EUROPE ATP 2015
Dublin, Ireland

The following white paper is related to a session at the 2015 E-ATP Gaining Advantage through Assessment Conference

Session Title:

Five Innovative Practices to Manage and Work with Big Data for the European Testing Industry

Session Presenter:

Manfred Straehle

Assessment, Education & Research Experts (AERE)

Big Data in the Testing Industry

Manny M. Straehle, Ph.D., GISF
Founder and President
manny@aerexperts.com

Big data is a term that has slipped into our daily conversations, but what does it really mean? And why does it matter, especially to the testing industry? This white paper will define big data and provide various examples of using big data in a meaningful way. It will also provide steps for working with and managing big data in the testing industry, discuss the importance of small data, and highlight the path to using data in a more efficient and meaningful way by focusing on understanding the research question.

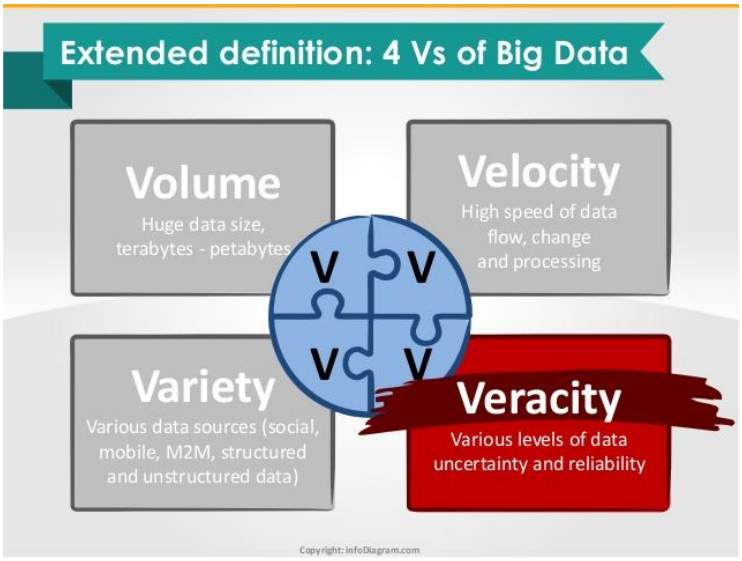
It should be noted that this white paper focuses on the definition and use of data, and will not discuss the technical server approaches in managing big data.

What is Big Data?

Since 1944, there have been many references to the concept of managing and working with large date sets. It was not until October 1997 at the *Proceedings of the IEEE 8th conference on Visualization*[1], when a few researchers presented on a lack of disk space when attempting to manage visual type data, that the term **Big Data** was first introduced. Initially this term referred to a hard disk space issue but it has evolved to using large data sets to intelligently guide and make decisions.

Today, most researchers define Big Data using the 4 Vs, defined the figure below. [2]

[1] https://www.evl.uic.edu/cavern/rg/20040525_renambot/Viz/parallel_volviz/paging_outofcore_viz97.pdf
[2] http://www.slideshare.net/infoDiagram/big-data-cloudappsvisualiconpptinfodiagramtoolbox



Where Does Big Data Come From?

Often big data is often collected electronically through sources which may include social media, banking, retail, healthcare, financial data and so forth. This data comes from activities which leave what researchers call a digital footprint.

In the testing industry, data can come from candidate applications, testing data, etc. A list of potential data sources and purposes can be found below.

Data Sets	Data Purposes
• Registration Data • Exam Preparation Data • Psychometric Data • Recertification Data • Marketing/Financial Data (Volume) • Other Data (e.g., Number of individuals taking courses related to exam)	• Volume • Financial • Satisfaction • Competency • Validity • Test Security

What Are Common Ways of Dealing With Big Data?

How Should I Approach Big Data?

There are essentially two ways to approach big data: exploratory and intentional.

