Data Analytics in Spectroscopy

Data Analytics in Spectroscopy

By

Joseph Dubrovkin

Cambridge
Scholars
Publishing

Data Analytics in Spectroscopy

By Joseph Dubrovkin

This book first published 2024

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-5275-7968-9
ISBN (13): 978-1-5275-7968-2

TABLE OF CONTENTS

PREFACE

"Ignorance is the curse of God, knowledge the wing wherewith we fly to heaven."
—Henry VI, Part 2 Quotes,
William Shakespeare

A person receives information about the world around us through the senses in different forms (sounds, smells, sights, tastes, and tactile sensations) and, from ancient times, saves it in analog (pictures) or coded representation by abstract symbols in handwritten, and since the Middle Ages in printed materials. Now, artificial sensors' automatic collection of information in the digital form is becoming dominant. Without going into terminological subtleties, let us say that humanity has received and is continuously receiving a huge amount of data on the processes occurring in nature and society, also generated by human and artificial intelligence, which cognizes and models our world in the broadest sense starting with atoms and finishing with galaxies. In other words, the final step is creating collections of information related to specific topics (knowledge bases) from raw data accumulated in databases.

Data as a collection of bits on its own has no value without analysis, allowing decisions concerning the solution to the problem under study. Decision-making is a major topic of Data Science called Data Analytics. Data analytics can reveal hidden trends and metrics emerging in the mass of information by encompassing diverse types of data analysis.

The brilliant application of data analytics is econometrics, which involves statistical and mathematical models to forecast future trends from accumulated historical data [1]. Chemometrics [2] and chemo- and bioinformatics [3] use statistical and computation tools to capture and interpret chemical and biochemical information from data analytics.

Reviews [4, 5] consider the features of applying Data Science in analytical chemistry.

As a derivative of Data Science, data analytics is the subject of different engineering and economics university courses requiring a solid mathematical base. In analytical chemistry and particularly spectroscopy applications, less attention is paid to mathematics since the powerful software tools allow solving analytical tasks without the headaches of

understanding cumbersome equations. Accurate formulation of a task, correct selection of experimental design, and processing methods often guarantee satisfactory solutions. Numerous conscientious studies published in periodical literature demonstrate this semi-blind approach to solving chemometrics problems. The situation is similar to a good driver who does not need to know how his car's engine works. However, this knowledge can improve the car's performance and lengthen its service life.

The goal (as in our previous books [6-8]) was to avoid, where possible, readers' blind faith in the validity of conclusions and recommendations.

The book's central idea is a top-down approach to data analytics presentation: from general concepts developed by data science to the corresponding chemometrics method used in spectroscopy.

Theoretical discussions on this issue involve various examples supplied by a simple program code on MATLAB, which non-professional users can easily modify. Readers who may wish to study the problem further can validate the numerical data from the book using computer calculations. Thus, they will be able to understand the details of the algorithm and, if necessary, modify computer programs.

References

1. Hansen, B. Econometrics. Princeton University Press. 2022.
2. Otto, M. (2016). Chemometrics: Statistics and Computer Application in Analytical Chemistry, 3rd Ed. ISBN: 978-3-527-34097-2, Wiley.
3. Chemoinformatics and Bioinformatics in the Pharmaceutical Sciences. 1st Ed. Editors: Sharma, N., Ojha, H., Raghav, P., Goyal, R. Elsevier. 2021.
4. Szymańska, E. (2018). Modern data science for analytical chemical data - A comprehensive review. Analytica Chimica Acta, 1028, 1-10. DOI.org/10.1016/j.aca.2018.05.038.
5. Gressling, T. Data Science in Chemistry. De Gruyte. 2021.
6. Dubrovkin, J. Mathematical processing of spectral data in analytical chemistry: A guide to error analysis. Cambridge Scholars Publishing. 2018.
7. Dubrovkin, J. Derivative Spectroscopy. Cambridge Scholars Publishing. 2021.
8. Dubrovkin, J. Data Compression in Spectroscopy. Cambridge Scholars Publishing. 2022.

About the Structure of the Book

The book consists of four parts. The first part, 'Information Age and Big Data,' introduces the reader to the basics of Data Science and general concepts of data processing and analytics. The last chapter briefly discusses the Big Data approach to analytical chemistry.

The second part introduces the data analytics theory, including clustering, data mining, classification, and regression. Numerical examples supplied with Matlab code help the user better understand the operation principles of numerous cumbersome algorithms. The purpose of this part is to make it easier to understand the practical methods of data analytics in spectroscopy, which is the subject of the following part.

The last part describes spectral databases (Appendix G considers the web databases). The author paid special attention to the compact description of the open database PhotochemCAD™ [1-3] and its exposition for demonstrating the clustering and classification of UV-VIS spectra.

The bibliography tables briefly describe hundreds of applications of spectral data analytics methods and related topics in industrial and research laboratories.

