AC17038

Big Data And Data Extraction: Tools For Solving Real World Problems

Vikkie McCarthy
Austin Peay State University

Jerry Plummer
Austin Peay State University

ABSTRACT

Management Information Systems (MIS) and Big Data continue to be of importance in private, public, and governmental organizations. Leveraging MIS technologies and Big Data can help organizations manage crises and produce sustainable solutions to real world problems. The purpose of this paper is to present a theoretical framework for using data available on the internet in order to address a real world problem. The significance of this research is that businesses and organizations have access to a plethora of data and can use data extraction tools. Readily available data from public websites is collected and analyzed by using data searching and extraction techniques such as Webcrawler and Webscraping. This case is one example of how organizations can leverage big data and data mining in decision making processes.

KEYWORDS: Management Information Systems, Big Data, Cloud Computing, Crowdsourcing, Kenya, Food Crisis, WebScraping, Python, ScraperWiki and Wayback Machine.

INTRODUCTION

The advent of the Internet and the following Big Data arrival introduced an entire series of new possibilities to business and government alike; as well as for-profit and non-profit areas Although uses of Big Data in the public eye has led to a negative connotation to its spread and usage; there are many instances of Big Data being used to a "good" end, (Davenport, 2015; Allison, Easterwood, & Power, 2013) Big Data, Information Systems and Cloud Computing can, and are, being used to manage crisis situations (Surdak, & Agarwal, 2014).. "Big data" refers to a data management process that enables organizations to "store, manage, and manipulate" large amounts of data in order to gain insight into problems, (Mellor, 2014). We discuss several ways Big Data is being used today in an effort to successfully reduce or alleviate these areas using Big Data for a variety of situations, including famines, food security, inflation rates, and poverty measurement.

BIG DATA

In 2008, Bryant, Katz and Lazowska argued that "just as search engines have transformed how we access information, other forms of big-data computing can and will transform the activities of companies, scientific research, medical practitioners, and our nation's defense and intelligence operations, (www.cra.org/ccc)." They go on to say that, "we have only begun to see its potential to collect, organize and process data in all walks of life. A modest investment by the federal government could greatly accelerate its development and deployment." Governments, non-governmental organizations and corporations have invested heavily in Big Data and data extraction methods.

Boyd & Crawford (2012) reported that "the era of Big Data has begun." In order to qualify as Big Data, as defined by Boyd and Crowford (2012), several factors must be present. These factors include: 1) Technology – "maximizing computing power in order to gather, analyze, link and compare large data sets;" 2) Analysis – "using large data sets to identify patterns in order to make economic, social, technical and legal claims;" 3) Mythology – a "widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy, (p. 663)" The ability to use Big Data in order to analyze societal ills such as future famines is also dependent on many factors. "Big Data stakes out new terrains of objects, methods of knowing, and definitions of social life," (Kitchin, 2014).

Data Kind is a non-profit organization that is dedicated to using Big Data to change the world. The goal of this organization is to promote social change by leveraging the latest data science research.

> *We bring together top data scientists with leading social change organizations to collaborate on cutting-edge analytics and advanced algorithms to maximize social impact. Our programs build upon one another and are designed to meet organizations where they are. From evening or weekend events to multi-month projects, all are designed to provide social organizations with the pro bono data science innovation team they need to tackle critical humanitarian issues in the fields of education, poverty, health, human rights, the environment and cities (DataKind, n.d...)*

One critical humanitarian issue is food insecurity (Cohen & Garrett, 2010). Food shortages caused by unexpected famine is a major global threat that, in many cases, leads to civil unrest

(Cribb, 2010). In the opinion of some, it is a far greater threat than global warming. In his book The Coming Famine: The Global Food Crisis and What We Can Do to Avoid It (2010), Julian Cribb notes, "The world has ignored the ominous constellation of factors that now make feeding humanity sustainably our most pressing task – even in times of economic and climatic crisis, (p. xi)." Food insecurity is a threat to global security for several reasons. For instance, in his book, Cribb argues that "the non-human environment created by lack of water and food leads to self-mutilating behaviors to include genocide, terrorism, and armed conflict," (p. xi). For this reason, hungry countries are the most likely to provide terrorist recruits or plunge into warfare. Most of the new conflicts take place in Africa, the Middle East, and parts of Asia because of famines and poverty (Cribb, 2008). There is a cycle of violence in these areas that leads to "more hunger, greater deprivation, and more vicious fighting, (Cribb, 2008)." Food security is essential for stable governance and human rights. The United Nations (UN) Food and Agriculture Organization report to the 2008 Committee on Food Security, reported an increase in the number of countries with food crises caused by war and conflict from 2% to 27% during the time period of the 1980's through 2007 (FOA, 2013). In these areas of conflict, food wars occur among the fighting factions. Control of food sources allows one party to gain a strategic advantage over the other. In this way, food is used to reward supporters and punish opponents.

