Issues in Information Systems
Volume 17, Issue III, pp. 209-217, 2016

CHALLENGES OF TEACHING DATA SCIENCE IN A BUSINESS SCHOOL

Jianfeng Wang, Indiana University of Pennsylvania, jwang@iup.edu
Linwu Gu, Indiana University of Pennsylvania, lgu@iup.edu

ABSTRACT

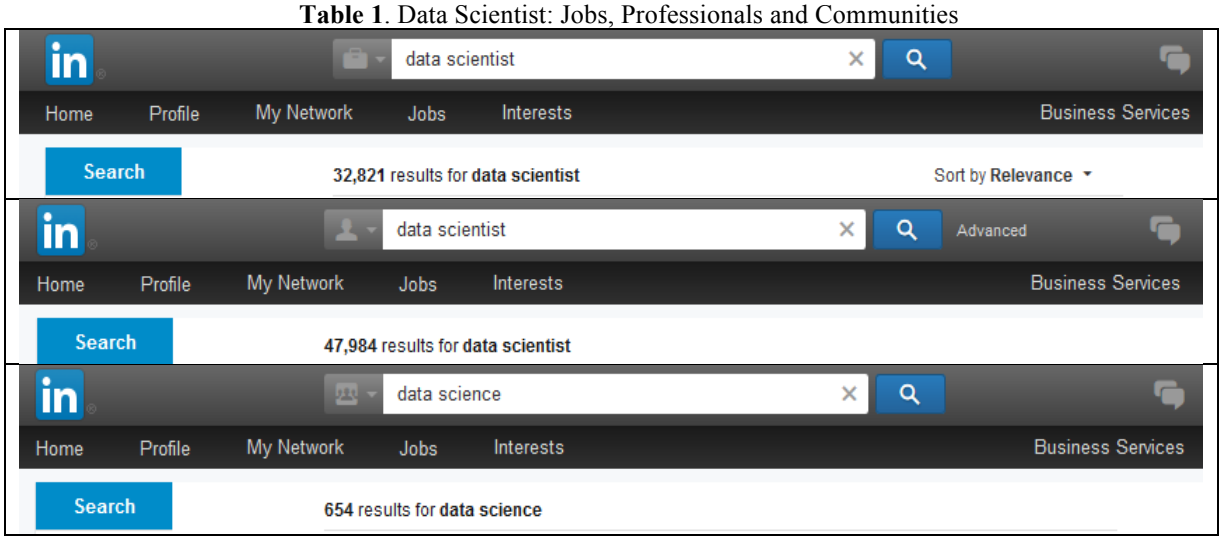
This paper discusses the challenges faced in teaching data science in a business school. Business school students with diverse academic backgrounds present unique challenges for faculty teaching data science as data science is such an emerging filed that it pre-requires students to have had some knowledge in the areas of computer science, mathematics, statistics, machine learning, and domain knowledge. Through our seminars of data science, we found many challenges and ways of dealing with such challenges.

Keywords: Data Science, R Programming, Python, Big Data

INTRODUCTION

Data science, an emerging field, is gaining momentum year by year. More jobs available require skills in data science. A search for job titled "data scientist" at www.linkedin.com will generate more than 30,000 returns of job posts. A similar search for people identified as "data scientist" at www.linkedin.com will display more than 49,000 data scientist professionals. Three or four years ago these numbers were close to zero. A search for data science groups at www.linkedin.com will return 654 group names relevant to data science. Data science communities are growing rapidly. The same is true of this emerging field itself, experiencing dynamic and fast change every year.

Table 1. Data Scientist: Jobs, Professionals and Communities



Retrieved from www.linkedin.com on July 9, 2016.

The basic ideas of data science as they are being discussed today are originated from traditional data warehouse, on top of which different decision support systems can be built. Such systems include OLAP, data mining, and model-driven, data-driven or hybrid decision support systems. Data-driven decision support systems have been with us for decades using both structured and unstructured data since both Wal-Mart and K-Mart built their decision support system in the 1980s.