We sincerely apologize to all those researchers whose outstanding works are not cited because the book lacks the free space to include a complete bibliography. The project "Data Analytics in Spectroscopy" (https://www.researchgate.net/profile/Joseph_Dubrovkin) provides a bibliographical supplement, updated as new information becomes available.

The author wanted to give each chapter the status of a self-contained article whose reading is independent of other sections. This material presentation allows readers to avoid cramming a previous text before moving on to another topic. MATLAB-based examples, which are focused on the subject matter, illustrate each chapter.

The book closes with appendices, which include supplementary materials necessary to facilitate the readers' understanding of the theoretical problems discussed in the main text. For example, a brief introduction to the mathematical methods such as singular value decomposition, discrete Fourier transform, and Tikhonov regularization is given. Reading requires knowledge of secondary school courses on differential calculus, linear algebra, and statistics. To perform the exercises, readers must have programming skills for beginners in MATLAB. Appendix D, which is

dedicated to translating the programming language R to Matlab, allows the users to rewrite the widespread R program in the Matlab environment. Appendix H describes a self-made database for intermediate-level Matlab programmers.

For simplicity, the captions of figures, tables, exercises, and expressions have the following structure: "part. chapter-current number."

The author would be very grateful for the criticisms, comments, and proposals about this book, which he hopes to consider in his future work.

References

1. PhotochemCAD™. https://www.photochemcad.com/
2. Du, H., Fuh, R.-C. A., Li, J., Corkan, L. A., Lindsey, J. S. (1998). PhotochemCAD. A Computer-Aided Design and Research Tool in Photochemistry and Photobiology. Photochemistry and Photobiology, 68, 141–142.
3. Wu, Z., Kittinger, A., Norcross, A. E., Taniguchi, M., Lindsey, J. S. PhotochemCAD Spectra Viewer for Web-based Visualization of Absorption and Fluorescence Spectra. Proc. S.P.I.E. BiOS 2021, Vol. 11660, Reporters, Markers, Dyes, Nanoparticles, and Molecular Probes for Biomedical Applications XIII, 11600I-19. DOI: 10.1117/12.2577840.

ABBREVIATIONS

AI-Artificial Intelligence
AIMS-Analytical Informational-Measurement System
AIC-Akaike Information Criterion
ANN-Artificial Neuron Network
AQP-Algorithms for Quantitative Pedology
AR-Association Rules
ARM-AR Mining
AS-Analytical Signal
ATR-Attenuated Total Reflection
BDA-Big Data Approach
BRANN- Bayesian Regularized ANN
BT-Boosted Tree
CFT-Continuous FT
CLODM-Clustering-Based ODM
CLSODM-Classification-Based ODM
CNN-Convolutional ANN
COS (2D-COS)-correlation spectroscopy
CP-CH-Calinski–Harabasz score
CS-Compressed Sensing
CWT-Continuous WT
DCL-Data clustering
DA-Discriminant Analysis
DALC- Data Analytics Lifecycle
DB-Data Base
DBSCAN-Density-Based Spatial Clustering of Applications with Noise
DCT-Discrete Cosine Transform
DFT-Discrete FT
DL-Deep Learning
DM-Data Mining
DOSC-Direct OSC
DSS-Decision Support System
DT-Decision Tree
DRS-Deep Residual Shrinkage
DW-Data Warehouse
DWT- Discrete WT
ELN-Electronic Lab Notebook
EMSC-Extended MSC
FB-ANN-feedback ANN
FT-Fourier Transform
FTIR-Fourier Transform IR
FTR-FT-Raman
FWHM-Full Peak Width at Half-Maximum
FCA-Fuzzy Clustering Algorithm
FCM-Fuzzy C-Means
FF-ANN-feed-forward ANN
FDA-factorial DA
FFT-Fast FT
GA-Genetic Algorithm
GC-Gas Chromatography
GUI-Graphic User Interface
GDH-Generalized Discrete Harmonics
GFT-Generalized FT
AGK-Gustafson-Kessel algorithm
HC-Hierarchical Clustering
HIS-Hyperspectral Imaging
HPLC-High-Performance LC
HT-Hyphenated Technique
HTML-Hypertext Markup Language
ICA-Independent Component Analysis
ICP-AES-Inductively coupled plasma atomic emission spectroscopy.
IFT-Inverse FT
ISC-Inverse SC
IT-Information Theory
IR-Infrared
KM-HMR- K-Means Hadoop MapReduce
KNN-k-nearest neighbor
LA-ICP-MS-Laser Ablation Inductively Coupled Plasma MS
LC-Liquid Chromatography
LDA-Linear Discriminant Analysis
LIBS-Laser Induced Breakdown Spectroscopy
LL-Lazy Learning
LLNB-Lazy Naive Bayes
LRA-Low Rank Approximation
LS-Least Squares
MALDI-TOF-Matrix-Assisted Laser Desorption/Ionization
MARS-Multivariate Adaptive Regression Splines
MLA-Machine Learning Algorithm

