Chapter 30
Visualization Tools for Big Data Analytics in Quantitative Chemical Analysis:
A Tutorial in Chemometrics

Gerard G. Dumancas
Louisiana State University – Alexandria, USA

Lakshmi Chockalingam Kasi Viswanath
Oklahoma Baptist University, USA

Ghalib A. Bello
Icahn School of Medicine at Mount Sinai, USA

Casey O'Neal Orndorff
Louisiana State University – Alexandria, USA

Jeff Hughes
RMIT University, Australia

Glenda Fe Dumancas
Louisiana State University – Alexandria, USA

Renita Murimi
Oklahoma Baptist University, USA

Jacy D. O'Dell
Oklahoma Baptist University, USA

ABSTRACT

Modern instruments have the capacity to generate and store enormous volumes of data and the challenges involved in processing, analyzing and visualizing this data are well recognized. The field of Chemometrics (a subspecialty of Analytical Chemistry) grew out of efforts to develop a toolbox of statistical and computer applications for data processing and analysis. This chapter will discuss key concepts of Big Data Analytics within the context of Analytical Chemistry. The chapter will devote particular emphasis on preprocessing techniques, statistical and Machine Learning methodology for data mining and analysis, tools for big data visualization and state-of-the-art applications for data storage. Various statistical techniques used for the analysis of Big Data in Chemometrics are introduced. This chapter also gives an overview of computational tools for Big Data Analytics for Analytical Chemistry. The chapter concludes with the discussion of latest platforms and programming tools for Big Data storage like Hadoop, Apache Hive, Spark, Google Bigtable, and more.

DOI: 10.4018/978-1-5225-3142-5.ch030

ANALYTICAL CHEMISTRY AND CHEMOMETRICS

Over the years, various Chemometric tools have emerged and have been utilized as data evaluation instruments generated by various hyphenated analytical techniques including their application since its advent today (Kumar, Bansal, Sarma & Rawal, 2014). Although its primary applications are geared toward Multicomponent Analysis, its applications have even been extended to the area of genetic epidemiology and Bioinformatics in the recent years (Dumancas, 2012; Dumancas et. al., 2014; Dumancas et. al., 2015).

The advances that are now visible in Process Analytical Technology (PAT) in Chemometrics can be attributed to the rapid development of both analytical instrumentation and mathematical methods involved in multivariate data analysis (Bogomolov, 2011; Dubrovkin, 2014; Kessler, 2013; Pomerantsev & Rodionova, 2012). Specifically, the rapid growth of a wide multitude of novel analytical methods and the continuous expansion in the area of their applications are the two driving forces that led to the success of PAT (Dubrovkin, 2014).

With the vast array of information emanating from various analytical instruments comes the challenge of processing these data in a rapid fashion. Thus, the process of Data Fusion, a subclass of Chemometrics is now considered an important topic (Esteban et. al., 2005; Ovalles & Rechsteiner, Jr., 2015). Data Fusion simply refers to the integration of data and knowledge from several sources (e.g. analytical instruments) (Castanedo, 2013). Many other definitions for data fusion exist in the literature. It is defined by the Joint Directors of Laboratories (JDL) as a "multi-level, multifaceted process handling the automatic detection, association, correlation, estimation, and combination of data and information from several sources" (Steinberg et. al., 1999). The corresponding informational models from data fusion should simulate extremely complex problems by fitting to the massive amount of empirical semi-structured and unstructured data (Isaeva et. al., 2012). Consequently, the algorithmic support and the interface of a computerized analytical system (often with limited computer resources) should be adjustable to systems with features of new types. Such challenge arising from analytical information management led to several perspective solutions such as the concept of Cloud Computing all of which is part of the development of "Big Data Approach" (BDA) (Dubrovkin, 2014).

In this chapter, the major aspects of Big Data utilization and processing in Analytical Chemistry (Chemometrics) will be discussed. Specifically, some commonly used algorithmic and instrumental techniques and aspects of computerized analytical systems will be discussed.

APPLICATIONS OF CHEMOMETRICS

Chemometrics is a fast spreading area which has many avenues of applications in both descriptive and predictive problems in experimental life sciences especially in Chemistry. It is considered to be a highly interfacial discipline employing Multivariate Statistics, Computer Science and Applied Mathematics using methods employed in core data analytics with the ultimate goal of addressing problems in Biochemistry, Medicine, Chemistry, Chemical Engineering and Biology (Khanmohammadi, 2014).

The biological and medical applications of Chemometrics encompass a wide area of expertise. Support Vector Machines (SVMs), Partial Least Squares Discriminant Analysis (PLS-DA) are widely used