A big driver of data science today is unprecedented amount of unstructured data available in various formats. That is what is so-called "volume, variety and velocity" of large data available. To analyze such huge amount of heterogeneous data sets, neither traditional statistics nor traditional machine learning nor the traditional database technology is sufficient to meet the challenges of analyzing big data, which is being generated in petabytes day by day. Second big driver there is the development of machine learning algorithm and that of deep learning in such a way that in some contexts machine learning may be able to extract intelligence from data to make a decision, better than a human being can do. Third driver is how cheaply huge amount of data can be stored, accessed and analyzed in a much faster way. Parallel computing based on hundreds of thousands of clusters using cheap commodity servers can be used to process huge amount of data in real quick way that it is reasonable to incorporate mining results from a data mining or machine learning process into decision making in a truly cost-effective way. Fourth driver is the open source development of core technologies and systems supporting the development. Open source efforts in the Hadoop ecosystem and other areas provide free software development and sharing supportive to the emergence and development of data science filed.

In the following section, we are going to discuss a few questions from our own experience teaching data science for three years:

1. What is data science?
2. What are challenges in teaching data science in a business school?
3. Which software to use in a data science class?
3. How to deal with big data in a data science class?
4. Conclusions

WHAT IS DATA SCIENCE?

We start with discussing this issue simply because this field is still at its very infant stage. Many things can be changed rapidly because of rapid changes in the supportive technologies and running environments. The definition of data science varies from author to author, book to book, and paper to paper. There is no agreed-on answer. It is important to have an appropriate understanding of what data science is as the answer will direct an instructor to arrange the contents for a data science class.

Dhar (2013) suggests that "data science is the study of the generalizable extraction of knowledge from data". A data scientist should able to integrate skill sets from "mathematics, machine learning, artificial intelligence, statistics, databases, and organization along with a deep understanding of the craft of problem formulation to engineer effective solutions". He also argues that big data promises automated actionable knowledge creation and predictive models for use by both humans and computers in settings where automated data-driven decision process can be better than human intuition and experience-based decision making.

Provost and Fawcett (2013) suggests that "Data science involves principles, processes, and techniques for understanding phenomena via the (automated) analysis of data… the ultimate goal of data science is improving decision making, as this generally is of direct interest to business." In their book, they "draw fundamental concepts of data science from many fields that study data analytics." According to their book website, as many as 133 universities have used their book in data science for business classes.

O'Neil and Schutt (2014) in their introductory book of data science, "Doing Data Science", makes a list of profile skills for a data scientist. These skills include knowledge from a few domains: computer science, math, statistics, machine learning, domain expertise, communication and presentation skills, and data visualization.

Boschetti (2015) thinks that data science core components include "linear algebra, statistical modelling, visualization, computational linguistics, graph analysis, machine learning, business intelligence, and data storage and retrieval", etc., which have been studied and researched for many year.

Baumer (2015) suggests that "Data science is an emerging interdisciplinary field that combines elements of mathematics, statistics, computer science, and knowledge in a particular application domain for the purpose of extracting meaningful information from the increasingly sophisticated array of data available in many settings. These data tend to be nontraditional, in the sense that they are often live, large, complex, and/or messy".

Common in these definitions of data science is that data science combines knowledge from computer science for data and processing programming, from linear algebra and statistics for correlation and causation detection and modelling, from machine learning or data mining for algorithms modifiable for big data analysis, from specific domains for domain-specific analysis, and from visualization and communication tools so that users can better understand the results. Data science is such an interdisciplinary area that at this stage it is hard to give a real unified and systematic definition. But basic components are data management and processing, data modelling and analysis, machine learning algorithms, visualization tools, and domain expertise.

Each of these components mentioned in the previous section can justify one or two semester discussions of core concepts and skill sets in the area. Data management and processing include munging and processing of relational database data, web-based structured and unstructured data, social media data, data from other sources, and technologies for integrating all these data together for analysis such as Hadoop systems, R, Python, Teradata, Microsoft Azure, Oracle, IBM Watson analytics, etc. Once data is ready for analysis, a supervised or unsupervised approach can be selected to analyze the data using algorithms from machine learning or data mining. The results can be visualized and plotted using tools available. To understand the algorithms used and to interpret the results from the analysis, basic knowledges from linear algebra and statistics are required. To understand the results from business perspectives, domain expertise is certainly a prerequisite.