MLP-multilayer perceptron
NIR-Near IR
NMR-Nuclear Magnetic Resonance
MS-Mass Spectroscopy
NBC- Naive Bayes Classifier
NNLS-Nonnegative LS
NoSQL-Not Only SQL
ODBC-Open Database Connectivity
ODM-Outlier Detection Methods
OLAP-Sophisticated On-Line Analytical Processing
OLS-Ordinary Least Squares
OMS-Optical Multisensor Systems
OP-Orthogonal Polynomial
OSC-Orthogonal Signal Correction
OSP-Open Scientific Platform
OTM-Orthogonal Transformation Method
QDA-Quadratic DA
QMF-Quadrature Mirror-Filter
QSAR-Quantitative Structure-Activity Relationship
PAT-Process Analytical Technology
PC-Principal Component
PCA-PC Analysis
PCD-S-Coherent Data Scatter
PC-DFA-PC-Discriminant Function Analysis
PCR-PC Regression
PCT-Phase Coherence Theory
PDV-Principle Discriminant Variate
RDMS-Relational DB Management Systems
PLS-Partial Least Squares
PLS-DA-PLS-Discriminant Analysis
RBF-Radial Basis Function
RDF-Random Decision Forest
RF-Random Forest
RGB-Red, Green, and Blue
RPD-Residual Prediction Deviation
RS-Raman Spectrum
RSDE-Random Subspace Discriminant Ensemble
RTM-radiative transfer model
RSEP-Relative Standard Error of Prediction
SC-Streak Camera
SERDS-Shifted Excitation Raman Difference Spectroscopy
SERS-Surface-Enhanced RS
SELDI-TOF-Surface-Enhanced Laser Desorption/Ionization- Time-Of-Flight
SFGL-LR-Sparse Fused Group Lasso Logic Regression
SFS-Synchronous Fluorescence Spectroscopy
SG-Savitzky-Golay
SIMCA-Soft Independent Modeling of Class Analogies
SMF-Soft Margin Formulation
S/N-Signal-to-Noise Ratio
SNV-Standard Normal Variate
SOM-Self-Organizing Map
SQL-Structured Query Language
SRD-Sum of Ranking Differences
SVD-Singular Value Decomposition
SVM-Support Vector Machine
SWANCAD-Spectral Wavelet-feature Analysis and Classification Assisted Denoising
TD-Tucker Decomposition
TMWDCA- Two-way moving window PCA
TR-Tikhonov Regularization
TRS-Time-Resolved Spectroscopy
TVD-Total Variation Denoising
UV-Ultraviolet
VIS-Visible
WAL-Weighted-Average Linkage
WMV-Ward Minimum Variance
WT-Wavelet Transform
XML-Extensible Markup Language

PART I:

INFORMATION AGE AND BIG DATA

INTRODUCTION

About ten years ago, the author of these lines, the Western Galilee College lecturer at the Faculty of Computer Science, began to teach two courses, “Human-Computer Interaction” and “Introduction to Databases,” for students who did not specialize in computer technology. The latter fact caused significant difficulties in teaching, especially the theme of “Big Data.” At the same time, the topic “Application of Big Data in Analytical Spectroscopy” caught our attention since it was related to the author’s old doctoral research on using mathematical methods in analytical signal processing. The first brief review article [1], dedicated to this topic, was further partly improved [2]. However, recent years have been marked by the rapid growth of scientific research and practical applications in this area, which has prompted us to look at the problem in a broader sense, both from the point of view of data processing and as a user of analytical instruments.

In order to introduce the reader, who is not a professional in the field of data and especially big data, to the course of the problem, it is necessary to answer many questions that arise while reading numerous works on this topic. First, this is due to the superficial interpretation of the terminology used and the insufficient understanding of the essence of the mathematical methods and the principle of operation of technical apparatus.

In preparing this book, the author set the task to eliminate these shortcomings as much as possible.

CHAPTER ONE

GENERAL CONCEPTS

Information Age

Out of the Darkness of Centuries came the predictions of the prophet Daniel:

"But thou, O Daniel, shut up the words, and seal the book, even to the time of the end: many shall run to and fro, and knowledge shall be increased" (The Book of Daniel) 12:3-5)[1].

Some interpreters consider the words "knowledge shall be increased" as a prediction that "to the time of the end," humanity's awareness of the universe as a whole will increase extraordinarily. All sorts of sciences and technologies will develop like never before, which we observe in our Information Age [2].

We cite these words not to start theological disputes but only to emphasize that philosophers of ancient times already foresaw the incredible development of human knowledge in the future.