KENYAN 2009 FOOD CRISIS

In 2009, Kenya's drought increased their need for food imports of corn and rice setting the stage for the 2009 Kenyan Hunger Crisis. Food prices were increasing at a rampant pace due to massive crop failure during the drought (Gosling, Warren, Arnell, Good, Caesar, Bernie, Lowe, & Smith, 2011). Even with significant increases in food imports, food insecurity was exacerbated by the ongoing global economic crisis increasing the potential for malnutrition on a country-wide and generational scale. Malnutrition in developing countries such as Kenya is a major problem for its most vulnerable populations. About 40% of the population ranged from infancy to 14 years at the time of the crisis (Cohen & Garrett, 2010) Malnutrition impacts survival, health, well-being and the developmental potential of these children. This, in turn, impedes progress in the least developed countries and has been shown to create potential breeding grounds for civil strife and havens for terrorism (Cohen & Garrett, 2010; Kick, McKinney, & Thompson, 2011).

Kenya is classified as a Low-income Country (LIC) by the World Bank (UN, 2008). Therefore, in order to relieve pressure on the national government and help alleviate fallout from the food crisis, the World Bank intervened with monetary policy. The inflation rate reported during the crisis was 25% (Crush, Hovorka, & Tevera, 2011). High inflation rates have huge economic consequences and can make food unaffordable to vulnerable groups. For these reasons, the World Bank focused its policy on the inflation data. The question then becomes, "is no data better than bad data?" Inflated interest rates or inaccurate inflation rates are example of bad data that can lead to bad monetary policies. The biggest sources of inflation are food and energy. One-third percent of inflation is food (FAO, 2013).

