

Summary

The amount, complexity, and speed of aggregation of biomedical and healthcare data will rapidly increase over the next decade. It's likely to double every 1–2 years. This is fueled by enormous strides in digital and communication technologies, IoT devices, and Cloud services, as well as rapid algorithmic, computational and hardware advances. The proliferating public demand for (near) real-time detection, precise interpretation, and reliable prognostication of human conditions in health and disease also accelerates that trend.

The future does look promising despite the law of diminishing returns, which dictates that sustaining the trajectory clinical gains and the speed of breakthrough developments derived from this increased volume of information, paired with our ability to interpret it, will demand increasingly more resources. Even incremental advances, partial solutions, or lower rates of progress will likely lead to substantive improvements in many human experiences and enhanced medical treatments. Figure 1 below illustrates a common predictive analytics protocol for interrogating big and complex biomedical and health datasets. The process starts by identifying a challenge, followed by determining the sources of data and meta-data, cleaning, harmonizing and wrangling the data components, preprocessing the aggregated archive, model-based and model-free scientific inference, and ends with prediction, validation, and dissemination of data, software, protocols, and research findings.

Our long-term success will require major headways on multiple fronts of data science and predictive analytics. There are urgent demands to develop new algorithms and optimize existing ones, introduce novel computational infrastructure, as well as enhance the abilities of the workforce by overhauling education and training activities. Data science and predictive analytics represents a new and transdisciplinary field, where engagement of heterogeneous experts, multi-talented team-work, and open-science collaborations will be of paramount importance.

The DSPA textbook attempts to lay the foundation for some of the techniques, strategies, and approaches driving contemporary analytics involving Big Data (large size, complex formats, incomplete observations, incongruent features, multiple sources, and multiple scales). It includes some of the mathematical formalisms,



Fig. 1 Major steps in a general predictive data analytics protocol

computational algorithms, machine learning procedures, and demonstrations for Big Data visualization, simulation, mining, pattern identification, forecasting and interpretation.

This textbook (1) contains a transdisciplinary treatise of predictive health analytics; (2) provides a complete and self-contained treatment of the theory, experimental modeling, system development, and validation of predictive health analytics; (3) includes unique case-studies, advanced scientific concepts, lightweight tools, web demos, and end-to-end workflow protocols that can be used to learn, practice, and apply to new challenges; and (4) includes unique interactive content supported by the active community of over 100,000 *R*-developers. These techniques can be translated to many other disciplines (e.g., social network and sentiment analysis, environmental applications, operations research, and manufacturing engineering).

The following two examples may contextually explain the need for inventive data-driven science, computational abilities, interdisciplinary expertise, and modern technologies necessary to achieve desired outcomes, like improving human health, or optimizing future returns on investment. These aims can only be accomplished by experienced teams of researchers who can develop robust decision support systems using modern techniques and protocols, like the ones described in this textbook.

- A **geriatric neurologist** is examining a patient complaining of gait imbalance and postural instability. To determine if the patient may have Parkinson's disease, the physician acquires clinical, cognitive, phenotypic, imaging, and genetics data (Big Healthcare Data). Currently, most clinics and healthcare centers are not equipped with skilled data analysts that can wrangle, harmonize and interpret such complex datasets, nor do they have access to normative population-wide summaries. *A reader that completes the DSPA course of study will have the basic competency and ability to manage the data, generate a protocol for deriving candidate biomarkers, and provide an actionable decision support system.* This protocol will help the physician understand holistically the patient's health and make a comprehensive evidence-based clinical diagnosis as well as provide a data-driven prognosis.
- To improve the return on investment for their shareholders, a **healthcare manufacturer** needs to forecast the demand for their new product based on observed environmental, demographic, market conditions, and bio-social sentiment data. This clearly represents another example of Big Biosocial Data. The organization's data-analytics team is tasked with building a workflow that identifies, aggregates, harmonizes, models and analyzes all available data elements to generate a trend forecast. This system needs to provide an automated, adaptive, scalable, and

reliable prediction of the optimal investment and R&D allocation that maximizes the company's bottom line. *Readers that complete the materials in the DSPA textbook will be able to ingest the observed structured and unstructured data, mathematically represent the data as a unified computable object, apply appropriate model-based and model-free prediction techniques to forecast the expected relation between the company's investment, product manufacturing costs, and the general healthcare demand for this product by patients and healthcare service providers.* Applying this protocol to pilot data collected by the company will result in valuable predictions quantifying the interrelations between costs and benefits, supply and demand, as well as consumer sentiment and health outcomes.

The DSPA materials (book chapters, code and scripts, data, case studies, electronic materials, and web demos) may be used as a reference or as a retraining or refresher guide. These resources may be useful for formal education and informal training, as well as, for health informatics, biomedical data analytics, biosocial computation courses, or MOOCs. Although the textbook is intended to be utilized for one, or two, semester-long graduate-level courses, readers, trainees and instructors should review the early sections of the textbook for utilization strategies and explore the suggested completion pathways.