Exploratory. One way to approach data is to consistently explore, harvest, mine, and analyze data with the hope that a meaningful relationship will occur among the many variables. This approach is usually very time consuming and costly and in many cases ineffective. This approach to data relies on chaos theory. Chaos theory essentially assumes "that randomness and determinism become somewhat compatible because of the long-term unpredictability." (Oestreicherp, 2007, p. 232). [3] The classic example of this is using various data sets to identify predictors of weather – by looking at data, one could infer that a butterfly in Brazil cause a tornado in Texas (referred to as the Butterfly Effect).

Intentional. Another approach is to manage and work with the data by using intention. Forming an operational research question allows researchers to use big data more effectively, efficiently, and more cost effectively. A helpful way to begin this approach is to ask stakeholders, **"At the end of the day, what question do you want the data to answer?"** The three examples in the section below use the intentional approach.

[3] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3202497/

What Are Some Examples of Using Big Data?

1. Personality Testing

In the testing industry, I/O and clinical psychologists have been relying on modern day personality testing initially developed by Sir Francis Galton to describe traits using adjectives from dictionaries in order to make critical and life impacting decisions. Consequently, many personality assessments used today like the 16PF and MMPI have consistently relied on a large volume of data to accurately assess an individual's personality. In fact, Cattell, who developed the 16PF was an early pioneer in conceptually thinking of Big Data when seeking to describe an individual's behavior. He described individuals' personality by developing these three categories of data:

- L-Data – Life Data
    - Include collecting data from everyday life events such as driving record, travel history, and grade point average
- T-Data – Experimental Data
    - Objective tests - measuring somebody's reaction in lab such as the George Miller studies on legitimate authority and pretending to administer an electrical shock to another person
- Q-Data – Questionnaire Data
    - Information obtained from questionnaires like Cattell's 16PF

2. Netflix and "House of Cards"

Netflix executives were interested in developing a successful television series, so they turned to data. They examined large sets of their own data such as members' review ratings and viewing and selection behaviors. The data showed that Kevin Spacey received high ratings, individuals were increasingly streaming video on electronic devices, and the British version of "House of Cards" did well. Informed with this data, the executives developed a streaming version "House of Cards" for American audiences, starring Kevin Spacey. The show was an instant success.

3. Meetings and Wasting Time

VoloMetrix is an organization that analyzes large data sets related to business and people's behavior. They performed an analysis of employee emails and calendar invites to determine what types of emails and meetings are effective. After analyzing these large data sets, the results indicated that certain behaviors were wasting individuals' time. Specifically, VoloMetrix found that Seagate employees were spending more than 20 hours at meetings per week, and one firm was producing nearly 3,700 emails and taking up 8,000 work hours a year from 228 Seagate

employees.[4] Another study showed that Senior Executives devote 2 days of the week simply attending meetings. Consequently, the data helped to provide specific recommendations to cut back on the types of meetings and emails that wasted the most time.

How Can Testing Organizations Work and Manage Big Data?

AERE staff has had many experiences working with and managing big data. Based on our experiences, we propose a five step approach in working and managing big data which is described in the figure below.

[4] http://www.inc.com/rebecca-borison/these-are-the-five-biggest-time-wasters-at-work.html ; https://hbr.org/2014/05/your-scarcest-resource

What about Small and Medium Data?

While Big Data has been receiving a great deal of media attention over the past few years, a smaller movement has risen to discuss the importance of small and medium data.[5] Small and medium data, including data sets as small as one observation, can be very important in contributing to major discoveries. Some historical examples include Jean Piaget, whose theories on child psychology are based on observations of his own children. Another example includes the serendipitous discovery of penicillin in Fleming's lab. There are countless examples of a single discovery based on a single observation. In fact, there is an entire research methodology based on case studies, proving the value of small data in breakthroughs, insights, and decision-making.

So What Really Matters?

The above white paper was intended to define big data and provide examples, approaches, and ways to manage and work with big data in the testing industry. In our opinion, despite the size of the data, it is important to use data where meaning can be discovered and data relationships can be validated and replicated. What really matters is finding relationships among data variables that answer the questions that stakeholders are seeking.

[5] http://www.forbes.com/sites/mikekavis/2015/02/25/forget-big-data-small-data-is-driving-the-internet-of-things/