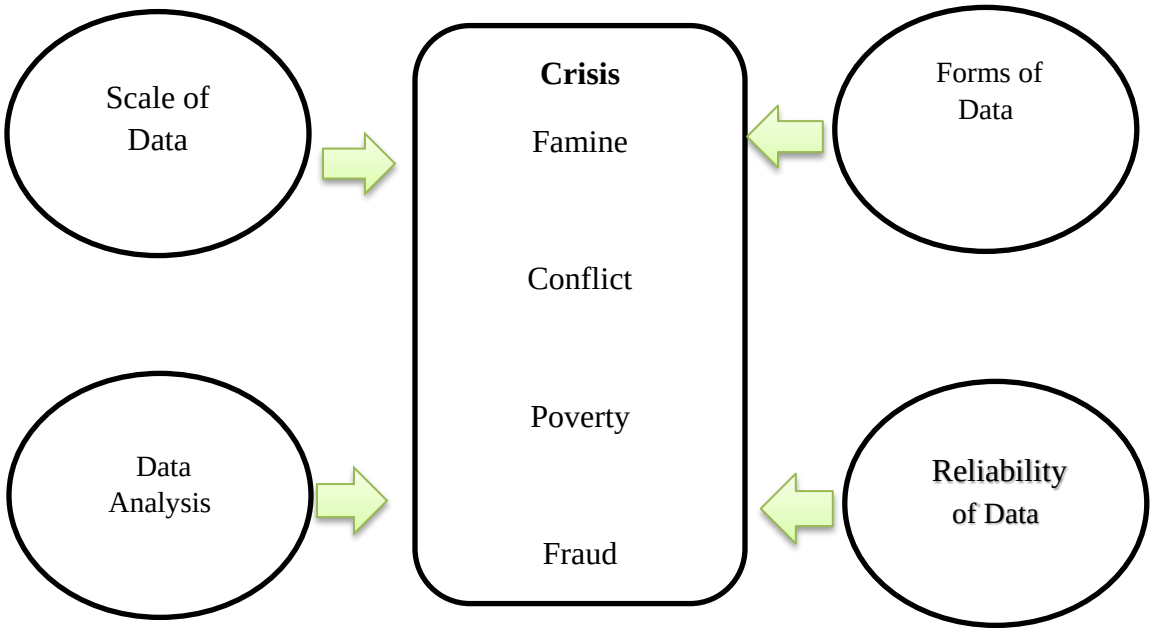
REAL SOLUTIONS FOR REAL WORLD PROBLEMS

Price inflation is a normal function of an economy over time. It is the phenomenon of rising prices in goods and services, to include food prices. To a degree, it is part of a normal

healthy economy. However, artificial or unusual supply side influences, such as, Warlord control of food supplies or unexpected famine, quickly produce unhealthy levels of inflation. Rampant or Hyper-Inflation results when prices increase at above-normal rates. For some of the reasons already stated, rampant inflation is not uncommon in developing countries. For example, the dramatic increase in food prices during 2007 and 2008 resulted in the eruption of civil unrest in some of the poorest countries across the world (Swan, Sierd, & Cichon, 2010).). Price increases for food staples such as corn, rice, soya, and wheat ranged from 31% to 130% (Onoja, 2010)... The countries that relied most heavily on food imports shouldered this burden the most. As has been shown, these "hungry" countries are at higher risk for becoming global security threats through artificial market forces. Therefore, national and international actors often intervene to help stabilize countries through both internal and external policies. One such international actor is the World Bank also known as the International Bank for Reconstruction and Development (IBRD). The World Bank is owned by 140 member countries; the voting power of each member country depends on its annual contributions which relate to the size of its economy. Its objectives are internal reconstruction and development and its purpose is to promote economic growth (Blay-Palmer, 2010). Another international player in the arena with the World Bank is the International Monetary Fund or IMF. The international resources and nature of these two entities makes them especially useful during crises that have potential global impacts. As their names indicate, the tools available to them are monetary in nature. The World Bank and IMF may promote their objectives is through fiscal policies in an attempt to help a troubled economy (Walsh & Jiangyan, 2012). One way to stabilize hyper-inflation is through price controls. Fiscal policy in developing countries often revolves around stabilization. However, the impact that monetary policy has on alleviating food crises varies (Cudjoe, Breisinger, & Diao, 2010). National governments and international bodies such as the World Bank intervene and attempt to mitigate the effects of rampantly increasing food prices. Some of these attempts help stabilize food prices while others seem only to help certain groups at the expense of the most vulnerable, e.g. children and pregnant women. Additionally, due to their artificial nature, these interventions may distort trade markets and produce more volatile food prices in the long run. Good data is vital for effective policy. This is especially true when dealing with monetary policy. Monetary policy is shaped by available data and depending on the quality of the data, the policies may help to stabilize food prices or increase the volatility of them.

Below is a theoretical framework for using Big Data based on the attributes of the data, data extraction, and data analysis. The amount of data available continues to grow exponentially. Data extraction techniques such as Webscraping increase the ease at which data can be collected for analysis. Advances in Big Data science have the potential to solve real-world problems and solve social ills.

Figure 1: A Framework for Using Big Data to Solve Real-world problems



TEAM NDIZI

The case of Team Ndizi demonstrates how the above framework works for solving a crisis. Team Ndizi was formed for a 2013 Data Drive Conference. The purpose of the data conference was for the volunteers to address two primary challenges (Das, 2013). First, find "new and innovative ways to measure poverty, (Das, 2013)". Secondly, volunteers were to sort "through World Bank procurement and program data to identify possible techniques to detect fraud and corruption, (Das, 2013)". Team Ndizi's project related to the 2009 Kenyan Hunger crisis. The 2009 Kenyan food crisis was the result of several factors to include drought, scandal and increases in food prices (FIAN, 2010). The team used web scraping techniques in order to pull food prices and consumption data from around the world. Web-scraping, also known as, web harvesting or web data extraction involves the automatic collection of information from websites.

Availability of data along with the extractions, reliability, form and analysis were all issues that Team Ndizi faced. The team had several important questions that they wanted answered. First, was there reliable data related to food pricing and consumption available? Also, could this data be extracted or "scraped" efficiently and organized so that banks and governments world-wide had the necessary information for managing poverty through monetary policy? One goal of this event was to "explore practical and tangible ways to demonstrate that open/big data can help improve poverty measurement," (Das, 2013).