As acknowledged in the front matter, this textbook relies on the enormous contributions and efforts by a broad community, including researchers, developers, students, clinicians, bioinformaticians, data scientists, *open-science* investigators, and funding organizations. The author strongly encourages all DSPA readers, educators, and practitioners to actively *contribute* to data science and predictive analytics, share data, algorithms, code, protocols, services, successes, failures, pipeline workflows, research findings, and learning modules. Corrections, suggestions for improvements, enhancements, and expansions of the DSPA materials are always welcome and may be incorporated in electronic updates, errata, and revised editions with appropriate credits.

Glossary

Table 1 Glossary of terms and abbreviations use in the textbook

Notation	Description
ADNI	Alzheimer’s Disease Neuroimaging Initiative
AD	Alzheimer’s Disease patients
Allometric relationship	Relationship of body size to shape, anatomy, physiology and behavior
ALS	Amyotrophic lateral sclerosis
API	Application program interface
Apriori	Apriori Association Rules Learning (Machine Learning) Algorithm
ARIMA	Time-series autoregressive integrated moving average model
array	Arrays are R data objects used to represent data in more than two dimensions
BD	Big Data
cor	correlation
CV	Cross Validation (an internal statistical validation of a prediction, classification or forecasting method)
DL	Deep Learning
DSPA	Data Science and Predictive Analytics
Eigen	Referring to the general Eigen-spectra, eigen-value, eigen-vector, eigen-function
FA	Factor analysis
GPU or CPU	Graphics or Central Processing Unit (computer chipset)
GUI	graphical user interface
HHMI	Howard Hughes Medical Institute
I/O	Input/Output
IDF	inverse document frequency
IoT	Internet of Things
JSON	JavaScript Object Notation
k-MC	k-Means Clustering

(continued)

Table 1 (continued)

Notation	Description
lm()	linear model
lowess	locally weighted scatterplot smoothing
LP or QP	linear or quadratic programming
MCI	mildly cognitively impaired patients
MIDAS	Michigan Institute for Data Science
ML	Machine-Learning
MOOC	massive open online course
MXNet	Deep Learning technique using R package MXNet
NAND	Negative- AND logical operator
NC or HC	Normal (or Healthy) control subjects
NGS	Next Generation Sequence (Analysis)
NLP	Natural Language Processing
OCR	optical character recognition
PCA	Principal Component Analysis
PD	Parkinson's Disease patients
PPMI	Parkinson's Progression Markers Initiative
(R)AWS	(Risk for) Alcohol Withdrawal Syndrome
RMSE	root-mean-square error
SEM	structural equation modeling
SOCR	Statistics Online Computational Resource
SQL	Structured Query Language (for database queries)
SVD	Singular value decomposition
SVM	Support Vector Machines
TM	Text Mining
TS	Time-series
w.r.t.	With Respect To, e.g., " <i>Take the derivative of this expression w.r.t. a_1 and set the derivative to 0, which yields $(S - \lambda_N)a_1 = 0$.</i> "
XLSX	Microsoft Excel Open XML Format Spreadsheet file
XML	eXtensible Markup Language
XOR	Exclusive OR logical operator

Index

A

Accuracy, 10, 211, 275, 276, 283, 301–303, 307, 323–325, 334, 335, 337, 339, 340, 342, 343, 377, 409, 424, 432, 463, 475, 479–482, 484, 485, 497, 500, 502, 504, 507, 508, 511, 561, 562, 573, 576, 583, 599, 605, 692, 698, 704, 726, 767, 781, 782, 784, 793, 800, 801, 806

Activation, 383–385, 403, 767–769, 774, 775, 781, 785, 799, 800

Activation functions, 384, 385, 767, 781

add, 16, 22, 24, 33, 41, 146, 155, 158, 159, 162, 225, 227, 230, 292, 332, 373, 386, 391, 402, 403, 418, 424, 454, 479, 530, 538, 595, 605, 633, 645, 712, 801

Alcohol withdrawal syndrome (RAWS), 3, 824

Allometric, 266, 817, 823

Allometric relationship, 817

ALSFRS, 4, 559, 733, 783

Alzheimer's disease (AD), 4, 149–151, 569, 823

Alzheimer's disease neuroimaging initiative (ADNI), 4, 823

Amyotrophic lateral sclerosis (ALS), 4, 140, 141, 559–569, 733, 783–784, 823

Analog clock, 816

Appendix, 56–60, 138–139, 149, 183–197, 420

Application program interface (API), 525, 784, 823

Apriori, 267, 268, 423–427, 431, 441, 472, 823

ARIMA, 623, 626, 628, 630–638, 823

array, 20, 25, 31–33, 145

array (), 18

Assessment, 282–286, 510–511

Assessment: 22. deep learning, neural networks, 816–817

assocplot, 40

assocplot(x) Cohen's Friendly graph shows the deviations from independence of rows and columns in a two dimensional contingency table, 40

attr, 27

Attributes, 26, 27, 144, 289, 311, 313, 315, 342, 530, 560, 561, 670

axes, 41, 46, 47, 131, 152, 154, 159, 171, 191, 219, 249, 258, 261, 368, 595, 648