The information era, or the era of digital technologies, originates from Shannon's famous work on information theory [3]. In principle, Shannon showed that any signals (messages) (e.g., sound and electromagnetic as radio and telephone), represented in binary form (zeroes and units), could be coded for storage and transmission without information loss in the unified framework.

However, practically, the information revolution began only in the 70s of the last century due to the Internet and continued with the advent of personal computers and cellular communication.

The father of the Internet, ARPANet, was created by the Information Processing Techniques Office in the Advanced Research Projects Agency (ARPA), part of the U.S. Dept. of Defense, at the height of the Cold War in 1969 [4].

ARPANet connected the computers of military and research institutions, providing a quick response to the sudden enemy attack. Later, a unified internetwork protocol for managing data transfer allowed combining different networks. So, an information system, the World Wide Web

(WWW), enabling web resources (commercial products and services, personal communications, and others) to be accessed over the Internet, was created.

The WWW is a distributed hypertext information system that contains interlinked Web (HTML) pages. HTML (Hypertext Markup Language) is a standard language of HTML documents (plaintext files) whose structure is controlled by standard tags embedded in the text [5]. Hypertext, unlike plaintext, allows one to read a document by jumping in any direction and even from document to document using special links.

In this connection, the following paragraph [6], in our opinion, is interesting for an inquisitive reader.

The glossed page of the Bible, created in the 12th century, "contains various texts keyed to the main text in a system of carefully arranged references. This is the medieval equivalent of the hypertext: texts linked up to other texts that the reader can access immediately without even lifting their eyes from the page. There can be up to four texts running simultaneously: the main text, two flanking columns of glosses in continuous prose, as well as a 'discontinuous' gloss of single words or explanatory passages written between the lines of the main text. Glosses often take up the upper and side margins and begin with a paragraph sign (similar to our ¶) to make them easy to locate. The glosses are placed alongside the passage they seek to explain."

A significant contribution to the expansion of the Internet was made by the invention of personal computers (PCs) in the 70-80s of the last century [7]. As a mass-market consumer electronic, PC was intended for individual interactive use, unlike expensive mainframe computers served by a team of professionals. PCs allow individual users to manage HTML pages containing links with addresses of other Web pages or on other Web servers (computers that run websites) worldwide.

The next step in the information progress was the development of cellular (radio) networks distributed over land areas called cells [8]. These networks allow internet connection practically in every place in the world. The fifth generation of mobile networks (5G technology) opens new perspectives in high-speed and high-capacity data communications, allowing customers to build a network connecting virtually everyone [9].

The fast connectivity of all devices, including miniature cellular instruments, handle more traffic at high speeds handles more traffic at high speeds.

A typical example is a sparsely populated table named Cloud Bigtable, which can scale to billions of rows and thousands of columns, enabling the user to store terabytes or even petabytes of data [11]. MATLAB software

package allows accessing and processing a huge data set in the cloud [12].

A novel approach to the development of the Internet was the concept and the term itself of the "Internet of Things" introduced by P. T. Lewis [13]: "The Internet of Things, or IoT, is the integration of people, processes, and technology with connectable devices and sensors to enable remote monitoring, status, manipulation and evaluation of trends of such devices." These devices with unique identifiers may transfer data over a network without requiring human-to-human or human-to-computer interaction [14]. A brilliant example is a Smart Home (home automation using IoT) [15]. "Better understand[ing] customers to deliver enhanced customer service, improve decision-making and increase the value of the business" [14].

In conclusion to this paragraph, we would like to note that the introduction of microelectronics and computer technology has had a significant impact on the development of the instrumental base of analytical spectroscopy and caused revolutionary achievements in the field of mathematical processing of data obtained from analytical instruments [16-18]. Automation systems for chemical analysis (robots) allow for performing three major integrated tasks: sample preparation, measurement of analytical signals, and signal processing [19].

In the last years, miniature spectrometers, based on the achievements of microelectronics, have been developed for portable applications [20, 21]. These devices may be entire systems on a single chip (on-chip). Also, they provided *in situ* optical analysis. For example, an on-chip, low-cost, high-spectral resolution spectrometer uses arrays of photodetectors (complementary metal oxide semiconductor (CMOS) sensors) [22]. A simple attachment to the PC, connected to the USB port, allowed reflectance spectra to be obtained using 18 AMS sensors in the 410-940 nm range [23].

Information and data

We turned to the terms "Information" and "Data" in the previous paragraph. However, the concepts 'Information' and 'Data' often overlap in common usage. Let us consider two simple examples: "Data obtained by analytical instruments allow obtaining valuable information about the chemical composition of the object under study" and "Analytical instruments give valuable information about the chemical composition of the object under study."

Let us try to understand the meaning of these concepts more deeply. The following table reflects definitions of information and data from the Merriam-Webster dictionary [24].

Exercise 1.1-1.

The reader is invited to find a similar use of these concepts by searching the Internet. E.g., "Multidimensional Information Visualization" and "Multidimensional Data Visualization."

