioned random fields algorithm, 298, 298t
  - Confounding factor, 67
  - Connectivity graph, 258
  - Connectivity Map, 269, 270
  - Constant-time algorithm, 12–13
  - Constant-time storage, 14
  - Consumer health informatics, 287
  - Contigs, 169
  - Contingency table, 31–33, 32t–33t, 129
  - Continuous data
    - mixing categorical and continuous data, 33b–34b
    - statistical tests on, 28–30
  - Continuous feature, 65
  - Contrast, 91, 91f
  - Controls, in GWAS, 128–129
  - Copy-number variants (CNVs), 142, 180, 182
  - Correction methods
    - Benjamini–Hochberg false discovery rate, 36, 121
    - Bonferroni, 35
    - Tukey, 36
  - Correlation, 39–43
    - covariance, 40
    - definitions, 128
    - described, 39–40
    - Pearson, 40–41, 42f
    - Spearman rank, 41, 42f
  - Correlation coefficient, 101–102
  - Cost function, 101
  - Counting rules, 19
  - Covariance, 40
  - Covariate analysis, 149
  - Cox proportional hazard models, 42
  - C programming language, 11, 15
  - C++ programming language, 9, 11
  - CPT (conditional probability table), 248, 252
  - CPU (central processing unit), 8, 8t
  - Cross-validation, 58–61, 59f
    - feature selection and, 61, 66
    - k*-fold, 59f, 60
    - leave-one-out, 60
  - CT (computed tomography), 88, 89f
  - cTAKES, 291t
  - Cubic algorithm, 13
  - Cubic space complexity, 14
  - Cufflinks, 170–171, 171t
  - Current procedural terminology, 293
  - Curse of dimensionality, 206
  - Cyanine 3, 109–111
  - Cyanine 5, 109–111
  - Cytobank, 203, 205
- ### D
- DAGs (directed acyclic graphs), 244
  - Data
    - categorical, 31–33, 33b–34b
    - censored, 41
    - clean, 64
    - learning biomolecular pathways from, 241–259
    - missing, 41–43
    - mixing categorical and continuous data, 33b–34b
    - output variables, 66–67
    - probability and, 17–21
    - quality, 64
    - statistical analysis of, 25–39
      - tests on categorical data, 31–33
      - tests on continuous data, 28–30
    - using machine learning algorithms, 62–67
    - visualization, 43–44
      - Anscombe’s plot, 41, 42f, 43
      - draftsman’s plot, 43, 44f
  - Databases, 261, 270
  - Data integration, 287
    - gene expression experiments, 261–281
    - investigative steps, 262f
    - paradigms
      - integrating expression data over “unrelated” contexts, 267f, 269–275, 270f
      - integrating expression data with other genome-wide modalities, 267f, 275–278, 276f
      - meta-analysis of gene expression data, 266, 267f, 268–269, 268f
    - programming exercise, 278–281
      - finding the data, 278
      - formulating a question, 278
      - integrating findings, 279
      - interpreting findings, 279
      - programming solution, 280
      - representation of differential gene expression, 278–279
    - question formulation, 262–263, 278
    - representation of differential gene expression data, 263–265, 264f, 278–279
  - Data representation, 263
  - Data snooping, 34–35
  - Data transforms, 24
  - DAVID, 265
  - DBNs (dynamic Bayesian networks), 244–245

- Decision tree, 71–72, 71f  
Degrees of freedom, 30  
De novo assembly, 169  
De novo sequencing, peptide identification and, 194–195, 195t  
Descriptive statistics, 21–24  
Differential equation models, 257b–258b  
Differential gene expression, representation of data, 263–265, 264f, 278–279  
Dimensionality, 91–92  
Dimensionality reduction, 78  
Directed acyclic graphs (DAGs), 244  
Discrete feature, 64–65  
Discrete ordered features, 64–65  
Dispersion, statistics of, 22  
Distributed systems, programs for, 9  
Distribution  
    conditional probability, 74, 246–248  
    F, 30  
    hypergeometric, 31, 32f  
    joint probability, 74, 245–248  
    null, 121, 140  
    skewness of, 24, 24f  
    symmetry of, 24, 24f  
    *t*, 28–30, 28f, 30f  
Divisive clustering, 77  
DNA-binding proteins, 172  
DNA sequencing  
    next-generation sequencing, 155–184  
        ABI SOLiD, 156t, 157, 157f  
        alignment, 157–158  
        BLAST use, 158–159  
        CHIP-seq, 172–179  
        454 FLX, 156, 156t  
        future of, 184  
        gene expression microarrays compared, 107  
        Illumina, 156, 156t  
        RNA-seq, 167–172  
        sequencing services, 183–184  
        short-read mapping, 159–167  
        variation detection, 180–183  
    1000 Genomes Project, 142  
Document classification, 286–287  
Draftsman's plot, 43, 44f  
Drug discovery, 287  
DrugNer, 291t  
Dye bias, 110–111  
Dynamic Bayesian networks (DBNs), 244–245  
Dynamic contrast ratio, of human eye, 83  
Dynamic susceptibility contrast perfusion imaging (DSC-MRI), 89
- E**
- Earley algorithm, 290  
Edge detection  
    Canny, 104b–105b  
    described, 103  
ELANDv2, 166, 166t  
Electronic medical records (EMRs), 287  
Electron microscope, 87–88, 88f  
*The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Hastie, Tibshirani, and Friedman), 79  
“Else if” statement, 10  
EM (expectation–maximization) algorithm, 119, 212, 214  
EMBL (European Molecular Biology Library), 265  
Emission probabilities, 75  
Encyclopedia of DNA Elements (ENCODE), 150  
Enrichment, identification of regions of, 175–176  
ENTREZ, 268, 270, 276, 277  
Entropy, 102  
Equivalence classes, 253b  
ERANGE, 170–171, 171t  
ERK, 246–252, 247f, 253b–254b  
Error  
    family-wise error rate, 35  
    in multiple hypothesis testing, 35  
    root mean square, 57  
    testing, 58–59, 63–64, 63f  
    training, 58–59, 63–64, 63f  
    type 1, 35  
    type 2, 35  
Estimate, biased, 23  
Euclidean distance, 68, 114–115, 115f  
European Molecular Biology Library (EMBL), 265  
Executable, 8  
Exomes, 142  
Expectation, 21  
Expectation–maximization (EM) algorithm, 119, 212, 214  
Experimental design, 57–61  
Exponential-time algorithm, 13  
Expression arrays, 107–122  
Expression data, 107–124  
Eye, human, 83  
Eye color  
    genetic association, 128–130, 128f  
    single-nucleotide polymorphisms (SNPs), 126, 127f
- F**
- False discovery, 34  
False discovery rate (FDR), 120–122  
    Benjamini–Hochberg, 36, 121  
    ChIP-seq, 177–178  
    expression data, 268, 273–275  
False negative, 54, 55, 55f  
False positive, 53, 55–56, 55f  
False positive rate, 35  
Family-wise error rate (FWER), 35, 120–121  
.fcs files, 202–204, 202f–204f, 202t  
F distribution, 30