axes=TRUE, 41

B

Bar, 15, 140, 143, 147, 159, 161, 162, 164

barplot, 39, 161, 162, 164, 463

barplot(x) histogram of the values of x. Use horiz=FALSE for horizontal bars, 39

Beach, 811–812

Big Data, 1, 4, 8–10, 12, 642, 661, 765, 819, 823

Biomedical, 8–9

Bivariate, 39, 40, 46, 77, 140, 153–156, 173, 238, 240, 252, 738–739, 766, 770

Black box, 383, 766

boxplot, 39, 70, 161

boxplot(x) 'box-and-whiskers' plot, 39

Brain, 4, 178, 286, 511, 769, 814–815

C

c(), 18–20, 552
 c (), seq (), rep (), and data.frame (). Sometimes we use list () and array () to create data too, 18
 C/C++, 13
 Cancer, 293, 294, 296, 298, 302, 303, 424, 427, 432
 Caret, 322, 477, 486, 487, 491, 492, 497–510, 554, 555, 564, 776
 Chapter, 13, 63, 69, 139, 143, 149, 164, 183, 201, 222, 245, 268, 271, 274, 289, 295, 298, 300, 301, 308, 317, 322, 329, 334, 336, 337, 342, 345–348, 353, 358, 361, 370, 373, 380, 383, 390, 392, 394, 398, 401, 409, 414–416, 420, 427, 442, 447–449, 465, 475–480, 488, 491, 492, 494, 527, 546, 553, 554, 557, 563, 564, 570, 573, 574, 585, 592, 599, 601, 623, 657, 659, 672, 674, 684, 689, 695, 697, 712, 713, 715, 717, 719, 720, 723, 727, 733, 735, 736, 738, 749, 753, 756, 763, 766, 795, 817
 Chapter 22, 415, 817
 Chapter 23, 164
 Chronic disease, 316, 330, 335, 383, 416, 476, 503
 Classification, 144, 267, 268, 281, 286–287, 289, 304–305, 307, 323, 331–332, 396–403, 477, 478, 498, 510, 533, 773–782, 795–805, 816
 Clinical, 258, 612, 614, 695
 Coast, 812
 Cognitive, 2, 4, 7, 149, 700, 820
 Color, 45, 46, 87, 132, 151, 154, 165, 167, 172, 269, 444, 649, 660
 confusionMatrix, 283, 322, 477, 480, 482, 485, 776, 787
 Constrained, 244, 587, 735, 740–747, 750
 Contingency table, 35, 40, 78, 500
 contour, 40
 contour(x, y, z) contour plot (data are interpolated to draw the curves), x and y must be vectors and z must be a matrix so that dim(z)=c(length(x), length(y)) (x and y may be omitted), 40
 coplot, 40
 coplot(x~y | z) bivariate plot of x and y for each value or interval of values of z, 40
 Coral, 815
 Cosine, 659, 685, 695
 Cosine similarity, 695

Cost function, 217, 503, 573, 586, 703, 735, 743, 747, 757, 758
 CPU, 553, 765, 775, 782, 800, 804, 805, 823
 Create, 19, 22, 76–78, 83, 132, 174, 202, 214, 222, 224, 273, 274, 299, 315, 318, 319, 370, 380, 383, 390, 450, 461, 489, 491, 504, 538, 607, 630, 638, 644, 645, 647, 661, 674, 688, 717, 775, 781
 Crossval, 776, 787
 Cross validation, 477, 599–601, 733–734, 823

D

Data frame, 19, 21, 22, 24, 28, 29, 31, 33–36, 39, 40, 47, 48, 66, 131, 132, 153, 164, 172, 174, 273, 274, 299, 300, 319, 438, 451, 490, 514, 526, 529, 537, 540, 547–549, 555, 561, 562, 565, 608
 data.frame, 19, 25, 83, 103, 164, 273
 Data science, 1, 9, 11, 661, 823, 824
 Data Science and Predictive Analytics (DSPAs), 1, 11–13, 198, 492, 623, 661, 819–821, 823
 Decision tree, 307, 310–316, 498, 510, 533
 Deep learning, 765–768, 816–817, 823, 824
 classification, 816
 regression, 817
 Denoising, 735, 756, 757, 760, 763
 Density, 46, 48, 49, 72, 98, 132, 133, 140, 141, 143–147, 173, 174, 198, 287, 289
 Device, 775, 800
 diagnosticErrors, 718, 776
 Dichotomous, 40, 271, 318, 459, 460, 478, 655, 698, 733, 746, 747, 770
 Dimensionality reduction, 233, 265–266
 Divide-and-conquer, 307, 311, 373
 Divide and conquer classification, 307
 Divorce, 443, 448–455, 467, 470
 dotchart, 39
 dotchart(x) if x is a data frame, plots a Cleveland dot plot (stacked plots line-by-line and column-by-column), 39
 Download, 15, 555, 806, 817

E

Earthquake, 132–135, 157, 159, 172
 Ebola, 5
 Eigen, 219, 823
 Entropy, 311–313, 342