Table 1.1-1. Information and data [24].

Information	1. "Knowledge obtained from investigation, study, or instruction. " 2. "…DATA. " 3. "A signal or character (as in a communication system or computer) representing data. " 4. "Something (such as a message, experimental data, or a picture) which justifies change in a construct (such as a plan or theory) that represents physical or mental experience or another construct. " 5. "A quantitative measure of the content of information. " 6. "Specifically: a numerical quantity that measures the uncertainty in the outcome of an experiment to be performed. "
Data	1. "Factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation. " 2. "Information in digital form that can be transmitted or processed. " 3. "Information output by a sensing device or organ that includes both useful and irrelevant or redundant information and must be processed to be meaningful. "

Items 5 and 6 in Table 1.1-1 involve Shannon's definition of the quantitative measure of the information [3] generated by an information source that produces successive symbols (letters) whose probabilities, in the simplest case, depend only on the preceding letter (Appendix A).

Shannon's information, based on the notion of entropy ("the degree of disorder or uncertainty in a system" [25]), may be understood qualitative in the Indian logic concepts as reducing the uncertainty of "knowing something by knowing something else" [26].

Two data categories used for day-to-day transactions and analytical purposes are Operational and Analytical. Analytical data is usually stored in Online Analytical Processing (OLAP) systems and databases (Fig. 1.1-1) [27].

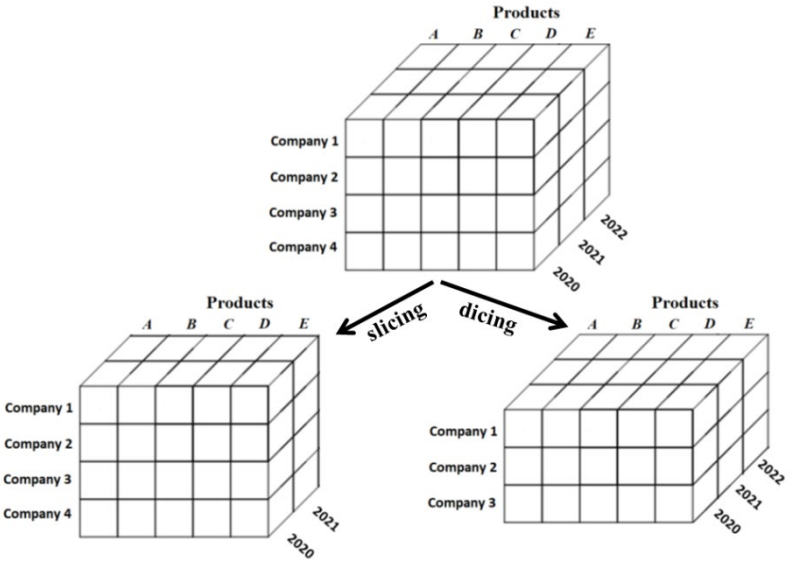

Figure 1.1-1. Online Analytical Processing System.

The core of any OLAP system is a "high-dimensional cube" (Hypercube).

The slice and dice procedures break a body of information into smaller pieces (views) or examine it from different perspectives for better understanding. A database consists of linked tables in various forms (cubes). The mathematical processing of cubes allows for observing a cross-section of data from different angles of an analyst's view.

A roll-up consolidation is aggregating information in one or more measures. For example, all sales of some companies' products are "rolled up" to the country's sales to forecast the product sales trends.

Typical applications of OLAP include business reporting for sales, marketing, management, reporting and business process management (BPM), budgeting and forecasting, financial reporting, and similar applications, with new applications approaching, such as agriculture.

Scientific data [28]

A huge amount of scientific data created in the information age presents special requirements for their processing, storage, and transmission:

- Efficient processing of huge amounts of data at many levels of detail.
- "Managing provenance, quality, proper attribution, and citation."
- Discovery, collection, and integration of relevant data from sources with different scientific significance, spatial and temporal scales, and others);
- Sharing data between multiple communities.

All these topics involve the problem of data formatting. A data format is a model of data representation in a digital store.

Binary and text formats map data concepts to bits (bytes) and character strings, respectively. Data formatting requires a description of the data ('metadata'), e.g., size of numbers, encoding scheme, and others.

Differentiation and integration of information flows

The most important mechanisms for the development of society are the differentiation and integration of information flows with a view to their application for subsequent implementation, both to deepen our knowledge about the world around us and for practical use in producing material goods. These mechanisms manifest themselves in different forms and with varying degrees of intensity, in fact, throughout the existence of society.

Differentiation of the accumulation of information flows is manifested, first of all, in the existence of a vast and ever-increasing number of unique, often heterogeneous areas of knowledge, which leads to a progressive narrow specialization of workers in the research sector of society. This factor, in turn, complicates communication between specialists working even in relatively close areas of science and technology. The situation becomes similar to the construction of the Tower of Babel, which was interrupted by a higher power that forced people to speak different languages.