## 310 Index

- Feature(s)
    - Boolean, 64–65
    - continuous, 65
    - discrete, 64–65
    - discrete ordered, 64–65
    - interactions between, 70
    - number of, 62, 65
    - ordinal, 64
    - sample, 49–50
  - Feature selection, 52, 61, 65–66
  - Feature space, 49
    - classification in 2D, 69–70, 69f
    - dimensional reduction and, 78
    - effective dimensionality of, 62
    - as joint probability distribution, 74
    - principal component analysis and, 78
  - Fisher, Ronald, 29
  - Fisher's exact test, 31–33, 141
  - Fisher's method, 268, 268f
  - Flow cytometry, 188, 200–218
    - analyzing, 205–206
    - background of, 200–201, 201f
    - comparing across samples, 216–218
      - informative event problem, 217–218
      - quantitative difference problem, 217
      - sample classification problem, 216–217
    - data visualization of flow data, 203–205, 203f
    - exploratory analysis, 218
    - .fcs files, 202–204, 202f–204f, 202t
    - future directions, 218–219
      - clinical applications, 219
      - data variability, 218–219
      - structured annotation for data sharing, 219
    - learning signaling pathway structure from, 256
    - preprocessing steps, 206–208, 206f–207f
      - compensation, 206–207, 206f
      - transformation of data, 207, 207f
    - probability distribution, 213b–214b
    - states, 201
    - subpopulation-finding and feature extraction
      - methods, 208–216
      - binning, 209–210
      - cluster analysis, 211–215
      - heatmaps, 209, 209f
      - histograms, 209–210, 210f
      - mixture models, 212–215
      - nonparametric population-finding methods, 215–216
      - 1D methods, 208–210
      - 2D methods, 210–211
  - Fluorescence microscope, 87–88, 88f
  - Fluorophores, 109–111
  - fMRI (functional MRI), 89
  - Focus, 83, 84f, 91
  - Fold enrichment, 176
  - Fork, 252
  - For-loops, 10
  - Frequentist approach, to statistical hypothesis testing, 25
  - Friedman, Jerome, 79
  - Friedman, Nir, 80
  - F statistic, 27
  - Function, 9–10
  - Functional class scoring approaches, 227–228
    - assessing statistical significance of pathways, 232
    - limitations, 228
    - overview, 227–228
    - tools for, 231t
  - Functional MRI (fMRI), 89
  - FWER (family-wise error rate), 35, 120–121
- ## G
- Gating, 205, 208
  - Gaussian smoothing, 104b
  - Gene Association Database (GAD), 277
  - GeneChip, 180
  - Gene expression
    - analysis of values, 114–122
      - metrics, 114–115, 115f
      - semisupervised clustering methods, 117–120, 119f
      - statistical approaches to data interpretation, 120–122, 121t
      - unsupervised clustering methods, 115–117, 116f–117f
    - machine learning and analysis, 49–51, 52f
    - meta-analysis and data integration, 261–281
    - microarrays, 107–122
      - analysis of gene expression values, 114–122
      - one-color, 108, 112–114
      - overview of, 107–109
      - two-color, 108, 109–111, 111b, 111f
  - Gene Expression Omnibus (GEO), 261, 262, 264, 269, 277
  - Gene-level statistics, 230, 231t–232t
  - Gene Ontology (GO), 292–293, 294t
    - incomplete and inaccurate annotations, 233–234, 235f
  - Generalization, 58
  - Genes, number of human, 130
  - Gene set association (GSA), 271, 272–273
  - Gene Set Enrichment Analysis (GSEA), 269–271
  - Genetic association, 128–130, 128f
    - interpreting, 144–145
    - testing, 137–141
      - $\chi^2$  test of statistical independence, 139–140
      - improving statistical power, 141–144
  - Genetic heterogeneity, 127
  - Genome sequencing, 142
  - Genome-wide association study (GWAS), 125–150
    - data quality, 145–146
    - fundamental concepts underlying, 126–130
    - goal of, 126
    - integrating expression data with, 276–277
    - rationale for, 130–135
      - development of GWAS as research tool, 131–132
      - linkage disequilibrium, 132–133, 134f
      - linkage studies different from, 133–135
      - what can be learned from, 130–131