- Error, 28, 47, 57, 60, 162, 163, 217, 254, 258, 270, 280, 281, 287, 302, 305, 311, 313, 316, 321, 324–325, 328, 329, 331, 332, 350, 361, 378, 388, 391, 393, 412, 478–480, 487, 491, 500, 501, 504, 507, 509, 562, 565, 573, 576, 579, 582–584, 586, 587, 599, 618, 640, 645, 648, 697, 701–703, 712, 714, 725, 733, 734, 784, 824
- Evaluation, 268, 282, 322, 335, 361, 443, 451, 475, 477, 491, 492, 501, 504, 507, 510, 543, 546, 554, 697, 703, 817
- Exome, 6
- Expectations, 11–12
- Explanations, 41, 510
- F**
- Face, 815–816
- Factor, 21, 24, 46, 79, 210, 219, 233, 255, 256, 259, 265, 287, 292, 294, 299, 319, 333, 352, 359, 412, 417, 438, 561, 570, 575, 588, 600, 608, 630, 638–640, 644, 676, 677, 703, 725
- Factor analysis (FA), 233, 242, 243, 254–256, 262, 265, 638, 639, 644, 823
- False-negative, 700
- False-positive, 325, 573, 574, 619
- Feature selection, 557–559, 571–572
- Feedforward neural net, 817
- filled.contour(x, y, z) areas between the contours are colored, and a legend of the colors is drawn as well, 40
- Flowers, 39, 63, 309, 383, 410, 411, 414, 510
- Format, 13, 17–18, 22, 36, 38, 427, 513–515, 522, 524, 525, 529, 537, 553, 665, 799, 801, 805
- Foundations, 13, 638–641
- fourfoldplot(x) visualizes, with quarters of circles, the association between two dichotomous variables for different populations (x must be an array with dim=c(2, 2, k), or a matrix with dim=c(2, 2) if k = 1), 40
- Frequencies, 29, 39, 46, 145, 193, 298, 429, 430, 439, 463, 484, 485, 667, 672, 685
- Function, 2, 4, 16, 20, 22, 28, 30, 32–35, 37, 47–50, 57–60, 66, 68–70, 76–78, 83, 131–133, 143, 145, 148, 149, 151, 153, 155, 157, 161, 162, 167, 172–175, 187, 202, 207, 208, 213, 216–219, 222, 224, 225, 234, 243, 246, 247, 251, 254, 255, 257, 260, 267, 269, 272–274, 289, 295, 299, 300, 308, 313, 314, 317, 319, 322, 323, 332, 334, 337, 351, 352, 356, 358, 361, 370, 375, 376, 378, 383–385, 390–392, 394–397, 401–403, 411–413, 427, 428, 432, 434, 438, 449–451, 455, 470, 475, 479, 480, 483, 490, 494, 499–501, 504–506, 508, 509, 514, 524, 526, 530, 532, 542, 547–554, 560, 561, 563, 569, 575, 579, 582, 586, 595, 600, 602, 607, 616, 625, 631, 632, 634, 637, 640, 644, 645, 649, 655, 660, 664–667, 673–676, 688, 702, 709, 713, 714, 716, 717, 735–741, 748, 749, 753, 767–770, 772, 774–776, 781, 782, 785, 799–801, 808, 823
- Functional magnetic resonance imaging (fMRI), 178–181, 623, 657
- Function optimization, 243, 735, 761–763
- G**
- Gaussian mixture modeling, 443
- Generalized estimating equations (GEE), 653–657
- Geyser, 174, 175, 813
- ggplot2, 14, 16, 131, 132, 157, 164, 172, 455, 648
- Gini, 311, 313, 335, 336, 342
- Glossary, 823
- Google, 383, 388–394, 396–398, 416, 491, 492, 494, 658, 697–700, 773, 784, 817
- GPU, 513, 553, 765, 775, 782, 804, 805, 823
- Graph, 14, 40, 47, 70, 75, 77, 164, 166, 198, 244, 287, 297, 305, 356, 376, 386, 391, 393, 399, 430, 431, 443, 448, 489, 528–533, 555, 562, 563, 570, 613, 626, 628, 649, 650, 658, 676, 775, 784
- Graphical user interfaces (GUIs), 15–16, 823
- H**
- Handwritten digits, 795, 799, 801
- HC, 135, 705, 824
- Heatmap, 134, 150–152
- Help, 16
- Heterogeneity, 11, 311
- Hidden, 135, 386, 391, 393, 394, 398, 416, 660, 765–767, 772, 774, 775, 781, 785, 799
- Hierarchical clustering, 443, 467–469, 727
- High-throughput big data analytics, 10
- hist, 39, 83, 144
- hist(x) histogram of the frequencies of x, 39