The team speculated that the 2009 Kenyan reported inflation rate was inaccurate. They formed this hypothesis primarily because banks in Kenya were lending money under 20%. If the inflation rate were really 25%, banks would have gone bankrupt lending at rate of less than 20% (Das, 2013). In addition, they believed that finding more accurate pricing information would lead to more accurate reporting of inflation rates.

Team Ndizi recognized several challenges involved in finding more accurate data. First of all, they were not sure if quality data even existed. The next challenge was whether or not the data would be reliable for estimating consumption. They knew that they could track prices for food, housing, fuel, and energy. However, they needed more expertise in order to determine the impact that fluctuating prices have on consumption. Additionally, the team wanted to build a new data set rather than use an existing foundation. In order to create the new datasets they used multiple web scraping tools to collect data from several sites, allowing the team to capture many layers of food pricing (Das, 2013).

Numbeo charts (see below) of the CPI Index and real-time food prices of popular groceries, show it is possible to collect the information needed by the team. Numbeo uses crowd sourcing. Crowd sourcing is the process of obtaining data with the help of others primarily through the internet. Food prices and other relevant data are reported by users from around the world to create these Numbeo charts.

Table 1: Recommended Minimum Amount of Money for food (2400 calories, Western food types)

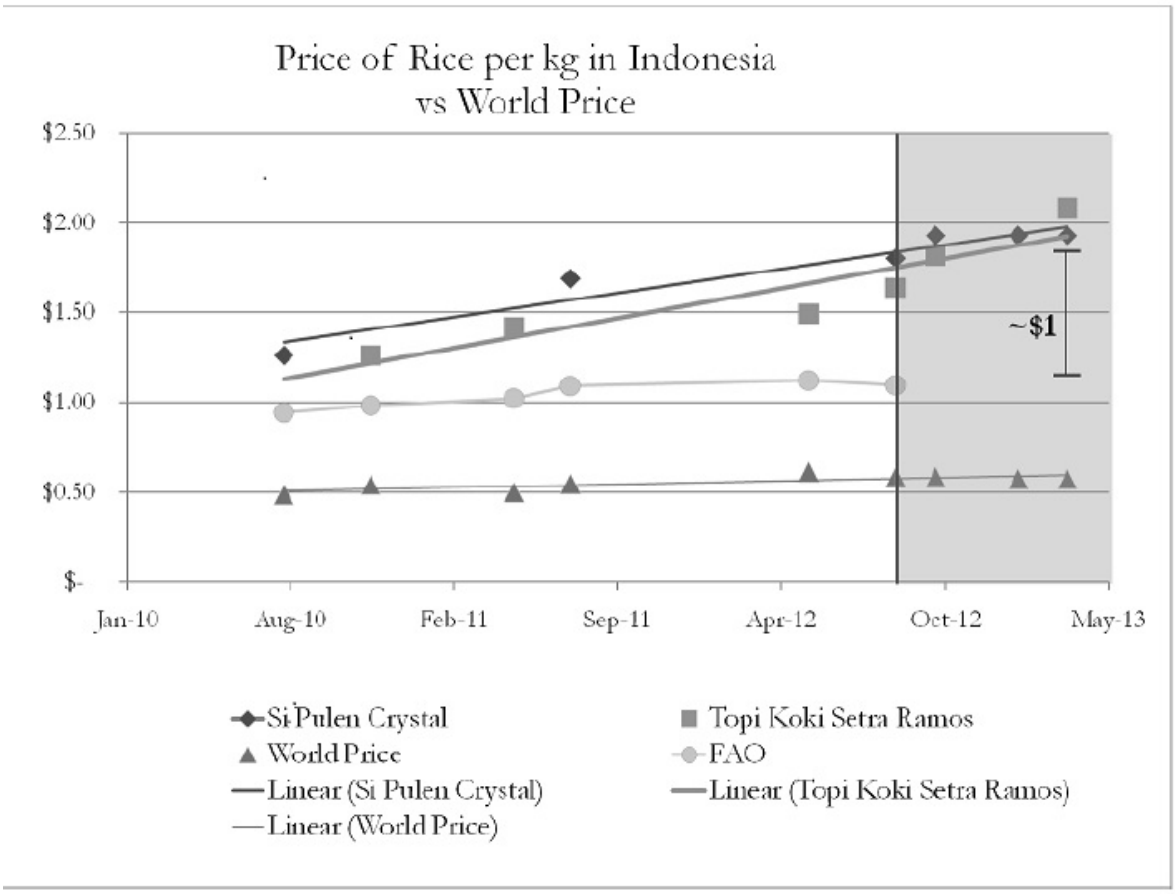
Milk (regular), (0.25 liter)	23.46 KSh
Loaf of Fresh White Bread (125.00 g)	13.32 KSh
Rice (white), (0.10 kg)	12.44 KSh
Eggs (2.40)	32.00 KSh
Local Cheese (0.10 kg)	63.71 KSh
Chicken Breasts (Boneless, Skinless), (0.15 kg)	84.84 KSh
Beef Round (0.15 kg) (or Equivalent Back Leg Red Meat)	65.62 KSh
Apples (0.30 kg)	87.99 KSh
Banana (0.25 kg)	36.23 KSh
Oranges (0.30 kg)	59.24 KSh
Tomato (0.20 kg)	19.97 KSh
Potato (0.20 kg)	19.10 KSh
Onion (0.10 kg)	9.76 KSh
Lettuce (0.20 head)	13.86 KSh
Daily recommended minimum amount of money for food per person	541.55 KSh
Monthly recommended minimum amount of money for food per person (assuming 31 days per month)	16,788.14 KSh

