Classroom Demonstrations of Big Data

Eric A. Suess[*]

Abstract

We present examples of accessing and analyzing large data sets for use in a classroom at the first year graduate level or senior undergraduate level. Larger and larger data sets are used. The beginning examples focus on in memory data sets and the later examples extend beyond the physical memory available on the computer. Simulated data sets are suggested and sources for real world data are given. The suggested examples are foundational for working with *Big Data*.

Key Words: Big Data, visualization, classification, prediction, R, Revolution Analytics

1. Introduction

We became interested in *Big Data* and the use of parallel computation (distributed computing) and to a lesser extent in the use of parallel data storage (distributed data storage) during the excitement of the NetFlix Prize (2009). Learning that the winners of the competition used the amazon EC2 cluster and the amazon S3 storage inspired us to investigate and learn about these new computing environments, now referred to as “cloud computing”.

The Heritage Provider Network Health Prize (2012) was also very exciting. And now with kaggle.com being a location on the Internet for these types of competitions, there are current open data analysis competitions posted regularly. Many of the kaggle completions focus on *Large Data* problems which benefit from the use of parallel computation and can be used as a stepping stone to preparing to work with *Big Data, Analytics, and Data Science*.

The computing techniques and hardware needed to work with *Big Data* currently seem far from the introductory course work in Statistics at all levels, such as, lower division and upper division Statistics courses, and first year graduate courses in Statistics. However, with some effort to motivate the ideas of *Big Data*, parallel computing, and distributed storage at earlier stages in Statistics Education, it would then be possible to

1. cover some practical applications of *Large Data*,
2. discuss the next steps toward working with *Big Data*,
3. show parallel computation in action,
4. and to introduce distributed storage much earlier in the Master’s level curriculum and undergraduate curriculum.

These additional topics would be very exciting to students of Statistics and they would connect their studies with the discussions that are going on in the media and beyond.

In introductory Statistics classes the data sets are usually in the hundreds of observations. A big data set in these classes might be in the tens of thousands of observations. Much of the examples are health related and are rarely in the millions of observations.

[*] Department of Statistics and Biostatistics, California State University, East Bay, 25800 Carlos Bee Blvd., Hayward, CA 94542

In the current business settings where the term *Big Data* is used there is no clear definition. It is common to hear about *Big Data* in terms of very large databases. The number of observations is replaced by the number of gigabytes or terabytes or petabytes or beyond. In this setting how the data is accessed is an important consideration. Accessing data from multiple computers concurrently or accessing datasets that are beyond the memory of the computer are the next steps for a student of Statistics. To start we suggest students need to deal with *Large Data*. These are datasets that can fit into the memory of a student's computer. While the word *large* may imply something that is bigger than something that is big, the term large seems appropriate. The term *big* as it is used in the term *Big Data* implies very very large datasets. These datasets might be consider *humungous*.

2. Why is Big Data not being discussed earlier in Statistics?

This is an important question! All of the core curriculum usually focuses on traditional methods, such as t-tests, ANOVA, Linear Regression, etc., and uses Confidence Intervals, Hypothesis Testing, and p-values. These are core topics that will remain the core curriculum of Statistics.

With the advances in computer technology, being able to access and analyze large datasets is now possible on PCs and laptops. However, opportunities to interact with such data is not currently a core part of the common Statistics curriculum.

Big Data is being discussed by students and is constantly a focus of questions (asked directly or indirectly) asked of faculty these days. The discussion of *Big Data* seems to be disconnect from the curriculum.

In the end, students of Statistics need to become much more capable with and more knowledgeable users of their own computers, which are now all inherently capable of storing large datasets for analysis because of larger amounts of RAM being installed (8, 16, 32 gigabytes), having very large harddrives (500 megabytes, 1 terabytes, beyond), and for parallel computation (Core2Duo, iCore3, iCore5, iCore7, 2, 4, 8 cores respectively).

Statistics educators should try to incorporate more use of *Large Data* sets and create exercises that include more data manipulation and experiences accessing data (i.e. data munging or data wrangling). The *size of the data* should become a topic of discussion when presenting standard statistical techniques, such as linear and logistic regression. This discussion should be expanded beyond the sample size n to include the size of the data stored on a harddrive and to the consideration of data stored in different locations and in difference formats, Cleveland (2001), Nolan (2010), and Hey (2009).

