Carlos Sathler | cssathler@gmail.com
INFO-I 535 – MGMT ACCESS USE BIG DATA (Fall 2016)

Big data in business

All answers are based on the following readings:

- *"A special report on managing information: Data, data everywhere," The Economist, February 25, 2010*
- *Data Scientist, The Sexiest Job of the 21st Century*, Thomas H. Davenport and D.J. Patil, Harvard Business Review, pp. 70-76, Oct 2012

1. Craig Mundie, past head of research and strategy at Microsoft made a powerful statement in the first article when he said, "What we are seeing is the ability to have economies form around the data – and that to me is the big change at a societal and even macroeconomic level." Economies forming around data. That implies data is becoming the new raw material of new businesses. Based on what you've read, **what are the new business efficiencies emerging from data**? A business efficiency is a change to a process brought about to reduce overall production costs.
   The articles mention a few examples directly: Google captures "data exhaust" as users search the web and improve search capabilities by narrowing down new searches to only show results that have been selected by a large number of prior users; doctors use data from medical records to select more efficient treatments, and to prevent diseases in patients; financial institutions use data to design better financial models that more efficiently identify financial risk and therefore prevent financial loss; companies use "abundant" data to explore opportunities on smaller markets (long tail marketing); statisticians perform market data analysis to "mine" for new business ideas; companies monitor real time data flow from mission critical machinery and identify trends that point to equipment failure ahead of time; information on premature babies is analyze to identify the risk of potential life threatening infections. In a general sense, data is being increasingly used to reveal insights about all business functions (e.g. marketing and operations). These insights inform actionable measures that have a very high probability of making the business more profitable through increased customer satisfaction and revenue, and/or cost reduction.

2. Identify and explain the multiple uses of data mining that business employees cited in the article, "A different game: Information is transforming traditional business."
   Some examples from the article: Wal-Mart was able to use data mining to predict and prepare for increased demand of pop-tarts when hurricanes approached; Cablecom was able to increase customer retention by offering special deals around the time customers were more likely to make the decision to discontinue their services; Li & Pung was able to predict an economic downturn on it's consumer markets by analyzing information on its retail partners; health data was used by a doctor of the University of Ontario and IBM to identify risk of fatal infections.
   All these examples speak of analysis done to data so as to identify causal relationships between different data elements, such that it can be determined with high confidence that in the presence of certain conditions it is expected that certain results will ensue. For example: "A hurricane is coming, yes? Well, rest assured people will come and look for pop-tarts. Let's stock up!"

3. "Data Scientist, The Sexiest Job of the 21st Century" makes instructive points about what makes the Data Scientist a new role in business. The data scientist can manipulate large amounts of unstructured data, can tell a story with the data, and can advise executives and product managers on the implications of the data for products, processes, and decisions. What training or set of classes do you see as essential background for this well-rounded person? (Hint: Look at the course list for IU Data Science program.)

A good knowledge of the big data technology toolset is very important, including cloud computing. I think a data scientist needs to know how to move and transform large volumes of data in the cloud, or at least know the basics of how that is done. Programming skills are also needed to be able to clean the data and perform exploratory data analysis. Descriptive statistics is also important for exploratory data analysis. Inferential statistics is important to draw inferences from the data and knowledge of applied machine learning is necessary in order to choose the right predictive-analytics solution for any given problem. Domain knowledge is also important as data sets tend to be complex and in the context of big data it is typically difficult to understand what each feature of the data really mean. Finally, it is important to understand ethical and legal issues relating to big data and analytics, for example, privacy and data ownership.

This is the list of classes I plan to take at IU:

ILS Z604	Topics in Library and Information Science: Informatics of Big Data
INFO I535	Topics in Informatics: Management, Access, and Use of Big and Complex Data
INFO I523	Big Data Applications and Analytics
ILS Z637	Information Visualization
INFO I524	Big Data Software and Projects
INFO I526	Applied Machine Learning
ILS Z639	Social Media Mining
INFO I520	Security for Networked Systems
INFO I590	Topics in Informatics: Data Science in Drug Discovery, Health, and Translational Medicine
INFO I590	Topics in Informatics: Real World Data Science

4. Referencing claims made in “The Open Society”, provide evidence for or against the statement: six years later (in 2016), public access to government data has released economic value and/or encouraged entrepreneurship.

I believe that since Feb 2010 when The Economist article was published, public access to government data has released economic value and encouraged entrepreneurship.