The narrowing of the area of specialization can create labor shortage problems since the number of workers in various fields is limited.

Some problems of the increasing specialization in chemistry for the student, teachers, and research workers have already been raised more than 70 years ago [29].

However, along with the differentiation of the accumulation of information flows by various highly specialized consumers, one can observe their integration. The integration is due to the emergence of new scientific disciplines and directions at the junction of existing ones, the creation of general approaches (even theories), and the discovered patterns inherent in many sciences.

Exercise 1.1-2.

The reader is invited to illustrate the above section with practical examples.

Human-Data Interaction

In human-computer interaction, studies have focused on the interactions between humans and computerized devices; in Human-Data Interaction (HDI), the human takes the center of the data flow (Fig. 1.1-2) [30, 31].

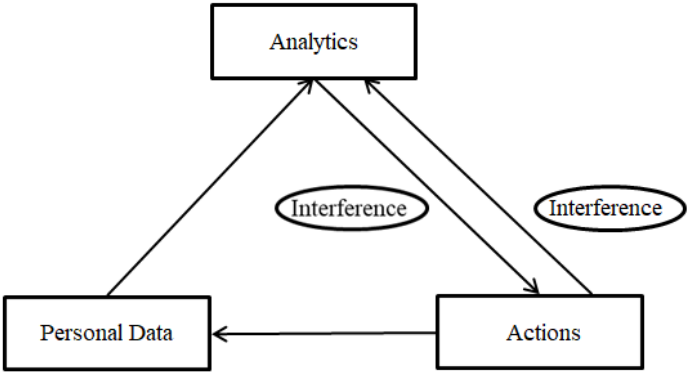

Figure 1.1-2. Human-Data Interaction (adapted from [30]). The inferences on the analytics algorithms by personal data change human actions in a complex way.

HDI was the subject of numerous studies [30]. This problem has been considered from different angles, the content of which often goes beyond the material presented in this chapter. The interested reader can get acquainted with various clever definitions related to the subtleties of the HDI. Let us mention only one subtitle of the review [30]: *"The Paradox of Privacy: The More We Reveal, the More Privacy We Desire."*

Indeed, Internet users are often forced to disclose their data to sales agents, employers, law firms, medical institutions, publishing houses, and other private and public organizations. Since privacy management is highly confidential, the more people reveal information on social media, the more privacy they want. They continue to publish their private data because they fear they will not be noticed!

The importance of tight control over our portfolios of personal data also relates to the trading of personal data between third parties.

How Much Data is there in the World?

The capacity of the storage units used in computer technology is based on the byte equal to the eight binary units (bit). 1 bit designates a memory cell that may contain 0 or 1. One byte is the amount of storage required to store a single character of English text. Universal coding standard UTF-8 includes 4 bytes which can represent 2^{32} different symbols.

Table 1.2 shows units derived from byte [32].

Table 1.1-2. Units derived from byte [32].

Metric	Value	Bytes
Byte (B)	1	1
Kilobyte (KB)	$1{,}024^1$	1,024
Megabyte (MB)	$1{,}024^2$	1,048,576
Gigabyte (GB)	$1{,}024^3$	1,073,741,824
Terabyte (TB)	$1{,}024^4$	1,099,511,627,776
Petabyte (PB)	$1{,}024^5$	1,125,899,906,842,624
Exabyte (EB)	$1{,}024^6$	1,152,921,504,606,846,976
Zettabyte (ZB)	$1{,}024^7$	1,180,591,620,717,411,303,424
Yottabyte (YB)	$1{,}024^8$	1,208,925,819,614,629,174,706,176

However, these incredibly huge numbers do not say anything by themselves. To understand them, let us estimate the memory required to store huge data.

For example, the number of classified print books in “The Library of Congress” as of 31.12. 2015 was 24,055,745 [33]. Therefore, the volume is 24,055,745x5MB/book ≈120 TB. Storing the other 120 million items in this library will need tens of terabytes. The amount of memory in all US academic research libraries is approximately 2 petabytes.

The market intelligence company (the “Global Datasphere”) predicts that global data creation and replication will experience a compound annual growth rate of 23% over the period 2021-2025 [34]. These values were: 6.5 ZP in 2012, 64.2 ZB in 2020, and the expected data is 181 ZB in 2025. Readers can understand these values because a single zettabyte equals about 250 billion DVDs [30]. If every person in the world begins to fill 250 billion x 181 DVDs from the Internet, it would still take approximately 100 days [35].

Table 1.1-3 shows that the world is filled with incredible digital information, even for partial storage, of which huge memory resources are needed [36].

Exercise 1.1-3.

Try roughly estimating how much personal data you generate in an average day (e.g., email messages, files for transfer, printed documents, photos, phone calls, and banking transactions).