significance criterion for, 145  
steps in, 135–150  
    association testing, 137–141, 138f  
    causal genetic factor, 149–150  
    genotype calling, 135–137, 136f  
    improving statistical power, 141–144, 143f  
    interpreting genetic associations, 144–145  
    population stratification, 146–149, 148f  
Genome-wide significance, 140–141  
Genomic inflation factor, 148–149  
Genomic variants, 180–183  
Genotype calling, 135–137, 136f  
Genotype imputation, 142–144, 143f  
Genotype–phenotype association, 128–130, 128f  
Genotypes  
    association with phenotypes (*see* Genome-wide association study)  
    described, 126, 127f  
    genotype–phenotype, 128–130, 128f  
Genotypic odds ratio, in GWAS, 129–130  
GEOquery, 266, 271–272  
GO. *See* Gene Ontology  
Golub, Todd, 51  
Gosset, William S., 29  
Grayscale image, 86–87, 87f  
    of a cell, 92–93, 92f  
Greedy random search, 250  
GSA (gene set association), 271, 272–273  
GSEA (Gene Set Enrichment Analysis), 269–271  
GWAS. *See* Genome-wide association study

## H

Haplotype blocks, 142  
Haplotype phase inference, 142  
Haplotypes, 142  
HapMap, 131–132  
Hard drive, 8, 8t, 14  
Hardy–Weinberg equilibrium, 137  
Hastie, Trevor, 79  
Health services delivery, 287  
Heatmaps, 209, 209f  
Heavy water, 198  
Heterozygous, 136  
Heuristic search, 250  
Hidden Markov model (HMM), 73f, 74–76, 298, 298t  
Hidden variables, 254  
Hierarchical clustering, 76f, 77, 116–118, 116f–117f  
Histograms, flow cytometry, 209–210, 210f  
HITECH Act, 287  
HomoloGene, 265  
Homozygous, 136  
Hoover Tower, photographs of, 83, 84f–85f, 87f  
Human Protein Reference Database, 276  
Hyperbolic arcsine transform, 207, 207f  
Hypergeometric distribution, 31, 32f  
Hypergeometric experiment, 31  
Hypergeometric test, 271

Hypothesis  
    alternative, 26  
    null, 26, 27  
    testing of, 27  
Hypothesis testing. *See also* Statistical hypothesis testing  
    described, 27  
    multiple, 140, 145  
Hysteresis thresholding, 105b

## I

ICAT (isotope coded affinity tags), 198  
ICD-10CM (international classification of diseases), 293, 294t  
ICP (iterative closest point), 102–103  
Identical-read stacks, 176  
“If” statement, 10  
Illumina, 112–113, 133  
    CASAVA, 182  
    sequencing, 156, 156t  
Image analysis, 83–106  
    biomedical images, 87–90  
    computational, 92–103  
    generating images for, 90–92  
    imaging basics, 83–87  
Image registration, 99–101  
    algorithms, 101–103  
        feature-based algorithm, 102–103  
        intensity-based registration, 101–102  
        mutual information theoretic technique, 102  
    multiple images, 99–100  
    spatial transformation, 100–101  
Images  
    biomedical, 87–90  
        computed tomography (CT), 88, 89f  
        magnetic resonance imaging (MRI), 89, 89f  
        microscope images, 87–88, 88f  
        positron emission tomography (PET), 90  
    bits in, 86b  
    dynamic, 100  
    imaging basics, 83–87  
    intermodality, 99–100  
    intramodality, 99–100  
    multiple, 99–100  
    serial, 100  
Image segmentation  
    defined, 93  
    *k*-means clustering, 95–96, 96f  
    Otsu’s method for, 94–95, 94f–95f  
Imputation, 43, 65, 272  
Indel, 142  
Indel alignment, 163–164  
Independent events, 19  
Indexed color, 90, 91f  
Influence, 244  
Informatics anxiety, 1  
Input, of the classifier, 50