According to Wikipedia article on data.gov (https://en.wikipedia.org/wiki/Data.gov) the website started with 47 datasets in May 2009 and now it has 186,142 datasets (https://www.data.gov/). These datasets include data grouped in 14 topics, including manufacturing, business, finance and consumer, to name a few. Additionally, the site offers api’s for developers and links to documentation on the data.gov open source platform. While I don’t know the rate at which new datasets have been made available at the site over the years, it is apparent that over the past 6 years much more data has become available for the general public.. A quick look at the “impact” tab on the data.gov website (https://www.data.gov/impact/) will further show many examples of “citizens leveraging open data” in organizations in both the public and the private sector. Some of these organizations have hundreds of employees.

5. In “Clicking for Gold”, Google employed statistical inferencing to vastly improve language translation over the previous approaches that attempted to teach a computer the structural and grammatical rules of a language. How is statistical inferencing employed for this use?

If we look for an English book translated into Spanish for example, we might find 6 translations of the book, done by 5 different bilingual experts, manually. These translations are assumed to be correct because experts did them. If we use these books as input to a program capable of locating words in the text, and mapping them to their equivalent word in the translated book, we might see that the word “car” was translated 4 times into the Spanish word “coche” and only 1 time into the word “auto”.

Statistical inference will allow the program to determine "coche" is a better (more probably "correct") translation for the word "car". If we think of a massive number of books, it is possible to envision a well-written program will have a lot of good data to draw inferences from. For its translation solution, Google used a collection of EU documents translated into 20 different languages, and thousands of books, including their translations (by experts, manually) into different languages. With such vast volumes of good data, it is possible to envision good inferences for translation purposes. Notice also that Google Translate (https://translate.google.com/) invites feedback. As a bilingual person (I speak English and Portuguese) I'm in a good position to offer contributions for better translation between English and Portuguese.

6. The article "A different game: Information is transforming traditional business" identifies multiple uses of data mining. These you identified in Question 2 above. From what you read in public literature today, how has the list changed since 2010 when the article was written?
Since the article was published there was an explosion in the adoption of smart mobile devices, such as smart phones and tablets. The use of these devices created opportunities for a variety of applications that collect data easily mapped to individuals. Some examples: application to analyze customer movement at retail stores [1], talent acquisition analytics in the field of human resources [2], forensics [3].
Another important development since the article was published with a tremendous impact to big data and analytics is the improvement of sensor technology, which made it possible for companies to embed sensors in a variety of devices for data collection and ultimately analytics (IoT). The company libellium is a good example of a company (www.libelium.com/company) making sensors available for a variety of applications with potential to have a huge impact to business of many kinds. In the document "50 Sensor Applications for a Smarter World" [4] libellium lists devices that enable "Intelligent Shopping Application", "Smart Product Management" [p 21] which libellium supports with the use of RFID (Radio Frequency Identification) and NFC (Near Field Communication). Libellium has a number of sensors designed to support "smart" management (through data mining and analytics on the data collected by its sensors) in several areas such as energy (smart meters), urban development (smart cities), farming (smart animal farming) and agriculture (smart agriculture).
Other changes that are noticeable in everyday lives: recommender systems seem more common for example at Amazon, Spotify, Pandora, AirBnb, not just Netflix. Target advertisement is now pervasive, with adds popping up as soon as one opens their first web page based, on prior navigation. Applications that use pattern matching (and the cloud) are also becoming more common (e.g. PlantNet http://m.plantnet-project.org/, automatic tagging in Facebook). Finally, big data analytics "as a service" [5] is a recent development that has the potential to make the power of analytics more accessible in the marketplace.

[1] Your Smartphone As Big Data Analytics Tool – InformationWeek, Web Page, http://www.informationweek.com/big-data/big-data-analytics/your-smartphone-as-big-data-analytics-tool/d/d-id/1108040? (Accessed on 09/06/2016)

[2] Mobile data analytics: not just for consumer any more, Web page, https://www.pwc.com/gx/en/communications/publications/communications-review/assets/data-analytics.pdf (Accessed on 9/06/2016)

[3] Mobile Device Analytics, Smartphone Forensics Software | Wynyard Group, Web Page, https://www.wynyardgroup.com/us/solutions/mobile-device-analytics/

[4] 50 Sensor Applications for a Smarter World, Web Page, http://www.libelium.com/top_50_iot_sensor_applications_ranking/download, (Accessed on 09/06/2016)

[5] Big Data-As-A-Service Is Next Big Thing, Web Page, http://www.forbes.com/sites/bernardmarr/2015/04/27/big-data-as-a-service-is-next-big-thing/#22c2aff3f9a3 (Accessed on 09/06/2016)}