- Histogram, 39, 46, 51, 68, 71–74, 87, 140, 143, 144, 146, 174, 180, 198, 222, 249, 250, 353, 356, 634, 792
- Horizontal, 39, 45, 46, 70, 151, 152, 159, 230, 356, 368
- Hospital, 346, 347, 513, 655, 656
- Howard Hughes Medical Institute (HHMI), 5, 823
- I**
- IBS, 789–792
- if TRUE superposes the plot on the previous one (if it exists), 41
- Image, 20, 24, 40, 83, 84, 176–178, 403, 404, 660, 781, 795, 796, 799, 801, 806–816
- Image classification, 817
- image(x, y, z) plotting actual data with colors, 40
- Independent component analysis (ICA), 233, 242, 243, 250–254, 265
- Index, 187, 313, 316, 388, 389, 392, 416, 513, 625, 641
- Inference, 1, 13, 201, 282, 289, 513, 573, 638, 655, 659, 735, 819
- Input/output (I/O), 22–24, 64, 765, 823
- interaction.plot(f1, f2, y) if f1 and f2 are factors, plots the means of y (on the y-axis) with respect to the values of f1 (on the x-axis) and of f2 (different curves). The option fun allows to choose the summary statistic of y (by default fun=mean), 40
- Interpolate, 48
- Intersect, 30
- Inverse document frequency (IDF), 659, 676–686, 695, 823
- Iris, 63, 64, 308–310, 409–411, 414, 727
- J**
- Java, 10, 13, 20, 72, 332, 334, 349, 534
- Jitter, 143, 157
- JSON, 198, 513, 514, 522, 525–526, 531, 533, 823
- K**
- k-Means Clustering (k-MC), 443
- k-nearest neighbor (kNN), 268, 269, 447
- Knockoff, 574, 621
- L**
- Lagrange, 401, 402, 735, 740–741, 749, 753–756, 762
- Lake Mapourika, 810–811
- Lattice, 46, 47
- Layer, 386, 388, 394, 765–768, 770, 771, 773–775, 781, 782, 785, 799–801
- Lazy learning, 267, 286–287
- Length, 5, 19, 21, 26, 28, 35, 37, 40, 46, 47, 63, 64, 132, 174, 230, 231, 235, 270, 273, 346, 374, 377, 409, 480
- Letters, 148, 193, 195, 215, 404, 530, 664
- Linear algebra, 201, 229–231, 345
- Linear mixed models, 623
- Linear model, 574–582, 621, 650
- Linear programming, 735, 748
- list (), 18–20
- lm (), 16, 225, 358, 553, 824
- log, 30, 31, 40, 313, 517, 587, 610, 611, 615, 616, 640, 716
- Log-linear, 40
- Long, 5, 13, 18, 36, 514, 547, 565, 676, 784, 819
- Longitudinal data, 40, 657–658
- Lowess, 824
- M**
- Machine learning, 2, 10, 267, 268, 289, 322, 383, 423, 443–444, 476, 477, 481, 497, 536, 549, 562, 659, 660, 667, 689, 765, 809, 816, 820
- Managing data, 63, 140–141
- Mask, 815–816
- matplot(x, y) bivariate plot of the first column of x vs. the first one of y, the second one of x vs. the second one of y, etc, 40
- Matrices, 20, 21, 24, 31, 149, 167, 201–203, 206–209, 213–216, 219, 220, 222, 229, 230, 233, 258, 478–480, 490, 549, 574, 640, 641, 645, 650, 667, 672, 698, 714, 716, 735, 782, 804
- Matrix, 13, 21, 26, 28, 31, 32, 40, 46, 47, 81, 132, 149–151, 153, 161–163, 166, 167, 174, 201–209, 211, 212, 214–217, 219–222, 224, 225, 227–231, 235, 236, 238–240, 242, 244, 245, 247, 251, 254–258, 260, 265, 295, 299, 300, 304, 305, 319, 322, 324–325, 350, 351, 356, 391, 427, 429–432, 450, 463, 478, 480, 483, 484, 501, 506, 507, 528–530, 537, 540, 552, 555, 574, 582, 607, 608, 620, 639–641, 648, 650, 654, 655, 660,

