Introducing Big Data Analytics in High School and College

Raja Sooriamurthi
Information Systems Program
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
raja@cmu.edu

ABSTRACT

In this teaching tip and courseware note we describe a series of hands on activities and exercises that we've used to introduce the notion of big data analytics to a wide range of audience. These exercises range in complexity from a paper and pencil thought exercise, to using Google Trends for simple explorations, to using a spread sheet to simulate the iterative nature of Google's PageRank algorithm, to programming with a Python based map-reduce framework. These exercises have been used in courses to train high school teachers in data science, full semester university courses (undergraduate and graduate), and CS education outreach efforts. Feedback has been positive as to their efficacy.

CCS CONCEPTS

• **Social and professional topics → Computing education programs**;

KEYWORDS

Big data, PageRank, Google Trends, Map Reduce, Outreach

ACM Reference Format:
Raja Sooriamurthi. 2018. Introducing Big Data Analytics in High School and College. In *Proceedings of 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education (ITiCSE'18).*. ACM, New York, NY, USA, 2 pages. https://doi.org/https://doi.org/10.1145/3197091.3205834

1 DATA DRIVEN DECISION MAKING

From high school, to college, to continuing education, the awareness of the term *big data* is ubiquitous. But what exactly is it? How does it provide its analytic power? These are questions that arise in a range of contexts and equivalently can be answered to varying depths. In this teaching tip and courseware note we describe a series of hands on activities and exercises that we've used to introduce the notion of big data analytics to a wide range of audience.

Perhaps one of the most succinct ways of describing what data science is about, is, *science done with data*. Students are conversant with the scientific method and the various approaches to science as determined by its tools: empirical, theoretical, and computational. We lay the foundation for the data driven spirit of data science (and consequentially big data) with the following thought exercise.

ITiCSE'18, July 2–4, 2018, Larnaca, Cyprus

ACM ISBN 978-1-4503-5707-4/18/07.
https://doi.org/https://doi.org/10.1145/3197091.3205834

Exercise 1. Consider the problem of fatal traffic accidents. What data could be collected and analyzed so as to inform policy that could mitigate traffic accidents? We provide an example along the lines of table-1.

For several years we have done this exercise with students with a range of backgrounds (high school to grad school). Some of the domains students have explored using this framework are recycling, public transportation, disaster relief, traffic management, energy conservation, river pollution, drug overdose, cyber bullying etc. We wrap up our context setting discussion of data driven decision making based on real world case studies such as Harrah's work on customer segmentation and Progressive Insurance's work on identifying effective risk groups for motorcycle insurance [1].

This exercise has also formed the basis for a multi-week team project where students follow up the thought exercise with a real world exploration of the questions: What is the problem? Why is it interesting / important? What is the significance of its solution? What motivated you to pursue this? How do you plan to address this problem? What data do you need? What data do you have?

2 DATA DETECTIVE: HOW MUCH TO TIP?

Data science is an applied field with the following broad stages analyze a data set, extract meaningful observations, and communicate the implications of the story. The case study we've used to illustrate this process is based on a fascinating story that was first reported on Bloomberg and then elaborated in more detail in the blog I Quant NY[4].

Exercise 2. The motivating question: What is the most common tip amount for cabs? A simple statistical analysis of tip amounts leads to a histogram with various peaks. Peaks at 20%, 25%, and 30% are to be expected given that these are preset tipping rates on payment devices in cabs. An intriguing observation from the histogram is that one also sees peaks at 21% and 22%! Who tips these odd amounts? Or do they? What is the data telling us? When this anomalous pattern was investigated further it turned out that the two taxi companies in New York City were calculating tips in different ways resulting in the skewed distribution. From a pedagogical view point, a charming aspect is that this data set can be

Table 1: Thought exercise in data driven decision making.

Problem	Fatal Traffic Accidents
Data	Fatalities, location, time of day, weather, vehicle details, driver details (age, BAC), speed, seat-belts?
Analysis	Location? Speed?
Policy	Education, Penalty

analyzed with just a spreadsheet. It forms a neat hands-on exercise into investigative data journalism that is accessible by a wide audience.

3 HOW DOES "BIG" MAKE A DIFFERENCE?

While most everyone has heard the marketing term of *big data*, two intriguing questions are (i) how big is *big* and (ii) how does big make a difference? In the spirit of Carl Sagan's Cosmic Calendar, one can convey an intuition about the size of big data sets by comparing a single byte to a grain of sand and a peta byte to a mile long stretch of beach and a zeta byte to all the sand in the coastlines of the world (source EMC). Yet, how does size make a difference? How does more data make a difference? We discuss this with another hands-on exercise.