WHAT ARE CHALLENGES IN TEACHING DATA SCIENCE IN A BUSINESS SCHOOL?

Students Insufficient Backgrounds

The challenges of teaching this new field can be seen from many aspects. A solid understanding of data science requires knowledge in linear algebra and statistics. Programming languages designed for statistical programming or data analysis such as R or Python can read and manipulate vectors or matrices or data frames easily (Wiley and Larry, 2015; Baumer, 2015). But many students who would like to take this class may lack the background in either linear algebra or programming languages or data management. Or it has been so long while ago that they barely remember anything about linear algebra or programming. Students in a business school usually will have more diverse backgrounds than those in an engineering program.

Traditionally business schools offer classes in business analytics using Excel. Excel with add-ins from Power BI and Solver can do lots of data analysis and some data mining. But the cost is not as cheap as to use R or Python while R is totally free to use. Excel is not so efficient in processing large size data as R.

There are some minor issues with using R. R is great and ranked No. 1 as statistical programming language. R is designed to handle data matrix and data frame in real efficient way and can handle very large size data with very complicated algorithms in machine learning or data mining. Usually one R function combine many steps of processing in just one function and one step. It is really neat that R can do such great and quick processing. R is a lot more wonderful than one could expect. On the other hand, students without programming background may just find it a little bit hard to understand and follow. While an R script processes data really quickly, a student may still try to figure out what an R function is doing at the step one.

The execution of data analysis has to be done at the computer using data mining or machine learning algorithm. Whether one uses SAS, SPSS, or open source R or Python, he needs special training in writing code. This is one of the most challenging part. In R there are thousands of R packages students can use for different purposes.

Though all the packages are always available for download and use, understand how to use appropriate functions from each package could certainly be a daunting task. The same is true of learning how to use SAS or SPSS for

serious data mining on large size data. But one can become quite good at using some of frequently used packages in R (Table 2).

Huge amount of data is available from very diverse sources such as those from www.kaggle.com, www.data.gov, Amazon.com/AWS and google cloud platform, social media, etc. Social media websites generate very large amount of data on a daily basis, ready for mining. Then how should we analyze such data? Why should we configure this or that model and try this or that algorithm? Model formulation is a process of applying knowledge from both statistics and domain area. Business analysis is such a broad area where any domain knowledge may be needed. Why we need such attributes in the analysis and collect such data? Why may there be any correlation or causality between the attributes? Students' lack of domain knowledge, statistics and linear algebra can make the discussion of modelling very boring and difficult.

Conditional probability and naive Bayes rule are very popular concepts and tools in machine learning. But many students have trouble understanding conditional probability, naive Bayes rule, and other statistical concepts, which makes it a bit hard when they need to choose a right algorithm to analyze data when data is available.

Even if they have clear grasp of linear algebra and statistical concepts, they may have trouble understanding the code to execute the data analysis in a computer system. Even if all these can be done, they may not have enough domain knowledge to interpret the results and communicate clearly to audience.

O'Neil and Schutt (2014) find that it may be pretty hard to find an employee who is good at the skills in all these areas required to do a good data science analysis. As a result a company often needs to have a team of working members with complementary skills in a big data analysis project. Classroom settings and requirements are much simplified. But the same situation is there. A student who can successfully take this class would have backgrounds in linear algebra, programming, database, statistics, and visualization. The contents for the class may focus on data mining algorithms, visualization, and business interpretation of the analytical results. A professors should be able to deliver such contents, meaning he should be at the master level of integrating these areas of skills and concepts. Our students are supposed to be able to integrate all the required skills in their assignments and classroom discussions. Otherwise they won't be able to follow the course. When a professor discuss a business case of data analysis, he should go from explaining why to collect such data for the business issue to algorithm selection or model formulation, to data munging, to coding, and to explaining the statistical results. Both teaching and learning can be pretty challenging.