- 667–668, 670–674, 676, 685, 688, 689, 695, 702, 716, 717, 727, 735, 739, 747, 748, 753, 766, 767, 782, 799–801
- Matrix computing, 201, 229–231, 345
- Michigan Institute for Data Science (MIDAS), 824
- Mild cognitive impairment (MCI), 4, 149, 151, 824
- Misclassification, 311, 324–325, 411, 418
- mlbench, 536, 774
- mlp, 774, 775, 781, 782, 785
- Model, 2, 10, 13, 47–48, 81, 93, 110, 120, 166, 201, 216, 217, 227, 230, 246, 252, 253, 260, 262, 267, 268, 274–276, 283, 286, 299–301, 345, 350, 356, 358–375, 377–381, 383, 385–387, 391–394, 397, 398, 405–409, 411–416, 418, 479, 488, 489, 510, 511, 571, 572, 658, 733, 734, 817
- Model performance, 268, 274–276, 300–301, 322–323, 333, 359–373, 377–380, 386, 392–394, 406–409, 412–414, 433–438, 451–454, 462–465, 475, 479, 480, 487, 488, 491, 492, 494, 495, 497, 501–503, 507, 564–569, 572, 605, 697, 698, 701
- Model-based, 2, 10, 345, 481, 566, 573, 660, 710, 819, 821
- Model-free, 2, 10, 481, 660, 689, 705, 819, 821
- Modeling, 1, 4, 9, 13, 48, 83, 201, 216–217, 233, 259, 307, 347, 349, 505, 513, 528, 582, 638, 640, 659, 668, 701, 703, 756, 775, 820, 824
- MOOCs, 821
- mosaicplot, 40
- mosaicplot(x) “mosaic” graph of the residuals from a log-linear regression of a contingency table, 40
- Multi-scale, 623
- Multi-source, 9, 514, 559
- MXNet, 774, 775, 782, 785, 799–801, 804, 805, 817
- N**
- NA, 22, 24, 28, 30, 38, 67, 69, 155, 287, 380, 427, 429, 538, 625
- na.omit, 28, 48
- na.omit(x), 28
- Naive Bayes, 289, 290, 299, 302–305, 476
- Natural language processing (NLP), 442, 659–668, 689–691, 694–695, 824
- Nearest neighbors, 267, 286–287, 719–720
- Negative AND (NAND), 771–772, 824
- Network, 383, 384, 386, 398, 533, 555, 730, 731, 773, 799–800, 804–806
- Neural networks, 383–388, 498, 510, 717–718, 765, 766, 816–817
- Neurodegeneration, 4–5
- Neuroimaging, 4, 7, 588–590, 608–621, 789, 817, 823
- New Zealand, 810–811
- Next Generation Sequence (NGS), 6–7, 824
- Next Generation Sequence (NGS) Analysis, 6–7
- Nodes, 164, 293, 307, 311, 316, 321, 336, 374, 376, 379, 383, 386, 391, 393, 394, 416, 524, 528–530, 532, 765, 766, 768, 775, 785
- Non-linear optimization, 752–753, 762
- Normal controls (NC), 4, 149, 151, 152, 167, 169–171, 824
- Numeric, 2, 19, 25, 46, 47, 66, 68, 71, 76, 77, 145, 149, 150, 212, 259, 273, 274, 299, 319, 370–371, 377, 396, 409, 503, 559, 570
- O**
- Objective function, 242, 250, 251, 401, 558, 573, 574, 579, 587, 592, 640, 641, 735–738, 740, 741, 747–749, 753, 754, 756–758
- Open-science, 1, 819, 821
- Optical character recognition (OCR), 383, 403–408, 795, 824
- Optimization, 13, 47, 243, 254, 401, 402, 513, 546, 573, 574, 579, 587, 592, 641, 735–753, 755, 756, 761, 762
- Optimize, 47, 337, 401, 739, 757, 758, 819
- P**
- Package, 30, 38, 46, 63, 78, 81, 131, 132, 138, 149, 157, 164, 167, 172, 174, 208, 247, 274, 294, 297, 299, 319, 322, 332, 356, 358, 375, 376, 378, 391, 410, 412, 413, 428, 434, 455, 467, 470, 483, 486–489, 491–493, 497, 501, 503, 505–507, 509, 514, 515, 517, 522–524, 526, 528, 531, 532, 534, 536, 547–555, 559–561, 588, 607, 608, 626, 627, 632, 641, 645, 648, 650, 661, 663, 666, 668, 672, 675, 686, 705, 723, 727, 741, 748, 754, 760, 765, 776, 804, 806, 824