One of the earliest examples of data driven science is John Snow's work on visualizing the spread of cholera in 1854 London. Google's Flu Trends is an interesting example of history repeating itself 150 years later where the browsing patterns of people were used to detect outbreaks of influenza before officially being declared by the Center for Disease Control (CDC) in the US.

Additional Google tools such as Google Trends [3] and the Google ngram viewer illustrate how more data can amplify weak signal to noise ratios and help us develop insights that otherwise would have been impossible.

Exercise 3. One such accessible exercise is to use Google Trends to analyze if there are web search patterns that are predictive of a movie or actor winning an award in the Golden Globes or Academy Awards. An interesting observation was that before the announcement for Best Picture in the 2017 academy awards *La La Land* was trending in web searches much more than the ultimate winner *Moonlight*. Students can also search for and interpret trends about objects of personal interest (e.g, what are the seasonal and country wide trends for the mango fruit?).

4 NEED FOR MASSIVE PARALLELISM

We have developed a spreadsheet tool that allows one to explore the impact of data size on processing time. For example, even with 100 Gbs speed it would take more than 2 hours to just *read* 100 tera bytes of data with a single computing node; a peta byte would take more than 23 hours to read. But if one were to process 1 peta byte of data with 1000 compute nodes in parallel, the time dramatically falls to 90 seconds!

Exercise 4. Perform what-if analysis with various other configurations with this spread sheet and motivate the need for horizontal scaling, the map-reduce paradigm, and the Hadoop platform.

5 MAPREDUCE AND PAGERANK

To illustrate how horizontal parallelism can be tamed with the map-reduce paradigm and the Hadoop framework, we explore the PageRank algorithm in detail [2]. PageRank is a way of measuring the importance of a web page. Viewing the web as a directed graph, we get to a page in one of two ways (i) by randomly jumping to the page or (ii) by navigating to the page from the page we currently are on. The PageRank algorithm determines the probability of landing on a page given the structure of the inter-connections.

Exercise 5. The core equation for determining the PageRank of a web page, x, is given by:

$$PageRank(x) = \underbrace{\frac{(1-\beta)}{N}}_{\substack{\text{probability of jumping} \\ \text{to 1 out of N pages}}} + \underbrace{\beta * \sum_{y \to x} \frac{PR(y)}{out(y)}}_{\substack{\text{probability of getting to page } x \\ \text{by clicking on a link on page } y}}$$

where β is the probability of clicking on a link on the current page, estimated to be 0.85, and 1-β is the probability of jumping to a random page; y is the set of links referring to page x; $out(y)$ is the number of links on page y, and N is the total number of pages.

The essence of this equation is the iterative convergence on a set of values. This can be elegantly simulated on a spreadsheet, thereby making one of the most famous big data algorithms accessible to a wide audience.

Exercise 6. With this background, PageRank can be explored further with the Python based map-reduce framework MrJob.[1] We have a series of exercises that takes students through various tasks that can be expressed with the map-reduce framework and with MrJob in Python. One such task is using parallelism and MrJob to solve a Jumble puzzle where scrambled letters such as *jeyno*, *gaile* have to be unscrambled to form valid words such as enjoy, agile etc. The culminating exercise is expressing the PageRank algorithm (essentially the above equation) in about 20 lines of code with the map-reduce paradigm.

6 REFLECTION

All of the pedagogical material discussed in this note has been successfully used in a number of contexts from high school outreach to college (UG and G) over several years. Exercise-1 is the most foundational demonstrating the data driven approach to decision making. Whereas exercise-1 is paper based, the spirit of data driven journalism can be conveyed with exercise-2 which only requires comfort with using a spread-sheet. In high school outreach workshops on data science we have used these two exercises in tandem to motivate data science. Exploring the potential of big data with Google Trends (exercise-3) is a lot of fun given its open ended nature. Exercise 5 typically takes an entire 80 minute class period to demonstrate and have the students work through. The combination of exercises 5 and 6 form a nice sequence on the map-reduce paradigm. Student feedback has been very positive on the pedagogical value of these two exercise sets in highlighting the nuances of the PageRank algorithm and its implementation.

REFERENCES

[1] Thomas H. Davenport and Jeanne G. Harris. 2010. *Analytics at Work: Smarter Decisions, Better Results.* Harvard Business Review Press.
[2] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report 1999-66. Stanford InfoLab. http://ilpubs.stanford.edu:8090/422/
[3] Seth Stephens-Davidowitz. 2017. *Everybody Lies: Big Data, New Data, and What the Internet can tell us about who we really are.* William Morrow.
[4] Ben Willington. 2015. How Software in Half of NYC Cabs Generates $5.2 Million a Year in Extra Tips. (January 2015). Retrieved Jan 21, 2018 from http://iquantny.tumblr.com/post/107245431809/how-software-in-half-of-nyc-cabs-generates-52

[1] https://pythonhosted.org/mrjob