Table 2. Example R Packages

K-means clustering	Package stats, with R base installation
Decision tree	Packages C50 and RWeka
Artificial neural network	Package Neuralnet
Logit analysis	Package Amelia, stats, Pscl
K nearest neighboring	Package class
Naïve Bayes	Package e1071
Support vector machine	Package kernlab or e1071 or KlaR
Market basket analysis	Package arules
Text mining	Package tm
Visualization	Package ggplot2, etc.

Textbook Selection

Provost and Fawcett (2013) is a popular book now but not enough. Their book has a nice introduction for statistics and their discussions of data mining concepts use business cases. But lots of such discussion is a bit of repetition of business analytics, data mining, or econometrics. The only difference may be in discussing social media data, on which econometrics and business analytics traditionally won't address. So if students already have backgrounds in business analytics or econometrics, not so much knowledge can be gained through such a class. Another problem with this book, is that they don't provide data sets or code script in whichever language.

O'Neil and Schutt (2014) is more technological and discuss more math formula. Readers certainly won't be able to read unless they have enough background in statistics, math and programming. The book has some code scripts in R readers can test.

There are probably more books available now with algorithms and data sets and scripts in R or Python. Such books all focus more or less on discussing the algorithms listed in table 1.

We think it is better to have some introductory materials as available in Provost and Fawcett (2013) to fit the backgrounds of the majority of the students in a business school class. If some students have true strong backgrounds in statistics and programming, then they can further read a machine learning or data mining book such as Lantz (2013). If all students in a class have very good backgrounds for the class, a book in machine learning or data mining probably will be better. This may more likely be a case in a data science program hosted in a computer science or applied statistics department.

Somehow to teach data science in a business school as suggested by Provost and Fawcett (2013) is much more of a compromise between students' insufficient background and diverse prerequisites for such an interdisciplinary field. Indeed business school students should be exposed to the ideas in advances in data science area as the drivers of data science are accelerating in the foreseeable future while more and more companies collect more data and become more data-driven in their decision making.

WHICH SOFTWARE TO USE FOR DATA SCIENCE, R OR PYTHON?

Both R and Python are very good for data analysis. We discuss which one may be better for a classroom setting. GitHub user Szilard edited a table about the machine learning scripts in www.kaggle.com. Below is the table (Table 3) he post at www.github.com.

We suggested our students to use R for their project. As we have tested, R processes data much faster than Excel when data files are bigger than 1 MB. R can process both regular-sized data file and very large data files in an efficient way. R, created for statistical processing, is very fit for vector, matrix and data frame. Rich functions in R packages are available to handle missing values, reformat data, transform from categorical features to dummies, and create subsets of data as analysis requires. For beginners R packages function more like a black box in processing data as long as students can follow to use appropriate functions from suggested packages. But at the end, students should be able to interpret the results themselves.

From tables 3 and 4, it is easy to see that R is currently the most popular language for data analytics. R is open source software, free downloadable, very easy to use and install. R base installation available from https://cran.r-project.org/. If additional packages needed to analyze some data sets, users can download immediately through R-project mirrors. R studio from https://www.rstudio.com/ is also free to use with easy features to edit and manage path, files and packages in a local computers. Before one can use R studio, he must have R base installation in his local computer first. R can analyze both structured and unstructured data and streaming data.

According to https://cran.r-project.org/web/packages/, the CRAN package repository currently features 8393 R packages. Users can download any packages they need to use when they need from a CRAN mirror. R is designed for statistical data processing with statisticians request in mind. So R is very good for machine learning and graphic use. Package "ggplots" provides very rich sets of graphical functions.

There are many blog and online discussions by R community. When students have questions, they can google for solutions possibly available anywhere online.

Overall both R and Python are pretty good for data analysis. Once students are familiar with R, it is pretty easy for them to learn and program with Python.

It is certainly fine for students to use Python for their projects. There are free python editors and IDEs that they can use. According to https://pypi.python.org/pypi, currently there are 80525 of Python packages, most of which are free to use. For beginners, free editors or IDEs such as enough canopy, anaconda, etc., are good and powerful enough to start with.