- pairs, 5, 40, 45, 153–156, 164, 191, 234, 237, 239, 311, 356, 357, 371, 424–427, 529, 532, 770, 796
 - pairs(x) if x is a matrix or a data frame, draws all possible bivariate plots between the columns of x, 40
 - Parallel computing, 548–553, 555–556
 - Parkinson’s disease (PD), 51, 135, 261, 262, 265, 511, 571, 600, 608–621, 642–647, 650, 705, 711, 824
 - Parkinson’s Progression Markers Initiative (PPMI), 245, 286, 511, 571–572, 588, 642, 656, 705, 711, 719, 824
 - Perceptron, 766, 769, 773, 775, 785
 - Perl, 13
 - persp(x, y, z) plotting actual data in perspective view, 40
 - Petal, 39, 64, 727
 - Pie, 39, 143, 147, 149, 167, 170, 198
 - pie(x) circular pie-chart., 39
 - Pipeline environment, 10
 - Plot, 39–47, 66, 70, 71, 73, 74, 77, 84, 98, 131–133, 136, 140, 141, 143, 145–148, 150, 153–160, 162–164, 166–177, 180, 188, 191–194, 226, 230, 231, 233, 235, 239–241, 243, 247, 249, 250, 256, 262, 285, 286, 298, 323, 325, 326, 331, 346, 347, 353, 356, 357, 359, 361, 366, 368, 375–377, 414, 430, 431, 436, 439, 448, 452, 454, 456, 459, 467, 469, 471, 472, 532, 534, 535, 537, 539, 562, 563, 565, 570, 571, 590, 592, 597, 602, 618, 626, 629–631, 727, 736, 739, 742, 757, 768, 776, 778, 779, 792
 - plot(x) plot of the values of x (on the y-axis) ordered on the x-axis, 39
 - plot(x, y) bivariate plot of x (on the x-axis) and y (on the y-axis), 39
 - plots.ts(x) if x is an object of class “ts”, plot of x with respect to time, x may be multivariate but the series must have the same frequency and dates. Detailed examples are in Chap. 19
 - big longitudinal data analysis, 40
 - Predict, 3, 4, 9, 10, 48, 81, 267, 283, 300, 322, 334, 346, 377–380, 389, 391, 392, 411, 412, 475, 476, 478, 500, 504, 505, 555, 582, 584–586, 600, 602, 623, 674, 679, 700, 703, 712, 714, 717, 783, 792, 796, 806, 808, 817
 - Predictive analytics (PA), 1, 9–10, 661, 823
 - Principal component analysis (PCA), 233, 241–249, 254, 256–258, 260, 263, 265, 266, 533, 824
 - Probabilistic learning, 304–305
 - Probabilities, 31, 173, 300, 322, 476, 482, 500, 600, 684, 715, 716, 767, 800, 801
 - Pruning, 307, 315, 316, 328, 330
 - Python, 13
 - Python, Java, C/C++, Perl, and many others, 13
- Q**
- QOL, 317
 - qqnorm, 40
 - qqnorm(x) quantiles of x with respect to the values expected under a normal law, 40
 - qqplot, 40
 - qqplot(x, y) quantiles of y with respect to the quantiles of x, 40
 - Quadratic programming, 824
 - Quality of life, 490, 792
- R**
- R, 1, 12–17, 20, 24, 25, 30–32, 37–39, 41, 46, 48–50, 56–60, 63–65, 67–69, 75–78, 84, 130, 131, 138–141, 143, 144, 149, 153, 161, 173, 175, 176, 178, 206, 208, 209, 211, 212, 214, 216, 219, 220, 224–227, 229, 230, 236, 240, 245, 246, 251, 254, 257, 258, 274, 294, 297, 307, 319, 332, 334, 349, 355, 356, 361, 374, 375, 378, 389, 409, 410, 420, 427, 428, 436, 438, 450, 460, 467, 470, 477, 479, 486, 488, 491, 501, 514, 515, 517, 524, 526–529, 532–534, 538, 540, 546–548, 550, 553, 559, 588, 617, 619, 624, 629, 641, 642, 650, 652, 659, 661, 694, 704, 714, 736–738, 748, 753, 754, 760, 765, 774, 782, 801, 806, 807, 820, 821, 823, 824
 - Regularized, 574–582, 621
 - Regularized linear model, 621
 - Relationship, 77, 78, 155, 226, 245, 314, 345, 349, 350, 369, 383, 394, 448, 454, 532, 581, 584, 640, 817, 823
 - rep(), 18
 - Require, 12, 400, 409, 432, 527, 550, 555, 563, 772, 815, 819
 - reshape2, 14, 16
 - Risk for Alcohol Withdrawal Syndrome (RAWS), 3

- Root mean square error (RMSE), 329, 477, 565, 698, 701, 703, 784, 817, 824
- RStudio, 13, 15–16
- RStudio GUI, 15–16
- S**
- Scatter plot
 scatter, 46, 153, 226, 230, 231
- Sensitivity, 305, 485, 486, 714, 734, 776
- seq (), 18, 38, 69
- Sequencing, 6
- set.seed, 37, 49, 287, 333, 492, 499, 782, 800
- setdiff, 30
- setequal, 30
- Silhouette, 443, 446, 451, 452, 456–459, 463, 464, 469, 477, 723, 725
- sin, cos, tan, asin, acos, atan, atan2, log, log10, exp and “set” functions union(x, y), intersect(x, y), setdiff(x, y), setequal(x, y), is.element(el, set) are available in R, 30
- Singular value decomposition (SVD), 233, 241, 242, 256–258, 265, 824
- Size, 16, 30, 46, 47, 49, 132, 135, 145, 154, 174, 192, 209, 210, 269, 315, 316, 323, 328, 336, 348, 390, 425, 426, 429, 450, 451, 495, 498, 500, 503, 510, 515, 534–536, 565, 566, 572, 592, 624, 676, 747, 767, 773, 774, 781, 784, 795, 819, 823
- sMRI, 4, 178
- softmax, 498, 767, 774, 781, 800
- Sonar, 774–781
- Sort, 28, 700
- Specificity, 305, 485–486, 734, 776
- Spectra, 231
- Splitting, 268, 307, 311, 315, 373, 374, 536, 584, 686
- SQL, 138–139, 513, 515–521, 537, 553, 824
- Stacked, 39, 46, 196
- stars(x) if x is a matrix or a data frame, draws a graph with segments or a star where each row of x is represented by a star and the columns are the lengths of the segments, 40
- Statistics Online Computational Resource (SOCR), 4, 10, 11, 50, 51, 56, 72, 79, 130, 140, 147, 171, 173, 178, 187, 193, 198, 230, 258, 305, 342, 349, 522, 524, 525, 531–533, 540–543, 555, 569, 584, 669, 817, 824
- stripplot, 39, 46
- stripplot(x) plot of the values of x on a line (an alternative to boxplot() for small sample sizes), 39
- Structural equation modeling (SEM), 623, 638–648, 824
- Summary statistic, 35, 40, 67, 76, 140, 187, 352, 549
- sunflowerplot(x, y) id. than plot() but the points with similar coordinates are drawn as flowers which petal number represents the number of points, 39
- Support vector machines (SVM), 398–403
- Surface, 132, 141, 174–176, 814–815
- Symbol, 47, 404, 490, 781, 799
- symbols(x, y, ...) draws, at the coordinates given by x and y, symbols (circles, squares, rectangles, stars, thermometers or “boxplots”) which sizes, colors... are specified by supplementary arguments, 40
- T**
- Table, 13, 22–24, 29, 30, 32, 35, 38, 40, 76, 78, 79, 140, 144, 148, 166, 208, 268, 274, 275, 282, 292, 300, 301, 311, 317, 322, 412, 426, 450, 463, 477–480, 482, 483, 486, 501, 504, 511, 529, 530, 548, 555, 614, 641, 686, 771
- TensorFlow, 765, 773, 784
- Term frequency (TF), 659, 676–686, 695
- termpart(mod.obj) plot of the (partial) effects of a regression model (mod.obj), 40
- Testing, 7, 268, 274, 282, 287, 299, 303, 318, 324, 342, 396, 414, 491, 505, 579, 581, 584, 599, 600, 639, 648, 679, 684, 686, 690, 691, 697, 701, 703, 704, 719, 765, 775, 782, 784, 795, 799–801
- Text mining (TM), 442, 659–668, 689–691, 694–695, 824
- The following parameters are common to many plotting functions, 40
- Then, try to perform a multiple classes (i.e AD, NC and MCI) classification and report the results, 816
- Training, 8, 141, 260, 267–270, 274, 281, 287, 289, 292, 295–297, 299–300, 303, 304, 311, 318–321, 332–333, 337, 358–359, 374, 375, 380, 390, 391, 395, 396, 398, 410–412, 416, 418, 432–433, 450–451, 461, 491, 493, 495, 501, 503–505, 507, 553, 554, 558–564, 579, 584, 599, 600, 679, 684, 686, 688, 697, 701–704, 715,