## 312 Index

- Input/output (I/O) devices, 8, 8t  
Intensity  
  computation of image intensity gradient, 104b  
  cutoff, 93–94, 93f  
  Otsu's method for image segmentation, 94–95, 94f–95f  
Intensity-based registration, 101–102  
Intensity plot, 136–137, 136f  
Intermodality images, 99–100  
International classification of diseases (ICD-10CM), 293, 294t  
International HapMap Project, 131–132  
International Society for Advancement of Cytometry (ISAC), 202  
Interquartile range, 22, 24  
Interventional data, 244, 253b–254b, 255  
Intramodality images, 99–100  
ISAC (International Society for Advancement of Cytometry), 202  
Iterative closest point (ICP), 102–103  
iTRAQ (isobaric tags for relative and absolute quantitation), 198
- ### J
- Jackknife technique, 38–39  
Java programming language, 9, 11  
Joint probability, 19–20  
Joint probability distribution, 74, 245–248
- ### K
- KEGG (Kyoto Encyclopedia of Genes and Genomes), 226, 233, 235, 236  
Kernel smoothing, 174  
Kernel trick, 70  
*k*-fold cross-validation, 59f, 60  
Klein, Dan, 80  
*k*-means  
  brain images, 97, 97f  
  cell images, 95–96, 96f  
  clustering, 77–78  
  flow cytometry, 212–215  
  semisupervised clustering, 117–120, 119f  
*k*-nearest neighbors algorithm, 67–68, 298t  
Knome, 183  
Koller, Daphne, 80  
Kolmogorov–Smirnov test, 217  
Kruskal–Wallis test, 30  
Kurtosis, 22  
Kyoto Encyclopedia of Genes and Genomes (KEGG), 226, 233, 235, 236
- ### L
- Language processing. *See* Natural language processing  
*LCT* gene, 147  
Leave-one-out cross-validation, 60  
Level set algorithm, 98–99, 99f–100f  
Libraries, 9  
Life Technologies, 184  
Light microscope, 87–88, 88f  
Likelihood ratio test, 141  
Linear classification algorithms, 69–70  
Linear time algorithm, 13  
Linear time storage, 14  
Linear transform, 207, 207f  
LingPipe, 291t  
Linkage analysis, 133–135  
Linkage criteria, 77  
Linkage disequilibrium, 132–133, 134f  
Linkage equilibrium, 133  
Local maximum, 251  
Location, statistics of, 22  
Locus  
  defined, 126  
  linkage analysis, 133–135  
Logistic regression, 70  
LOINC, 293, 294t  
Lowess normalization method, 110–111, 111f
- ### M
- Machine learning, 47–80  
  algorithm use, 62–78  
  data, 62–67  
  probabilistic models, 72–76, 73f  
  supervised learning algorithms, 67–72  
  unsupervised learning algorithm, 76–78, 76f  
  defined, 47  
  experimental design, 57–61, 59f  
  performance evaluation, 53–57, 55f, 56f  
  resources, 79–80  
  terminology, 49–53  
MACS, 178, 179f  
Magnetic resonance imaging (MRI)  
  brain images, 96–100, 97f–98f, 100f  
  described, 89, 89f  
  functional, 89  
  perfusion, 89, 96–97  
  structural, 89  
Manhattan distance, 67–68, 114  
Manhattan plots, 146, 147f  
Mann–Whitney U test, 30  
*MA* plots, 111, 111b  
Mapping  
  next-generation sequencing, 158–167  
  short-read, 159–167  
  alignment programs, 166–167, 166t  
  characteristics of short reads, 159–160, 160f  
  indel alignment, 163–164  
  mapping output, 165–166  
  mapping quality/posterior probability, 162–163, 162f  
  paired-end alignment, 164–165, 165f  
  practical considerations, 166–167



- quality score use in mapping, 163
- repetitive reads, 161
- scoring and filtering, 161–162
- seeding, 160–161
- Marginal likelihood, 249
- Markov assumptions, 247
- Mascot, 192
- Mass differential equation model, 257b
- Mass spectrometry (MS), 188–199
  - overview of, 188–189, 189f
  - peptide identification, 191–194
    - accurate mass and time tag (AMT) approach, 193–194
    - database-driven approaches, 191–193
    - de novo sequencing, 194–195, 195t
    - estimating false positives, 193
    - target-decoy approach, 193
  - peptide quantitation, 196–199
    - labeled, 197–198, 198f
    - label-free, 196–197
    - selected reaction monitoring (SRM), 198–199, 198f
  - protein digestion for, 190
  - protein identification, 195–196
    - false positives, 195–196
    - one peptide mapped to many proteins, 196
  - protein quantitation, 199
  - sample preparation, 189–190
  - spectra example, 191, 191f
  - tandem (MS/MS), 190–191, 191f
- Mass-to-charge ( $m/z$ ) ratio, 190
- Mating, assortative, 137
- Matlab machine learning algorithms, 79
- MATLAB programming language, 9, 11
- Maximum, 22
- Maximum entropy algorithm, 298, 298t
- Mean
  - central limit theorem and, 29
  - comparing means between groups, 28–30, 28f
  - population, 22, 23
  - sample, 21, 22
  - standard error of, 24
  - in unimodal, symmetric distribution, 24
- Measures of central tendency, 22
- Median
  - sample, 21, 22
  - in unimodal, symmetric distribution, 24
- Median Polish summation, 113
- Medical dictionary for regulatory activities (MedDRA), 294t
- Medical subject headings (MeSH), 293, 294t
- MedLEE, 291t
- Medline, 285, 286f, 288
- MEK, 242f, 243, 246–252, 247f, 253b–254b
- Memory
  - defined, 8, 8t
  - space complexity and, 13–14
- Meta-analysis
  - described, 144
  - gene expression experiments, 261–281
    - for increasing statistical power, 144
- MetaMap, 291t
- Metathesaurus, 292–293, 294t
- Metrics, microarray analysis, 114–115, 115f
- Microarrays, 107–122
  - analysis of gene expression values, 114–122
    - metrics, 114–115, 115f
    - semisupervised clustering methods, 117–120, 119f
    - statistical approaches to data interpretation, 120–122, 121t
    - unsupervised clustering methods, 115–117, 116f–117f
  - blocks, 109
  - high-density DNA, 132
  - next-generation sequencing compared, 107
  - one-color
    - overview, 112
    - preprocessing and normalization, 112–114
    - two-color compared, 108
  - overview of, 107–109
  - tiling, 107
  - two-color
    - MA plots, 111, 111b
    - one-color compared, 108
    - overview, 109–110
    - preprocessing and normalization, 110–111, 111f
- Microscope images, 87–88, 88f
- Minimum, 22
- Minkowski equation, 114
- Mismatch probes, 112
- Missing data, 41
- Mixture models, flow cytometry and, 212–215
- Model averaging, 251
- Modeling assumptions, 242
- Models
  - pathway
    - Bayesian networks, 245–259
    - Boolean networks, 242–244, 242f, 245
    - challenges in, 241
    - differential equation models, 257b–258b
    - dynamic Bayesian networks, 244–245
    - modeling assumptions, 242
    - network inference, 241
    - robust, 244
  - probabilistic, 72–76, 73f, 241–242
    - Bayesian network, 73f, 74–76
    - hidden Markov models (HMM), 73f, 74–76, 298, 298t
    - proportional hazard, 42
- MoDIL, 183
- Module networks, 255–256
- Monty Hall problem, 18
- Moore, Andrew, 80
- Moore, Gordon, 47
- Moore's law, 47–48
- MOSAIC, 166t, 167
- MRI. *See* Magnetic resonance imaging

