Chapter 4
Tools, Technologies, and Methodologies to Support Data Science:
Support Technologies for Data Science

Ricardo A. Barrera-Cámara
https://orcid.org/0000-0002-3170-4671
Universidad Autónoma del Carmen, Mexico

Ana Canepa-Saenz
https://orcid.org/0000-0003-0583-439X
Universidad Autónoma del Carmen, Mexico

Jorge A. Ruiz-Vanoye
https://orcid.org/0000-0003-4928-5716
Universidad Politécnica de Pachuca, Mexico

Alejandro Fuentes-Penna
https://orcid.org/0000-0002-4303-3852
Centro Interdisciplinario de Investigación y Docencia en Educación Técnica, Mexico

Miguel Ángel Ruiz-Jaimes
https://orcid.org/0000-0002-2585-9896
Universidad Politécnica de Morelos, Mexico

Maria Beatriz Bernábe-Loranca
https://orcid.org/0000-0003-3014-4139
Benemérita Universidad Autónoma de Puebla, Mexico

ABSTRACT

Various devices such as smart phones, computers, tablets, biomedical equipment, sports equipment, and information systems generate a large amount of data and useful information in transactional information systems. However, these generate information that may not be perceptible or analyzed adequately for decision-making. There are technology, tools, algorithms, models that support analysis, visualization, learning, and prediction. Data science involves techniques, methods to abstract knowledge generated through diverse sources. It combines fields such as statistics, machine learning, data mining, visualization, and predictive analysis. This chapter aims to be a guide regarding applicable statistical and computational tools in data science.

DOI: 10.4018/978-1-7998-3053-5.ch004

Data science was initially proposed as a set of areas with a technical point of view made up by operation research, data modelling and data methods, pedagogy, tool evaluation and theory (Cleveland, 2001). Data science encompasses mathematics, automated learning, artificial intelligence, statistics, databases and optimization (Dhar, 2013).

All activities related to data science professionals are classified as follows (Donoho, 2017): 1. Data collection, preparation and exploration, 2. Data representation and transformation, 3. Data calculation, 4. Data modelling, 5. Data visualization and presentation, 6. Science on data science. Furthermore, different roles or professionals with profiles and skills linked to data science have emerged (Government of Spain, 2018) (UC Regents, 2019): 1. Data scientist. A professional with ability to extract, clean and present data through exploration. These professionals aim to find unanswered questions and the data required to answer them. 2. Data analyst. These professionals act as a liaison between a data scientist and a business analyst. They translate the technical analysis into qualitative type data elements and communicate results. 3. Data engineer. These professionals design set up and administer the necessary infrastructure required for the transformation and transfer of data for inquiry.

This chapter is organized in different sections, which are the Related Works section, data analysis, Data Visualization, Dataset, Project Managements in data Science, Data Science Platforms, Machine Learning, Future Research Directions, and finally the Conclusion section.

BACKGROUND

Related Works

This section presents some works related to applicable technology applications in data science. Platforms: Performs an analysis of hardware platforms considering specific features and software framework used in them, as critical elements that must be present for the execution of big data algorithms (Singh & Reddy, 2014); Learning Machine: Some criteria are proposed and analyzed for the selection of opens source tools for learning machine with big data. The experience of processing, libraries and machine learning framework is also considered (Landset et al., 2015); Software: Open source data mining tools are analyzed considering their operational characteristics, license, programming languages, web support, type, domain that are also used in data science (Barlas, 2015); Vizualization: Various tools and techniques of data visualization oriented to large volumes of data are analyzed, presenting their functional and non-functional characteristics (Caldarola & Rinaldi, 2017); Dataset: The availability of data, exchange, access, use recovery, searches make possible the emergence of data stores or data sets available in public access dataset services but from a company with information search services on the internet (Chapman et al., 2019).

In Figure 1, presents a timeline related to the year of launch of the technologies identified in the background of this work.

Data Analysis

Data analysis (Izabella et al., 2019) is the process of inspection, cleaning, transformation and modelling of data with the purpose of finding useful information, reporting conclusions and providing ground for