Advanced Analytics for Big Data

Stephen Kaisler
SHK and Associates, USA & George Washington University, USA

J. Alberto Espinosa
Kogod School of Business, American University, USA

Frank Armour
Kogod School of Business, American University, USA

William Money
School of Business Administration, George Washington University, USA

INTRODUCTION

Big Data has emerged as one of the most challenging aspects of business data and scientific processing within the past 20 years. A major problem is that although we may be able to collect the data, we often have inadequate means for analyzing it. Kaisler, Armour, Espinosa, and Money (2013) identified some of the issues and challenges associated with using Big Data. INFORMS defines *analytics* as "the process of transforming data into insight for the purpose of making decisions" (2013). It involves formulating specific problems or questions; identifying, gathering and organizing the relevant data; and selecting and applying the appropriate methods, algorithms, heuristics and procedures to solve the problems or answer the questions. Analytics are quantitative and qualitative, linear and non-linear, large and small, numerical versus symbolic, and vary along other dimensions as well.

BACKGROUND

"Big Data" originally meant the volume of data that could not be processed (efficiently) by traditional database methods and tools. The original definition focused on structured data, but most researchers and practitioners realize that most of the world's information resides in unstructured data, largely in the form of text and imagery, both still and video, and in audio. Today, big data refers to data volumes in the range of tens of petabytes (10^{16}) and beyond. Such volumes exceed the capacity of current on-line storage and processing systems. Big Data was originally described by the 3Vs (Laney 2001), but Kaisler, Armour, Espinosa, and Money (2013) have suggest two more.

Up until 30 years ago, simple business models often sufficed for international business. Globalization, brought on by advances in digital technology, information available at our fingertips, and a rapidly changing, even chaotic, international political environment, has up-ended these models. It has increased the diversity and uncertainty in outcomes when complex systems such as financial flows and markets, regional economies and political systems, and transnational threats involving multiple actors are in constant flux.

Traditional big data analysis has focused on data mining. The explosion of social media has provided massive amounts of data that can be analyzed and exploited for a wide variety of purposes. Understanding what patterns exist and determining how to use them has been the primary problem for the past decade. Identifying the relevant information from the plethora of ambiguous and often contradictory data will require sophisticated diagnostic, predictive, and prescriptive analytical methods.

Advanced analytics is the application of multiple analytic methods that address the diversity of big data to provide descriptive results and to yield *actionable* predictive and prescriptive results that facilitate decision-making. Advanced analytics go beyond data mining and statistical processing methods to encompass logic-based methods, qualitative analytics, and non-statistical quantitative methods. Advanced analytics

DOI: 10.4018/978-1-4666-5888-2.ch747

Table 1. Five Vs of Big Data

V	Description
Data Volume	The amount of data collected and available for use. It is estimated that over 2.5 Exabytes (10^{18}) of data are created every day as of 2012 (Wikipedia 2013).
Data Velocity	The rate at which data is accumulated or the speed at which the data arrives, and how quickly it gets purged, how frequently it changes, and how fast it becomes irrelevant or outdated.
Data Variety	The different types of data required for analysis, which can be either *structured*, such as RDF files, databases, and Excel tables or *unstructured*, such as text, audio files, and video.
Data Value	The value derived from processing the data that contributes to decision making and problem solving. A large amount of data may be valueless if it is perishable, late, or imprecise.
Data Veracity	The accuracy, precision and reliability of the data. A data set may have very accurate data with low precision and low reliability based on the collection methods and tools used.

represents a diverse set of techniques and requires new software architectures and application frameworks to solve complex problems. New metrics that focus on contributions to the value of the analysis as a holistic result are required to assess and evaluate the outcomes of advanced analytics.

COMPLEX PROBLEMS

The complexity of today's problems – in science, business, and the social sciences – often requires more than a single analytic to solve. These problems, including the so-called wicked problems (Horst and Rittel 1973, Ritchey 2005), are all about *discovery*, which is not about finding the right answer to a question, but determining what is/are the right question(s) to ask and finding the set of solutions to those questions. The more complex a system, the more likely it is that there is no one "right answer," but a set of answers which depend on the initial boundary conditions. Sometimes, it is not the answer that matters, but the progressive set of decisions made to reach the answer. Variations in the decision sets leading to different answers based on the boundary conditions, the existence of exogenous inputs, and the potential for non-deterministic decision processes are likely to characterize the problem space. A few examples include: what happens when the oil and gas run out? How do we ensure sustainable development and ecosystems in third world regions? How do we develop resilient infrastructures for energy provision? Can we predict the next "Arab Spring"? How do we mitigate shocks in the global economy?

Complex problems are often computationally intensive and expensive – relying on big data, large scale cyber infrastructure, and parallel processing to process and analyze data in a reasonable amount of time. But, numerical computation is only a partial solution. Analysis has evolved from data mining to data discovery to generating actionable intelligence. While the former two have been associated with understanding big data, the latter type of analysis is associated with solving problems through prediction and prescription. Complex problems require an integrated set of analytics to solve them. An *analytic architecture* is a software architecture supporting the integration of a set of analytics (Kaisler 2005). An essential element of an analytic architecture is the ability to infuse domain knowledge to inform and direct the problem-solving process and assess the results.

ANALYTIC TAXONOMIES

Analytics can be classified based on their applicability to different problem spaces. Three taxonomies have received considerable attention. A fourth focuses on their morphology. Each class is representative of a suite of algorithms and methods that can be applied to different types of problems.

Within the past decade, much of the popular literature has revolved around business analytics, such as data mining and statistical techniques, including statistical machine learning. Business analytics are often classified as operational, tactical, and strategic analytic. This taxonomy is based on the applicability