Table 1.1-3. Activities in different information media [36, 37].

Origin	Activities
Internet	[3.5 (Google) + 1.5 (another search engines)] x 10^9 searches per day
Internet of Things	2 x10^{11} devices, including 33x10^6 voice-first devices (i/o interfaces)
Facebook	32 x 10^9 active users daily
Instagram	4 x 10^8 million active users daily; posted ≈49,000 pictures/min
Twitter	456,000 tweets/min
LinkedIn	120 new users/min
Skype	≈ 154,000 calls/min
Spam emails	103 x 10^6/min
Digital Photos	4.7 x 10^{12} are stored
Wikipedia	600 new page edits/min

Physiological abilities of a person and information perception

The above fantastic growth amount of information circulating in the world raises an obvious question:" What are the physiological abilities of a person in terms of information perception?" Alternatively, in other words: "how messages are handled inside human beings" [39].

It was found that during "intelligent" or "conscious" activities (e.g., reading), the human nervous system transfers electrical pulses through the nervous channels producing a maximum of 50 bits per second [38]. However, our senses can gather millions of bits per second from the environment (Table 1.1-4). Therefore, our brain demonstrates huge compression ratios up to $10^7/50=2 \times 10^5$.

Table 1.1-4. Information transmission rates of the senses [39].

Sense	Bits/ sec.
Eyes	10^7
Skin	10^6
Ears	10^5
Smell	10^5
Taste	10^3

These compression ratios pose several problems:

1. What is a distortion of the original data due to the compression?
2. What is the processing time of the compression?
3. What is the human "hardware" used for compression?

Generally, "much of human learning, perception, and cognition may be understood as information compression and often more specifically as "information compression via the matching and unification of patterns (ICMUP). Basic ICMUP means lossy compression of information." [40]. Discussion of this problem is out of our main text. The interested reader is referred to the review [40].

According to [39], the time required for processing and compressing sensory input is approximately 0.5 sec. The answer to the third question is the net of the brain which contains 100 billion cells; each with connections to thousands of other brain cells executes more than 100 billion operations per second [39].

Information Crisis

A huge number of circulating information reduces the role of modern man to the simple function of absorbing information. Sets of information and data become the essence of a person.

A novel philological-philosophical analysis of the notation of 'Information' [41] defined this concept as "something that at the process level allows you to make a variety of transitions, and also displays reality using codes and signs, while not being an object itself, but allowing information to be exchanged between objects." Therefore, we can say that the world is filled with a huge amount of related data, some of which are false. False data can generate crises in society. "Cleaning" information from false connectivity is the primary step in data processing.

The authors of the scientific report [42] formulated "The Five Giant Evils of the information crisis:"

- *Confusion*. "What is true and whom to believe?" It is increased by 'information pollution at a global scale.'
- *Cynicism*. "Citizens are losing trust, even in trustworthy sources."
- *Fragmentation* "means that the pool of agreed facts on which to base societal choices is diminishing."
- *Irresponsibility* arises because 'informational power' belongs to organizations (including state ones) with no developed ethical code of responsibility.
- *Apathy*, "whereby citizens disengage from society and begin to lose faith in democracy."

Exercise 1.1-4.

The reader is invited to illustrate these “Evils” through real-life examples.

Scientific research also suffers from false information: from hyperbole to publication, bias and misquoting, predatory publications, and filter bubbles (or echo chambers) (“what we want to hear and not always what we need to know”) [43].

The interested reader is referred to the article [44] dedicated to the problems of false information and fake science.

In conclusion, the author wants to express his admiration for the successful launch of ESA’s Euclid space telescope (01.07.2023) for a six-year mission to chart the largest-ever map of the universe. The spectral instruments of the telescope accumulating about 520 GB of visible and 290 GB of near-infrared data every day [45] will allow us to expand our knowledge of the universe’s structure to an extraordinary extent.

CHAPTER TWO

DATA PROCESSING

Numerous references contain slightly different definitions of data processing. Here we reproduce a modified definition [1]: "Data processing is the method of collecting raw data from the data sources and translating it into a form that contains valuable information for further use."

Data processing (DP) is a part of data management (DM) which describes tools and methods to organize, sort, and process static datasets. DP enables real-time processing of data streams generated by sensors, instruments, and simulation experiments.

Due to the textbook's content, what follows is also concerned with the mathematical processing of spectral data in analytical chemistry [2].

DP is inextricably linked with the organization of data stored in databases (DB). DBs have gone through a long evolution from small spreadsheets in the 1980s to relational SQL-DBs containing large two-dimensional tables linked by an identifying descriptor and to NoSQL DBs (Appendix B).

DP depends on data types which are classified in different ways. Here, let us consider data types as they are usually presented to students in courses on the use of computer programs (e.g., SPSS for statistical calculations) (Fig. 1.2-1) [3, 4]. This classification differs from data types of the variables declared by computer programs to restrict the type of data to be stored, e.g., character, integer, floating point, and others.