- 719, 733, 765, 768, 769, 775, 776, 778,
782, 784, 795, 796, 799, 800, 804, 805,
815, 819, 821
- Transdisciplinary, 9, 819, 820
- Trauma, 163, 443, 459–467
- ts, 1, 31, 40, 47, 77, 80, 533, 629, 631
- ts.plot(x) id. but if x is multivariate the series
may have different dates and must have
the same frequency, 40
- U**
- Unconstrained, 735–741, 761
- Union, 30, 145, 424
- Unique, 1, 7, 29, 144, 150, 210, 334, 429,
463, 495, 527, 639, 667, 688, 698, 736,
774, 820
- Unstructured text, 289, 659, 660, 795
- V**
- Validation, 9, 267, 280, 281, 283, 287, 329,
340, 396, 414, 446, 448, 475, 477,
491–495, 501, 562, 574, 581, 598–600,
675, 679, 689–691, 697, 698, 700–704,
708, 709, 711, 712, 718, 733, 765, 784,
796, 799, 815, 819, 820, 823
- Visualization, 4, 143–145, 164, 198–199, 657
- Visualize, 4, 70, 77, 132, 149, 173, 178, 201,
222, 226, 247, 249, 252, 256, 280, 297,
305, 323, 356, 429, 434, 453, 528, 565,
571, 590, 592, 626, 648, 726
- Volcano, 175, 812–813
- W**
- which.max, 27
- Whole-genome, 6
- Whole-genome and exome sequencing, 6
- Wide, 13, 18, 36, 48, 381, 427, 514, 583,
656, 820
- With, 1, 2, 4, 5, 8, 9, 12, 16–20, 22–26, 28–31,
33, 36, 37, 39–41, 45–50, 57, 67, 69–71,
77, 80, 81, 83, 84, 130, 132, 133, 135,
139, 140, 143, 147, 149, 150, 153,
155–159, 161–163, 166, 171, 173–175,
202, 210, 212–216, 219, 220, 222, 230,
231, 233, 237, 242–244, 254–256, 258,
267, 269–271, 273, 289, 295, 299,
302–305, 307, 310–319, 322, 324, 325,
327, 331–334, 336, 337, 339, 340, 342,
347, 371, 374, 377–380, 385–390, 394,
398–403, 408, 411, 412, 416, 420,
423–425, 427, 429, 430, 432, 435, 438,
439, 441, 442, 444–446, 448–451,
453–460, 463–466, 469, 472, 475, 476,
484, 495, 497, 500, 502–506, 508, 511,
513–536, 540, 543, 546–559, 563–570,
572–574, 576, 584, 586, 599, 605–608,
612–614, 616, 617, 620, 625, 626, 628,
630, 634, 637–639, 642, 643, 648, 653,
655, 656, 663, 665, 668, 670, 672, 674,
676–678, 684, 686, 688, 689, 698, 700,
704, 710, 711, 715, 717, 718, 720, 734,
736, 746, 756, 769–774, 781, 784, 785,
800, 805, 808, 812, 815, 817, 819, 820
- X**
- XLSX, 526–527, 824
- XML, 7, 24, 513, 522–524, 555, 824
- Exclusive OR (XOR), 770, 771, 824
- Y**
- Youth, 271, 443, 465–467, 498