## 314 Index

- MS. *See* Mass spectrometry
- Multiple hypotheses, correction for, 21, 35–36, 230t–232t, 232–233
- Multiple hypothesis testing, 140, 145
- correction methods, 35–36
    - Benjamini–Hochberg false discovery rate, 36
    - Bonferroni, 35
    - Tukey, 36
  - errors, 35
  - problems with, 34–35
- Multiple testing, 140, 145
- Mutual information theoretic technique, 102
- ### N
- Naive Bayes algorithm, 68–69, 298, 298t
- Named entity recognition (NER), 290–291, 291t
- National Cancer Institute (NCI)
- Enterprise Vocabulary Services, 292
  - Thesaurus and Metathesaurus, 292–293, 294t
- National Center for Biotechnology Information (NCBI), 265
- Natural language processing, 285–301
- annotation, 292
  - applications, 286–287
  - machine learning, 297–300, 298t–300t
  - named entity recognition, 290–291, 291t
  - ontologies in biomedicine, 295–297, 296f
  - preprocessing raw text, 288–290, 288f–289f
  - standard terminologies, 292–295, 294t–295t
- NCBI (National Center for Biotechnology Information), 265
- NCI. *See* National Cancer Institute
- Negative predictive value (NPV), 54, 55f
- NER (named entity recognition), 290–291, 291t
- Network device, 8, 8t
- Network inference, 241
- Networks
- Bayesian, 73–74, 73f, 245–259
    - in action, 255–256
    - chain rule, 247
    - joint probability distribution, 245–248
    - learning signaling pathway structure from flow cytometry data, 256
  - Markov assumptions, 247
  - model properties, 251–255
  - model semantics, 245–246
  - notation, 246–249
  - structure learning, 249–251
- Boolean, 242–244, 242f, 245
- dynamic Bayesian networks, 244–245
- module, 255–256
- Neurosphere cells, 116, 117f
- Next-generation sequencing, 155–184
- ABI SOLiD, 156t, 157, 157f
  - alignment, 157–158
  - BLAST use, 158–159
  - CHIP-seq, 172–179
    - advantages, 178
    - algorithms, 173–177
    - features of software packages, 179f
    - overview, 172–173
    - practical considerations, 177–178
  - 454 FLX, 156, 156t
  - future of, 184
  - gene expression microarrays compared, 107
  - Illumina, 156, 156t
  - RNA-seq, 167–172
    - advantages, 167–168, 171–172
    - applications, 167–168
    - approaches to identifying transcript structure, 168–170, 168f
    - overview, 167–168
    - transcript quantification, 170–171, 171f
  - sequencing services, 183–184
  - short-read mapping, 159–167
    - alignment programs, 166–167, 166t
    - characteristics of short reads, 159–160, 160f
    - indel alignment, 163–164
    - mapping output, 165–166
    - mapping quality/posterior probability, 162–163, 162f
    - paired-end alignment, 164–165, 165f
    - practical considerations, 166–167
    - quality score use in mapping, 163
    - repetitive reads, 161
    - scoring and filtering, 161–162
    - seeding, 160–161
  - variation detection, 180–183
    - copy-number variants, 180, 182
    - detecting large-scale variants, 182–183, 183f
    - detecting nucleotide-level variation, 180–182, 181f
    - genomic variants classified by scale, 180
- Ng, Andrew, 80
- Nimblegen, 181
- Noise, 64
- Nominal significance threshold, 140
- Nonmaximum suppression, 105b
- Nonparametric population-finding methods, 215–216
- Nonparametric statistics, 30, 41
- Normalization
- one-color microarrays, 112–114
  - two-color microarrays, 110–111, 111f
- NPV (negative predictive value), 54, 55f
- Null distribution, 121, 140
- Null hypothesis
- described, 26
  - test statistic and, 27
- ### O
- Oases, 171, 171t
- Object-oriented languages, 11
- Octave, 79
- Odds ratio, in GWAS, 129–130
- Odds ratios, 37b–38b