Structured data are organized in a special format, e.g., transaction data of banks, spreadsheets, and a multidimensional array of data-OLAP (online analytical processing data cubes). These structures are widely used in chemistry databases. For example, the axes of a three-dimensional OLAP cube are atoms by number, time, and protein structure [5].

Semi-structured data are usually saved in textual files to enable parsing, e.g., XML files (see Appendix C).

Quasi-structured data are textual data that, in principle, can be formatted using some methods, e.g., data about webpages a user visited, their order, formats, and other information.

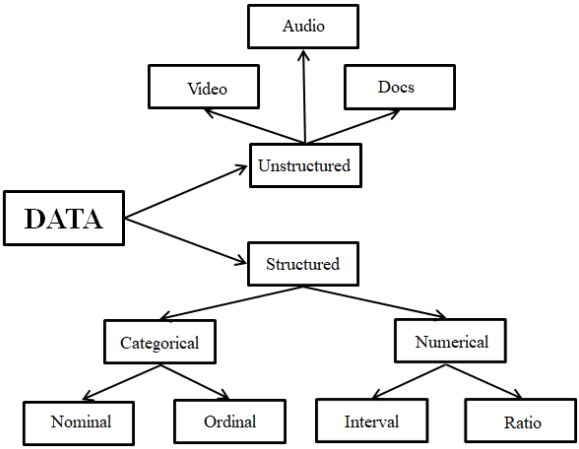

Figure 1.2-1. Data types.

In practice, unstructured type includes semi- and quasi-structured data.

Our modern world of data is filled with unstructured data: files of texts, images, audio records, films, and others. Usually, these data are not stored in a structured database format, simply in the tabular form in a relational SQL database. Unstructured data are managed using the NoSQL database (Appendix B).

Categorical data describes groups of non-ranking nominal data and ordered (in a quality sense) ordinal data. Corresponding examples are gender and education level (primary school, high school, BS, MS, PhD.).

Numerical data includes values in an interval scale with no true zero (e.g., the temperature in Celsius degrees) and a ratio scale with a real zero (e.g., the temperature in Kelvin degrees).

However, the non-zero ratio data may be only positive (e.g., age).

Exercise 1.2-1.

The reader is invited to give examples of different data types.

For further discussion, it is necessary to answer the question of what are the specific properties of databases (DB) in analytical spectroscopy (e.g., [6-8). Table 1.2-1 represents some technical details of these relational DB, which includes the text and image files used only as reference data.

(Spectral databases are the subject of Chapter 4.1).

So, we conclude that managing spectral DB does not principally differ from that of another relational DB. The major difference is possibly preprocessing the one- and multi-dimensional spectra (see below).

Before answering the question, “How does the type of data affect their processing” we will briefly consider the principles of DP.

Fig. 1.2-2 is a general diagram of DP. A raw data source may generate unstructured and structured data. In analytical applications, it is an object (sample) under study. However, we suppose the analytical data management system is supplied with additional information, such as sample properties, measurement conditions, staff, references, etc. The origin of this information is not essential at this stage of our discussion.

The source itself may originate a raw analytical signal (e.g., a star), or the spectral device measures some spectral characteristics of an object by influencing it in a special way (e.g., by heating or irradiation). However, in this measurement process, data is not simply collected, but it can be subjected to various actions, such as de-cluttering and removing unexpected signals (outliers). Also, compressed sensing allows us to collect sparse data economically, which requires much fewer resources to store it [10].

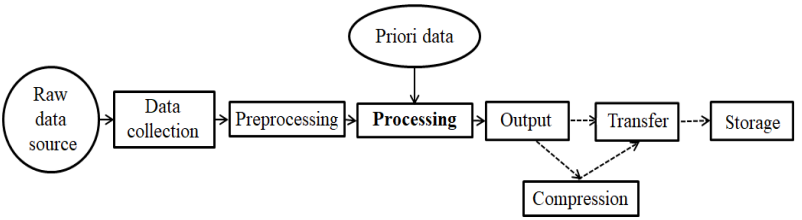

Figure 1.2-2. Simplified block diagram of data processing.

The second preprocessing step, called in data science “Data cleaning (cleansing)” [1], aims to remove “bad” data (redundant, incomplete, or incorrect). However, the question remains: “How can one determine the unsuitability of some data without a full analysis of the problem under solution”?

Data redundancy problems may appear if the same data is located in more than one cell within database tables [3, 4]. Table 1.2-2 illustrates the problem.

A solution is decomposing this complex relation into two independent tables describing universities and students separately. However, one must define links between these tables. Since this topic is out of the book subject, we refer the interested reader to the textbook [3].

Data accuracy must be checked in each cell and inside all rows and columns. For example, does the value “female” of the attribute gender correspond to the correct name? [4].