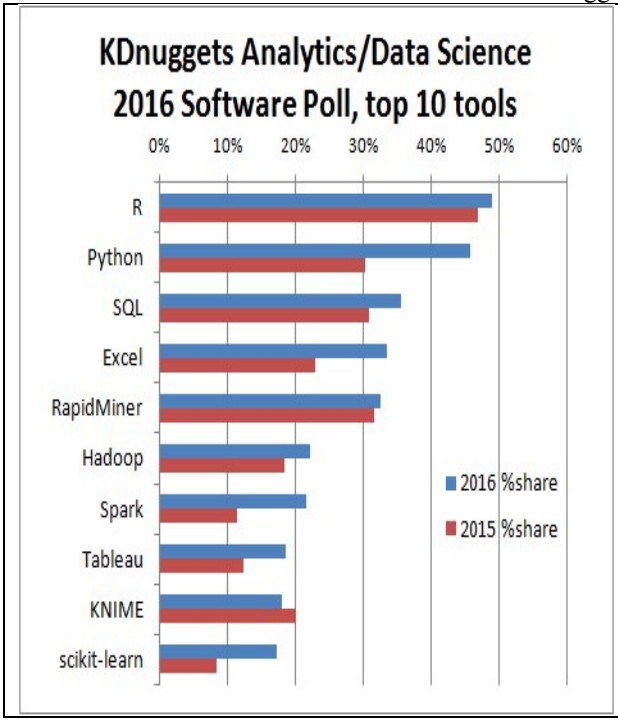
Both R and Python can handle unstructured data such as text messages from social media. There are a few packages available in R to deal with text mining and analysis. There are rich sets of text mining packages in Python. Beginners can use R packages as black boxes to analyze data sets without understanding the coding details of a package. Beginners need to start with Python basics before they can truly understand Python for analytics. We recommend R for beginners to try and learn in a classroom setting of data science class.

Table 3. Kaggle.com Script Counts

	scripts	votes	views
Julia	30	50	11059
Python	945	4558	1095967
R+Rmd	975	5212	1223700

Retrieved at www.github.com on May 10, 2016.