(Source: Numbeo http://www.numbeo.com/food-prices/country_result.jsp?country=Kenya)

Still, a question remained, “can big data crunching help feed the world?” (Rubens, 2014). For this reason, Team Ndizi also decided to look at the prices of rice. Soaring prices create questions of whether people will be able to afford just enough food to survive. Therefore, the team wanted to compare the Indonesian price of rice versus rice prices world-wide. Once again, Team Ndizi used web scraping techniques to focus on the price of Indonesian rice compared to World Prices. Through the use of web scraping techniques and software they created a

"definitive chart" of food prices that would plot the price of Indonesian rice from May of 2010 to May of 2013.

Figure 2: Indonesia Rice Prices v. World Price



(Source: https://github.com/mjrich/ndizi/commit/3d9b77cf905be7e8fe2137f0c5a65ae3b6c9b9cd)

The significance of Team Ndizi's comparison of actual prices for Indonesian rice versus the World Price was to illustrate that monetary policy must be more customized in order to have an impact on a local crisis. The monetary policy employed by the World Bank for the 2009 Kenyan Food Crisis was based on bad data. The result was an exacerbation of the problem.

SUMMARY

Data and knowledge have been growing exponentially over the past fifty years-especially over the Internet time frame (Larose, 2015). An example of this knowledge usage is Team Ndizi; attempting to harness some of this data and knowledge in order to solve real world problems such as hunger, famine and conflict. They used readily available data from public

websites; collected and analyzed using data searching and extraction techniques such as Webcrawler and Webscraping. This is just one example that demonstrates how organizations can leverage big data and data mining in decision making processes-whether for-profit or foe humanitarian usages.

Big Data's usages in famines, food security, inflation rates and poverty measurement are all in an effort to successfully reduce or alleviate these crises. As an example, food security is vital for stable governance and the protection of human rights. In 2008, the UN's Food and Agriculture Organization reported to the Committee on Food Security that there was a 25% increase in the number of countries in which food crises caused war and conflict since the 1980's (FAO, 2013). Sharp increases in food prices over this time period were a primary cause of these food crises. Common and well defined goals were: reduce poverty, inequality and unemployment; provide minimum levels of education, health, housing, and food to every citizen; broaden economic and social opportunities; form a cohesive national state. Big Data is a viable tool to help meet these goals.

LIST OF FURTHER READINGS

Allison, C., Easterwood, K. & Power, J. (2013). *Transforming business [electronic resource]: big data, mobility, and globalization.* Imprint: Indianapolis, IN: John Wiley and Sons.

Blay-Palmer, A. (2010). *Imagining sustainable food systems [electronic resource]: Theory and practice.* Burlington, VT: Ashgate.