- ODEs (ordinary differential equations), 257, 257b–258b
- OMSA, 192
- One-color microarrays
  - overview, 112
  - preprocessing and normalization, 112–114
  - two-color compared, 108
- One-strand peaks, 176
- Online mendelian inheritance in man (OMIM), 294t
- Ontologies in biomedicine, 295–297, 296f
- ORA. *See* Overrepresentation analysis
- Ordinal features, 64
- Ordinary differential equations (ODEs), 257, 257b–258b
- Otsu's method for image segmentation, 94–95, 94f–95f
- Outliers, 22, 24
- Output, of the classifier, 50
- Output variables, 66–67
- Overfitting, 60, 62–64, 63f, 249
- Overlapping reads, 169
- Overrepresentation analysis (ORA), 224–227
  - assessing statistical significance of pathways, 232
  - correction for multiple hypotheses, 230t, 232
  - limitations of, 226–227
  - overview, 224–226, 225f
  - tools for, 230t
- P**
- Pacific Biosciences, 184
- Paired distance, 164–165
- Paired-end alignment, 164–165, 165f
- Pairs of reads, 180
- Palette indexing, 90, 91f
- PAM (prediction across microarrays), 217
- Paradigm, 263
- Parallelizability, of algorithms, 14–15
- Parameters, 250b
- Parsing, 288f–289f, 289–290
- Partial differential equations, 257, 257b–258b
- Partitioning algorithms, 71–72, 118–119
- Part of speech tagging, 288–289, 288f
- PathBLAST, 276
- Pathway, defined, 223–224
- Pathway analysis
  - comparison of existing tools, 229–233, 230t–232t
    - assessing statistical significance of pathways, 232
    - correction for multiple hypotheses, 230t–232t, 232–233
    - gene-level statistics, 230, 231t–232t
    - pathway-level statistics, 230, 232
  - current challenges in, 233–236
    - inability to model and analyze dynamic response, 236
    - inability to model effects of external stimulus, 236
    - incomplete and inaccurate annotations, 233–234, 235f
    - low-resolution knowledge bases, 233, 234f
    - missing condition- and cell-specific information, 234–236
    - weak interpathway links, 236
  - functional class scoring approaches, 227–228
    - assessing statistical significance of pathways, 232
    - correction for multiple hypotheses, 230t, 232
    - limitations, 228
    - overview, 227–228
    - tools for, 231t
  - knowledge base-driven, 223–238
  - overrepresentation analysis (ORA), 224–227
    - assessing statistical significance of pathways, 232
    - correction for multiple hypotheses, 230t, 232
    - limitations of, 226–227
    - overview, 224–226, 225f
    - tools for, 230t
  - pathway-topology-based approaches, 228–229, 232t
  - utility and confidence of, 236–237
- Pathway-level statistics, 230, 232
- Pathway models
  - Bayesian networks, 245–259
  - Boolean networks, 242–244, 242f, 245
  - challenges in, 241
  - differential equation models, 257b–258b
  - dynamic Bayesian networks, 244–245
  - modeling assumptions, 242
  - network inference, 241
  - robust, 244
- Pathway-topology-based approaches, 228–229
- Pattern Recognition and Machine Learning* (Bishop), 79
- PCA (principal component analysis), 78
- PCR, quantitative (qPCR), 172–173
- Peak finding, 174
- Pearson, Karl, 29
- Pearson correlation, 40–41, 42f, 114–115, 115f
- PEMer, 183
- Peptide identification, by mass spectrometry (MS), 191–194
  - accurate mass and time tag (AMT) approach, 193–194
  - database-driven approaches, 191–193
  - de novo sequencing, 194–195, 195t
  - estimating false positives, 193
  - peptide modifications, 95
  - target-decoy approach, 193
- Peptide modifications, mass spectrometry and, 95
- Peptide quantitation, mass spectrometry and, 196–199
  - labeled, 197–198, 198f
  - label-free, 196–197
  - selected reaction monitoring (SRM), 198–199, 198f
- Percentile, 22
- Perfect-match probes, 112
- Performance evaluation, 53–57, 55f, 56f
- Perfusion MRI, 89, 96–97
- Perl programming language, 11
- Permutation testing, 39, 141
- Personalized medicine, 131

## 316 Index

- Perturbation factor, 229  
PET (positron emission tomography), 90  
pFDR, 121  
Pharmacogenetics, 131  
Phenotypes  
  association with genotypes (*see* Genome-wide association study)  
  defined, 127–128  
  genotype-phenotype, 128–130, 128f  
Pindel, 183  
Pixel  
  bits per pixel (bpp), 85  
  defined, 84  
  pixelation, 85f  
  volumetric, 91–92  
PolyPhen, 150  
Population mean, 22  
Population stratification, 146–149, 148f  
Population variance, 23  
Positive predictive value (PPV), 54, 56  
Positron emission tomography (PET), 90  
Posterior probability, 68, 162  
Power, statistical, 127–128, 141–144  
PPV (positive predictive value), 54, 56  
*Practical Flow Cytometry* (Shapiro), 200  
Precision, 27, 54, 55f  
Prediction across microarrays (PAM), 217  
Pred probability, 159  
Pred score, 159  
Preprocessing raw text, 288–290, 288f–289f  
  chunking and parsing, 288f–289f, 289–290  
  part of speech tagging, 288–289, 288f  
  stemming, 290  
  stop word removal, 290  
  tokenization, 288, 288f  
Primer extension, 135  
Principal component analysis (PCA), 78  
*Probabilistic Graphical Models: Principles and Techniques* (Koller and Friedman), 80  
Probabilistic modeling, defined, 72  
Probabilistic models, 72–76, 73f, 241–242  
  Bayesian network, 73f, 74–76  
  hidden Markov models (HMM), 73f, 74–76, 298, 298t  
Probability  
  Bayes' Rule, 20–21  
  conditional, 18, 20, 68  
  described, 17–21  
  emission, 75  
  expectation, 21  
  joint, 19–20  
  *Monty Hall* problem, 18  
  notation, 18  
  posterior, 68, 162  
  transition, 75  
Probability distribution, 213b–214b  
Probes, for genotype calling, 135–136  
Processors, 8, 8t, 9  
Programming language  
  choosing best suited, 15  
  described, 9, 11  
  effect on running time, 12  
  object-oriented, 11  
Programs  
  control flow in, 10  
  described, 9–11  
  for distributed systems, 9  
  ease of implementation, 15  
  executing, 8–9  
  functions and, 9  
  variables and, 10  
Proof-of-principle, 2  
Proportional hazard models, 42  
Prostate cancer, meta-analysis of, 266, 268, 268f  
Protein identification, by mass spectrometry (MS), 195–196  
  false positives, 195–196  
  one peptide mapped to many proteins, 196  
Proteins  
  mass spectrometry (MS) of  
    peptide identification, 191–194  
    peptide quantitation, 196–199  
    protein identification, 195–196  
    protein quantitation, 199  
  variations in, 188  
Proteomics, 187–219  
  flow cytometry, 188–218  
  mass spectrometry (MS), 188–199  
  reasons for studying, 187  
Proton, in MRI, 89  
PubMed, 288  
*p* value  
  interpretation of, 27  
  multiple-testing correction, 35–36  
  significant, 25  
  *t* distribution and, 29  
Pyrophosphates, 156  
Pyrosequencing, 156  
Python programming language, 9, 11, 15, 79
- ## Q
- QPALMA, 169, 170, 171t  
qPCR (quantitative PCR), 172–173  
qRT-PCR (quantitative real-time-polymerase chain reaction), 108  
Quadratic algorithm, 13  
Quality score, 159, 162–163, 162f  
Quantile normalization, 113  
Quantile-quantile (QQ) plot, 148, 148f  
Quantitative PCR (qPCR), 172–173  
Quantitative real-time-polymerase chain reaction (qRT-PCR), 108  
QuEST, 178, 179f  
Question, formulation of, 25–26  
*q* value, 36–37

