Data Literacy: 
Developing Skills on Exploring Big Data Applications

Dimitar Christozov, American University in Bulgaria, Blagoevgrad, Bulgaria
Katia Rasheva-Yordanova, University of Library Studies and Information Technologies, Sofia, Bulgaria

ABSTRACT

The article shares the authors’ experiences in training bachelor-level students to explore Big Data applications in solving nowadays problems. The article discusses curriculum issues and pedagogical techniques connected to developing Big Data competencies. The following objectives are targeted: The importance and impact of making rational, data driven decisions in the Big Data era; Complexity of developing and exploring a Big Data Application in solving real life problems; Learning skills to adopt and explore emerging technologies; and Knowledge and skills to interpret and communicate results of data analysis via combining domain knowledge with system expertise. The curriculum covers: The two general uses of Big Data Analytics Applications, which are well distinguished from the point of view of end-user’s objectives (presenting and visualizing data via aggregation and summarization [data warehousing: data cubes, dash boards, etc.] and learning from Data [data mining techniques]); Organization of Data Sources: distinction of Master Data from Operational Data, in particular; Extract-Transform-Load (ETL) process; and Informing vs. Misinforming, including the issue of over-trust vs. under-trust of obtained analytical results.

KEYWORDS

Big Data, Competency, Education, Literacy, Literacy Divide

1. INTRODUCTION

One way to understand the evolution of civilizations is to study the way people solve problems connected to the use of data and information. Every stage of human history is marked by specific ways of exploring facts, which involve learning from collected data and preserving and disseminating the acquired knowledge. Examples like Stonehenge or Talmud are perfect illustrations of this concept. During almost the entirety of human history the individual amount of data needed to be grasped has depended on personal cognitive capacity. This limitation inflicted considerable influence on the kind of data which is selected and stored and the way it is presented. Till the middle of the last century all recorded data passed a careful screening, verification and editing. The revolution of data processing, made by the introduction of computer technology, which merged a bit later with communication technologies, changed radically the way data is being handled.

DOI: 10.4018/IJDLDC.2017040102
When making a parallel with another important technological revolution - that of energy conversion - we can see how the tremendous impact of data processing technology to all kinds of human activity. But we can also observe certain side effects, which result from the rapid penetration of data processing technologies in all facets of human life. One of these side effects is pollution. Tremendous growth of energy conversion technologies and their utilization, such as the thermal power plants or nuclear power stations, gave rise to numerous ecological problems. Availability of technologies for data storing and data processing resulted in availability of (1) enormous amount of data, (2) data, which is not always verified and trustworthy and (3) which is constantly changed and updated. This phenomenon is marked today by the term “Big Data”. As Gartner (see Gartner IT Glossary) defines it “Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing which enable enhanced insight, decision making, and process automation”.

Big Data phenomenon sets a new line of division between people according to their literacy and the competencies required nowadays, similar to the competences to make use of electricity and electrical devices. A new aspect to the problem known as “digital divide” has appeared – the division based on exploring Big Data. We can define Big Data Exploration as (1) ability to search, identify and retrieve data, relevant to a given problem, (2) ability to use different techniques to verify reliability and relevance of obtained data, and (3) ability to use different techniques to represent huge amount of data in a meaningful and comparable with one’s cognitive capacity way and to understand specific limitations, requirements for applicability, and quality of information generated through these techniques. Those competences help to understand data represented implicit properties of objects or events and enhance decision making (further we will refer to them as “entities attributes”). These abilities, define the major aspects of Big Data literacy, and are essential to business entities and individual citizens and their survival in the current globalized world. From this perspective Big Data literacy can be considered as one of the key components of “information literacy” (Girard, Klein & Berg, 2015, p. 162).

Learning from Big Data faces significant difficulties. The major one comes from the inability to observe directly the entire set of entities properties, because of their volume. They can be observed only via summarizing statistics. Validity of information obtained depends on whether data can satisfy particular set of requirements, for example, whether different parameters are mutually independent. Proving independence usually is a tricky problem and assuming independence without proof may result in misleading and wrong decisions. The three categories of requirements in this respect are: (1) to know what requirements must be met by data in order to obtain valid results by applying certain statistical technique; (2) to possess skills necessary to check whether data satisfies those requirements; and (3) to understand what impact the unsatisfied, or partially satisfied, requirements have on obtained results and to be able to map this understanding to the problem, which needs solution. In other words, effective exploration of Big Data makes it necessary for the user to possess deep knowledge of statistics, skills to apply statistical methods by using sophisticated software, and extensive domain knowledge. And also, the requisite use of computer technology.

Today all these competences are indispensable to educated professionals in every field and combining all those aspects of training create significant challenges for the universities. This paper shares authors’ experience in introducing courses related to developing Big Data competences at undergraduate level of college education. Usually such training is offered within master programs to students possessing background, experience and maturity, which contribute to their understanding of the complexity of problems and helps the learning process. For the authors, it is clear that including special Data Science courses in the curriculum of a particular major cannot solve the above problems.
Synthesizing Technological and Pedagogical Knowledge in Learning Design: A Case Study in Teacher Training on Technology Enhanced Learning
[www.igi-global.com/article/synthesizing-technological-and-pedagogical-knowledge-in-learning-design/152606?camid=4v1a](www.igi-global.com/article/synthesizing-technological-and-pedagogical-knowledge-in-learning-design/152606?camid=4v1a)

Beyond Angry Birds™: Using Web-Based Tools to Engage Learners and Promote Inquiry in STEM Learning
[www.igi-global.com/chapter/beyond-angry-birds/188955?camid=4v1a](www.igi-global.com/chapter/beyond-angry-birds/188955?camid=4v1a)