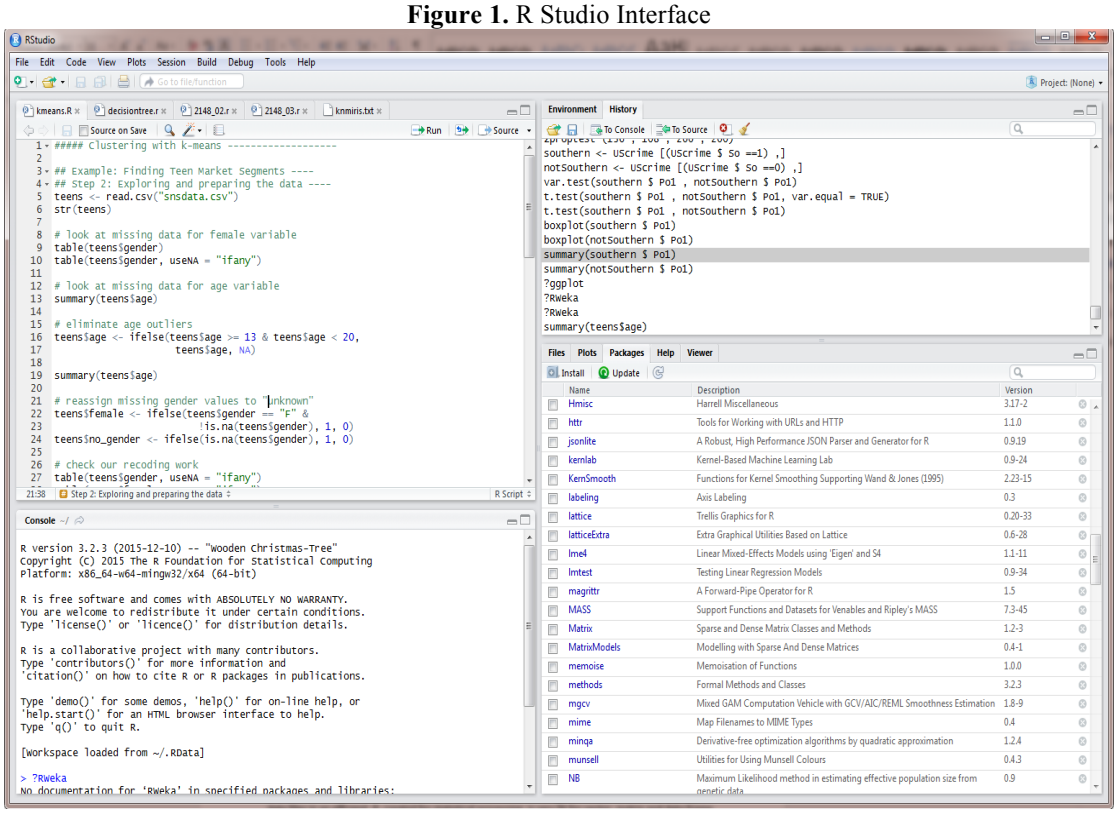
Table 4. KDnuggets.com Survey 2016



Tool	2016 % share	% change	% alone
R	49%	+4.5%	1.4%
Python	45.8%	+51%	0.1%
SQL	35.5%	+15%	0%
Excel	33.6%	+47%	0.2%
RapidMiner	32.6%	+3.5%	11.7%
Hadoop	22.1%	+20%	0%
Spark	21.6%	+91%	0.2%
Tableau	18.5%	+49%	0.2%
KNIME	18.0%	-10%	4.4%
scikit-learn	17.2%	+107%	0%

Retrieved from http://www.kdnuggets.com on July 10, 2016.

Figure 1. R Studio Interface



How to Use Excel in the Class

Most universities have group licenses from Microsoft and install Microsoft Office in their computer labs. It is free for students to get a copy when they are enrolled.

In Year 2014 and 2015, we offered data analysis classes at both undergraduate and graduate levels using Excel, PowerBI and Microsoft SQL Server. The classes were very welcomed by our students. Students demonstrated huge interests and were very excited to know that Excel could be used to do so many advanced data analysis jobs. There are a large number of statistical functions and chart functionalities available in Excel. PivotTable provides great tools for summarizing data in multi-dimensions in an easy and quick way. PowerMap is pretty good to display summary data in a map, based on www.bing.com map. PowerQuery can be used to retrieve data from relational databases, Internet, and Microsoft Azure. Users can edit data in PowerQuery and then load the data set to an Excel spreadsheet or PowerPivot. All the data sets in an Excel workbook can be integrated and analyzed together. SQL Server Data tools provide wizards to create data cubes using SQL Server analysis service from a data warehouse built within the SQL Server. And Data cubes can be analyzed using Excel data mining add-in. Though one Excel spreadsheet can only handle about 1,048,576 rows by 16,384 columns, PowerPivot and SQL server can handle fair large data sets. The problem is that a lab computer will slow down dramatically when a user try to process 10 MB of data. For a lab computer with 12 GB of prime memory, it is a big job for the computer to process 10 MB of data. If there is only 6 or 8 GB DRAM in the computer, 4 MB of data can cause trouble to the computer.

Another technical issue lies in the lab computer and network configuration. In our university students are allowed to read and write data only through network storages. We store data sets in the designated network storage. During a class time there can be jam issues with the lab network when all the students try to read the same data at the same time from the same server of the network storage. This is especially a problem when you have a big class with 25-40 students.

Excel can be very helpful in preparing and training data for beginners. Data saved in Excel can be read to R session through some packages and can then be analyzed. Both numeric and textual data can be organized using Excel and fed to R session using XLSX or XLSconnect packages. There are functions in Excel that can be used to reformat data or deal with missing values. Microsoft SQL server data tools carry some mining algorithms for classification, regression, segmentation, association, and sequence analysis. Excel together with Solver add-ins and SQL server data tools can provide meaningful learning experiences in data analysis on structured data. Though they are not good tools to analyzing unstructured data, it is a good approach to introduce data analysis first using Excel and other easy-to-control tools in a data science class. It is easier for beginners to understand because most steps we do in data analysis using Excel are visible in Excel. Many basic concepts in data science can be demonstrated using small data sets in Excel. So Excel can be used to help students develop a solid understanding about basic concepts.