## R

Raf, 242f, 243–244, 246–252, 247f, 253b–254b  
Random forests algorithm, 298, 298t  
Random start, 251  
Rank Product (RP) method, 122  
Read length, 155–156, 156t  
Reads  
  overlapping, 169  
  pairs of reads, 180  
  short-read mapping, 159–167  
Recall, 54  
Receiver operating characteristic curves (ROC),  
  56–57, 56f  
Recombination  
  linkage disequilibrium and, 133  
  number per meiosis, 133  
Reference panel, 143  
Reference transcriptome, 168  
Region finding, 174  
Regression, example of, 63, 63f  
Regression task, 50, 57  
Regression tree, 71  
Regulation of actin cytoskeleton pathway, 235  
Representation of differential gene expression data,  
  263–265, 264f, 278–279  
Reproducibility, statistical hypothesis testing and, 25  
Resampling methods  
  bootstrapping, 38  
  jackknifing, 38–39  
  permutation testing, 39  
Resolution, 84, 84f, 90  
Reusability, in computer science, 11  
RGB image, 85, 87f  
RMA (Robust Multiarray Analysis) model, 113–114  
RNA-seq, 167–172  
  advantages, 167–168, 171–172  
  applications, 167–168  
  approaches to identifying transcript structure,  
    168–170, 168f  
  with reference genome, 168–169, 168f  
  without reference genome, 169–170  
  overview, 167–168  
  transcript quantification, 170–171, 171f  
Robust, 244  
Robust Multiarray Analysis (RMA) model, 113–114  
ROC (receiver operating characteristic curves),  
  56–57, 56f  
Root mean square error, 57  
Root nodes, 246  
RPKM, 170  
RP (Rank Product) method, 122  
R programming language, 9, 11, 113, 203, 261, 265  
  Bioconductor, 113, 265, 266  
  GEOquery, 266, 271–272  
  machine language algorithms, 79  
  programming exercise, 278–281  
  finding the data, 278

  formulating a question, 278  
  integrating findings, 279  
  interpreting findings, 279  
  programming solution, 280  
  representation of differential gene expression,  
    278–279

Running time analysis of the algorithm, 12–13  
RxNORM, 293, 294t

## S

SAM (Significance Analysis of Microarrays),  
  120–122, 270, 270f, 277  
SAM file format, 165–166  
Sample mean, 21, 22  
Samples, independent, 64  
SAMtools, 166, 182  
Sanger, Fred, 155  
Sanger sequencing, 155  
Scikit-learn, 79  
Scoring, Bayesian, 249–250, 250b  
Search  
  greedy random, 250  
  heuristic, 250  
Seed, 160  
Seeding, 160–161  
Seed matches, 161  
Segmentation. *See* Image segmentation  
Selected reaction monitoring (SRM),  
  198–199, 198f  
Self-organizing maps (SOMs), 120  
Self-self hybridization, 110  
Semisupervised clustering methods, 117–120, 119f  
Semisupervised learning, 51, 52f  
Sensitivity, 54–57, 55f, 56f, 244  
Sequencers, DNA, 155  
SEQUEST algorithm, 191–192, 192b  
Sex chromosomes, 126  
Shifting, ChIP-seq reads, 174–175, 175f  
SHOGUN Machine Learning Toolbox, 79  
Short Oligonucleotide Analysis Package (SOAP),  
  166, 166t  
Short-read mapping, 159–167  
  alignment programs, 166–167, 166t  
  characteristics of short reads, 159–160, 160f  
  indel alignment, 163–164  
  mapping output, 165–166  
  mapping quality/posterior probability,  
    162–163, 162f  
  paired-end alignment, 164–165, 165f  
  practical considerations, 166–167  
  quality score use in mapping, 163  
  repetitive reads, 161  
  scoring and filtering, 161–162  
  seeding, 160–161  
SIFT (sorting intolerant from tolerant), 150  
Signaling pathway, learning structure from flow  
  cytometry data, 256

