**A NEW FRAMEWORK FOR STATISTICAL THINKING IN TIMES OF BIG DATA AND DIGITAL ECONOMY**

**ABSTRACT**

The interest in big data is growing exponentially in today’s society. Commercial insights, government initiatives and even research calls, all seem to be focused on exploiting the potential of technology to capture and analyze massive amounts of data in increasingly powerful ways. Big data, that is, data that are too big for standard database software to process, is everywhere. For some, big data represents a paradigm shift in the ways that we understand and study our world, and at the very least it is seen as a way to better utilize and creatively analyze information for public and private benefit. The concept of big data “refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze” (Manyika et al., 2011, p.1). Additionally, big data is often associated with key characteristics that go beyond the question of size, namely the 5 Vs: volume, velocity, variety, veracity and value (Storey & Song, 2017). Big data is dispersed among various platforms that operate with different standards, providers and degrees of access (Ferguson, 2012). For example, a lot of work in big data focuses on Twitter, the blogosphere, and search engine queries. All of these activities are not undertaken equally by the whole population, which raises concerning issues around the question of whose data traces will be analyzed using big data. There are also a number of practical issues related to working with big data. These include, among others, issues we cannot afford to ignore, such as implications for the training of future teachers regarding handling and analysis of big data.

**Existing System**

The concept of big data “refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze” (Manyika et al., 2011, p.1). Additionally, big data is often associated with key characteristics that go beyond the question of size, namely the 5 Vs: volume, velocity, variety, veracity and value (Storey & Song, 2017). Big data is dispersed among various platforms that operate with different standards, providers and degrees of access (Ferguson, 2012). For example, a lot of work in big data focuses on Twitter, the blogosphere, and search engine queries. All of these activities are not undertaken equally by the whole population, which raises concerning issues around the question of whose data traces will be analyzed using big data.

**Disadvantages**

1. these activities are not undertaken equally by the whole population,
2. We cannot afford to ignore, such as implications for the training of future teachers regarding handling and analysis of big data.

**Proposed System**

Many researchers (e.g., Wild & Pfannkuch, 1999; delMas, 2002; Watson, 2017) consider statistical thinking as the practice of statistics through the enactment of the different thought processes involved in statistical problem solving and statistical investigations. For us, in this digital era, statistical thinking processes do not follow the Problem-Plan-Data-Analysis-Conclusion (PPDAC) cycle (Wild & Pfannkuch, 1999) anymore, due to the shift in the way we work with data set by the arrival of big data analytics. In fact, the PPDAC cycle is a question-then-answer research method, focused on data gathered for a purpose using planned processes, chosen on statistical grounds to justify certain types of inferences and conclusions. However, in times of big data, this is actually a weakness of the PPDAC cycle, because most of the data available is opportunistic (happenstance or “found”) data (including “big data”): huge amounts of data already collected by others and hosted somewhere. Nowadays, many companies have data teams exploring large sets of raw opportunistic data, looking for new connections and identifying significant correlations, while refining their analysis until they arrive at valuable understandings. This approach reverses the question-then-answer process of the PPDAC cycle. It starts with strong, data-first answers, and then works backward to find the questions that should have been asked.

**Advantages**

1. Statistical problem solving and statistical investigations.
2. Huge amounts of data already collected by others and hosted somewhere.

# Hardware Requirements:

# Processor - Pentium –IV

* Speed - 1.1 GHz
* Ram - 256 MB
* Hard Disk - 20 GB
* Key Board - Standard Windows Keyboard
* Mouse - Two or Three Button Mouse
* Monitor - SVGA

**Software Requirements:**

* Operating System - Windows XP
* Coding Language - Java