In spring 2016 when we officially offered the data science class, we started with introducing basic R syntax, keywords and functions, and some Excel features. When students had trouble understanding in R, we used Excel to tell the same story and then went back to R. This proved to be an effective way of helping students who had a little bit trouble to catch up.

HOW TO DEAL WITH BIG DATA?

When we talk about big data, we refer to a data set bigger than one terabyte. Huge amount of unstructured data from social media makes it a daunting task to analyze using traditional approaches in data analytics. To manage and analyze this kind of data, Hadoop systems have been available since 2009 as open source software (www.apache.org). Leading vendors such as Cloudera, Hortonworks, MapR, Facebook.com, Linkedin.com, Twitter.com, Amazon.com, IBM, and a few others created their own version of Hadoop systems to manage and process the big data generated from their services. Hadoop was designed to be run on clusters of cheap servers. Such cheap servers often are configured with memory ranged from 16 GB to 96 GB with 1TB hard disk storage. Since they were available in 2009, Hadoop systems have been further developed and enhanced year by year by the Hadoop communities worldwide. Today there are more options to run Hadoop systems for cluster-based distributed data processing than year 2009. For students to practice Hadoop, they can download sandboxes of Hadoop systems from Hortonworks, Cloudera, or MapR to run in virtualbox or VMWare and practice through local computers. Google cloud platform provides two-month free-trial of BigTable. Google BigQuery uses Colossus, a proprietary Google technology, and provide free data processing service under one terabyte. There may be some other free service available for students to try.

Apache Spark is a large-scale data processing system, free downloadable to your clusters. Spark can process data from HDFS, Cassandra, HBase, and Amazon S3, etc. Users can run R, Python, Java or Scala on top of Spark systems. As part of Hadoop ecosystem, Spark is gaining momentum with time. Spark has proved to be more efficient than MapReduce. Spark is built on top of YARN, a more general resource manager and task coordinator than MapReduce. Users can download Spark to their single computers to test some features of Spark. But just like Hadoop, Spark is designed to run on clusters. Spark is becoming a data analytics operating system on which big data from different sources, whether structured or unstructured, stored or streamed, can be pulled together to be analyzed and complement each other.

Revolution Analytics Inc., now a Microsoft company, provides some R packages to run on Hadoop MapReduce framework. Revolution Analytics provides RHadoop with three basic packages: RHDFS, RHBASE, and rmr.

CONCLUSIONS

In this paper we identified a few challenges we experienced in teaching data science in our business school. The classes have been very popular among our students. The skill sets and concepts learned from our classes indeed help many students get job offers as data analysts or data scientists. The major challenge of teaching data science comes from the fact that data science is such an applied field requiring comprehensive knowledges from many different

areas. But business school students often lack the sufficient backgrounds in math or programming. Sometimes it is even hard to decide how to sequentially introduce some concepts or packages such that most students in the classes would not be confused.

We summarize in this paper our considerations and suggestions of which textbook to use and which software to use and how to deal with big data. Early this year we noticed many books about big data analysis and those about machine learning in R or Python are coming and will be available later. Hopefully there will be more books faculty can choose when they teach.

Both R and Python are being improved and enhanced with contributions from researchers and programmers all over the world. It can be good and bad. It is good because R may become more powerful and easier to program with. It may be bad for teaching because lots of scripts may have to be rewritten or modified so that they can be interpreted with newly improved packages. Hadoop systems are progressing every year. It is truly necessary we share our experiences in teaching such pioneering and dynamic topics when the industries move forward to adopting many new technologies.

REFERENCES

Boschetti, A. (2015). Python Data Science Essential, Packt Publishing, Birmingham, UK.

Baumer, B, (2015) A Data Science Course for Undergraduates: Thinking with Data, *American Statistician, 69*(4), pp. 334-342.

DHAR, V. (2013). Data Science and Prediction, Communications of the *ACM. 56*(12), pp. 64-73.

O'Neil, C. & Schutt, R. (2014). Doing Data Science, O'Reilly Media, Sebastopol, CA.

Provost, F. & Fawcett, T. (2013). Data Science for Business, O'Reilly Media, Sebastopol, CA.