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,

I. D. Dinov, *Data Science and Predictive Analytics*,
https://doi.org/10.1007/978-3-319-72347-1

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 staistical 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)

I. D. Dinov, *Data Science and Predictive Analytics*,
https://doi.org/10.1007/978-3-319-72347-1

Table 1 (continued)

Notation	Description
lm()	linear model
lowess	locally weighted scatterplot smoothing
LP or QP	linear or quadratic programming
MCI	mildly cognitively impared 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., "*Take the derivative of this expression w.r.t.* a_1 *and set the derivative to 0, which yields* $(S - \lambda I_N)a_1 = 0$."
XLSX	Microsoft Excel Open XML Format Spreadsheet file
XML	eXtensible Markup Language
XOR	Exclusive ***OR*** logical operator

Index

I. D. Dinov, *Data Science and Predictive Analytics*,
https://doi.org/10.1007/978-3-319-72347-1

N

O

P

Q

R