## 318 Index

- Signal shifting, 174–175, 175f
  - Significance
    - error and, 35
    - experiment-wide level, 35
  - Significance Analysis of Microarrays (SAM), 120–122, 270, 270f, 277
  - Significant, 25
  - SILAC (stable isotope labeling by amino acids in cell culture), 198
  - Silhouette plot, 118–119, 119f
  - Single-nucleotide polymorphisms (SNPs)
    - in cluster plot, 136–137, 136f
    - cost of SNP genotyping, 132
    - defined, 126
    - eye color and, 126, 127f
    - genotype imputation, 142–143, 143f
    - GWAS, 130–137, 134f, 144–146
    - interpreting genetic associations, 144–145
    - linkage disequilibrium and, 132–133, 134f
    - number in human genome, 132
    - small effect sizes, 131
    - tag, 133
  - Skew, 22, 24, 24f
  - Sliding window, 173–174
  - Smoothing, 173–174, 173f
  - SNOMED-CT (systematized nomenclature of medical terminologies-clinical terms), 293, 294t
  - SNP caller, 182
  - SNP genotyping, 180
  - SNPs. *See* Single-nucleotide polymorphisms
  - SNVMix, 182
  - SOAP (Short Oligonucleotide Analysis Package), 166, 166t
  - SOMs (self-organizing maps), 120
  - Space complexity analysis, 13–14
  - SPADE, 218
  - Spatial transformation, 100–101
  - Spearman rank correlation, 41, 42f, 114–115
  - Specificity, 54–57, 55f, 56f
  - Spectral resolution, 206
  - Split-read alignment, 168–169, 168f
  - Square errors criterion, 118
  - SRM (selected reaction monitoring), 198–199, 198f
  - SSAHA2 (sequence search and alignment by hashing algorithm), 166t, 167
  - SSD (sum of squared differences), 101–102
  - Standard deviation, 23–24
  - State of the process, 75
  - Statistical analysis of data, 25–39
  - Statistical approaches to data interpretation, 120–122, 121t
  - Statistical hypothesis testing, 25–28
    - multiple hypothesis testing, 34–36
    - steps in, 25–28
      - assumptions, 26
      - interpretation, 27–28
      - null hypothesis, 26
      - simple question, 25–26
      - summarizing data to test the statistic, 26–27
      - testing the hypothesis, 27
  - Statistically independent, 139
  - Statistical power of a study, 127–128
    - improving, 141–144
  - Statistical significance
    - described, 130
    - of pathways, 232
  - Statistical tests
    - on categorical data, 31–33
    - on continuous data, 28–30
  - Statistics
    - bias, 23
    - descriptive, 21–24
    - of dispersion, 22
    - of location, 22
    - nonparametric, 30
    - odd ratios, 37b–38b
    - q* value, 36–37
    - resampling methods, 38–39
    - summary, 22
    - variance, 22–23
  - Stemming, 290
  - Stop word removal, 290
  - Storage
    - constant-time, 14
    - linear time, 14
    - space complexity and, 13–14
  - Storey, J.D., 121
  - Structural MRI, 89
  - Student's *t*-test, 29
  - Study sample, 142
  - Subtractive color mixing, 85
  - Summary statistics, 22
  - Sum of squared differences (SSD), 101–102
  - Supervised learning, 50, 52–53, 52f
  - Supervised learning algorithms, 67–72
    - decision tree, 71–72, 71f
    - k*-nearest neighbors, 67–68
    - linear classification, 69–70
    - naive Bayes, 68–69
    - partitioning, 71–72
    - regression tree, 71
  - Support vector machines (SVM) algorithm, 70, 298, 298t
  - SVDetect, 183
  - Systematized nomenclature of medical terminologies-clinical terms (SNOMED-CT), 293, 294t
- ## T
- Tag SNPs, 133
  - Target-decoy approach, 193
  - t* distribution, 28–30, 28f, 30f
  - Test error, 63–64, 63f

Testing error, 58–59  
Test of statistical independence, 139  
Test set, 58–61, 59f  
Test statistic, 139–141  
     $\chi^2$ , 32  
    described, 26–27  
    *t*, 28–30  
    in tests on continuous data, 28–30  
Text. *See* Biomedical text  
1000 Genomes Project, 142  
Tibshirani, Robert, 79  
Tokenization, 288, 288f  
TopHat, 169, 170–171, 171t  
Training error, 58–59, 63–64, 63f  
Training phase, of machine learning algorithm, 50  
Training set, 58–61, 59f, 62–64  
    feature selection and, 66  
    multiple, 72  
Trans-ABYSS, 171, 171t  
Transcription factor, binding to promoter,  
    258, 258b  
Transformation, 100–101  
    affine, 101  
    nonrigid (elastic), 101  
    rigid, 100–101  
Transformation of data, 207, 207f  
Transition probabilities, 75  
tRMA method, 114  
True negative, 54, 55f  
True positive, 53, 55f  
*t* statistic, 28–30  
*t*-test, 29, 110, 120–122  
Tukey method, 36  
Two-color microarrays  
    MA plots, 111, 111b  
    one-color compared, 108  
    overview, 109–110  
    preprocessing and normalization, 110–111, 111f  
Type 1 error, 35, 120–121, 121t, 140  
Type 2 error, 35

## U

Unified medical language system (UMLS), 293,  
    294t–295t  
Unsupervised clustering methods, 115–117,  
    116f–117f  
Unsupervised learning algorithms, 76–78  
    dimensionality reduction, 78

    hierarchical clustering, 77  
    *k*-means clustering, 77–78  
Unsupervised learning tasks, 51, 52f, 53

## V

Variability, measures of  
    ANOVA, 30  
    F distribution, 30  
    interquartile range, 24  
    *t* statistic, 30  
    variance, 23  
Variable  
    described, 10  
    hidden, 254  
    sensitivity, 244  
Variance  
    bias and, 23  
    covariance, 40  
    described, 22–23  
    population, 23  
    sample, 23, 27  
    standard deviation, 23–24  
VariationHunter, 183  
Velvet, 171, 171t  
Visualization, 43–44, 44f  
Volex, 91–92  
v-structure, 252–253

## W

Water, heavy, 198  
Web Ontology Language (OWL), 297  
Weka (program), 79  
Welcome Trust Case-Control Consortium (WTCCC),  
    131, 132  
While-loops, 10  
Wilcoxon rank-sum test, 30  
*Wnt*/ $\beta$ -catenin pathway, 235  
WordNet, 297

## X

X! Tandem, 192

## Y

Yates, John, 191