Boyd, D. & Crawford, K. (2012) Critical questions for big data, *Information, Communication & Society*, *15*(5), 662-679, DOI: 10.1080/1369118X.2012.678878

Bryant, R. Katz, R. & Lazowska, E. (2008). Big-Data Computing: Creating revolutionary breakthroughs in commerce, science, and society. Paper presented at the Computing Community Consortium. Retrieved from http://cra.org/ccc/wp-content/uploads/sites/2/2015/05/Big_Data.pdf.

Cohen, M., & Garrett, J. (2010). The food price crisis and urban food (in)security. *Environment and Urbanization*, 22(2) 467-482.

Cribb, J. (2010). *The coming famine: The global food crisis and what we can do to avoid it.* Berkley, CA: University of California Press.

Crush, J., Hovorka, A., & Tevera, D. (2011). Food security in southern African cities. *Progress in Development Studies*, 11(4), 285-305. DOI: http://dx.doi.org/10.1177/146499341001100402.

Cudjoe, G., Breisinger, C., & Diao, X. (2010). Local impacts of a global crisis: Food price transmission, consumer welfare and poverty in Ghana. *Food Policy*, 35, 4, 294-302.

Davenport, T. (2014). Big Data at Work: Dispelling the Myths, Uncovering the Opportunities. Harvard Business Press.

Das, R. (2013) Scenes from a dive: What's big data got to do with fighting poverty and fraud? Data Blog. Retrieved from http://blogs.worldbank.org/opendata/scenes-from-a-dive-what-s-big-data-got-to-do-with-fighting-poverty-and-fraud

DataKind (2013). Scraping websites to collect consumption and food prices. Retrieved from http://www.datakind.org/projects/food-price-scraping/.

DataKind (n.d.) Our story. Retrieved from http://www.datakind.org/our-story.

FIAN International, (2010). Kenya's hunger crisis – The result of right to food violations. Report of a joint international mission by RAPDA and FIAN International. Retrieved from http://www.rtfn-watch.org/fileadmin/_migrated/content_uploads/Kenya_s_hunger_crisis_-_the_result_of_right_to_food_violations.pdf.

Food and Agriculture Organization of the United Nations (FAO), (2013). Monitoring African Food and Agricultural Policies project. Retrieved October 3, 2014 from http://www.fao.org/fileadmin/templates/mafap/documents/technical_notes/KENYA/KEN

Github.com (2013). Findings from the DataKind DataDive #TeamNdizi. Retrieved from https://github.com/mjrich/ndizi.

Gosling, S., Warren, R., Arnell, N., Good, P., Caesar, J., Bernie, D., Lowe, J., & Smith, S. (2011). A review of recent developments in climate change science. Part II: The global-scale impacts of climate change. *Progress in Physical Geography*, 35, 4, 443-464.

Kick, E., McKinney, L., & Thompson, G. (2011). Intensity of food deprivation: The integrative impacts of the world system, modernization, conflict, militarization and the environment. *International Journal of Comparative Sociology*, 52, 6, 478-502.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. Big Data & Society, 1 (1) 2053951714528481; DOI: 10.1177/205395171452848.

Larose, D. (2015). *Discovering knowledge in data: An introduction to data mining*. Hoboken, NJ: John Wiley & Sons.

Mellor, A. (2014). Little insights about big data. *Xconomy*. Retrieved from http://www.xconomy.com/wisconsin/2014/03/24/little-insights-about-big-data/).

Onoja, A. (2010). Contemporary conflicts in Africa. In *War and peace in Africa*. Falola, T. and Njoku, R. Editors. Durham, NC: Carolina Academic Press.

Rubens, P. (2014). Can big data crunching help feed the world? BBC News. Retrieved from http://www.bbc.com/news/business-26424338.

Surdak, C. & Agarwal, S. (2014). The Benevolent Side of Big Data, Finance & Development, December, 51(4).

Swan, S., Sierd, H., & Cichon, B. (2010). Crisis behind closed doors: Global food crisis and local hunger. *Journal of Agrarian Change*, 10(1), 107-118.

Walsh, J. and Jiangyan, Y. (2012). Inflation and income inequality [electronic resource]: is food inflation different? Washington, D.C.: International Monetary Fund.