3. What exactly is Big Data?

The term *Big Data* has meaning in business settings and in the media, but it is not clearly defined among Statisticians. The idea of *Big Data* seems to focus on very large databases containing datasets that may include many data tables that include *very large numbers* of variables and *very very large* numbers of observations. Or contain unstructured data that does not fall into a nice format (natural language, images, for example), this is a next step in the *Big Data* discussion.

These types of datasets may be stored in .csv files and/or databases (other formats are also used). They may be stored on multiple data servers. The size of the datasets *far far exceeds* the RAM in a usual PC or laptop and often *far exceeds* the space on a usual harddrive.

The *humungous* size of *Big Data* makes giving student access to such data very difficult or not possible when the *students are providing their own computer hardware*.

4. What can be taught in the classroom?

While hands on experience with *Big Data* is not yet easily accessible for most students at the introductory undergraduate level and difficult at the MS level, there are many foundational computing experiences that could be included into the current curriculum that would be very valuable to students to build their experiences with *Large Data* to eventually be prepared to work with *Big Data*.

The idea of *Large Data* is for students to work with data that can be accessed in memory on their single computer and possibly in parallel on two computers.

5. Computing platform(s) and software.

In the field of Statistics there are two primary software packages for accessing *Large Data*, these programs are `R` and SAS. Both software packages are fully capable of accessing and analyzing *Large Data*. It is debatable whether one or the other is more commonly used in business settings. We believe it is very common to see `R` being used in businesses that are developing *Big Data* solutions to problems. We focus on the use of `R` in this paper.

`R` is an **open source** software package. Using it on an open source operating system, such as gnu/linux, is the natural next step. Another topic, related to *Big Data* is the use of unix operating systems for performing this kind of work.

Students should be encouraged to learn `mysql`, `mariabd` or `sqlite`. Student need to become familiar with the SQL commands such as *SELECT, FROM, WHERE*. SQL should become common knowledge among Statistics senior undergraduate and MS level students.

Finally, beyond `R` students should be introduced to the basics of `perl` and `python`. Both of these languages are commonly used for working with data.

Next we present examples that can be used by students to develop experience with *Large Data* and databases as a foundational experience for eventually working with *Big Data*.

6. Examples

6.1 Simulated data for cluster analysis.

As a starting point for working with *Large Data*, simulating many datasets that have millions of observations with different clusters, gives an excellent point to start working with such data. Ideally the datasets can be loaded into memory, within `R`, so a final analysis can be performed.

Here are a collection of ideas that students can investigate with simulated data.

1. A first question for students to investigate is how large a dataset can be simulated and stored on their computer harddrive.

2. After simulating a dataset, split the file so the parts can be loaded quickly into memory, within `R`.

3. Figure out how to randomly split the data into separate data files.

4. Figure out how to important a simulated dataset into a database and then how to access the same data in a database, from within `R`.

Here are a collection of ideas that students can investigate with analyzing simulated data.

1. Propose various forms of analysis and perform them on a sample.
2. Propose a overall analysis and perform it on the entire dataset.
3. Evaluate the analysis.
4. Communicate the results clearly.

For this example, the analysis could be as follows: Consider simulating data for a cluster analysis. See the the R library MixSim, Melnykov (2012). Use the first 1000 observations to develop the R code. Export the simulated data set out of R and import it to a mysql database. Start with the first 1000 observations. Use the

```
write.csv()
```

function in R. Use the R library RODBC to connect to the mysql database.

Propose a sampling procedure, use the

```
sample()
```

function. For example,

```
A$X[sample(nrow(A$X), 3),]
```

Since we are considering cluster analysis, try the

```
kmeans()
```

function with different numbers of clusters. Try to find the best number of clusters with the entire data set. Compare with the simulated values for the groups. Compute the errors in classification. Examine plots. Consider exporting the data to a .csv file and loading it into rattle (2014) to use the *Partition* and the *Evaluate* tabs.

Communicate the results. Try posting the code on GitHub and writing a blog post about the data and analysis. Try to make the final plot using the cloud based software tableau public (2014).

The code giving in Melnykov (2012) is an excellent place to have students start simulating data with clusters. Here is a nice example from the paper.

```
library(MixSim)
n.size = 500000000
Q <- MixSim(MaxOmega = 0.20, BarOmega = 0.05, K = 5, p = 2)
A <- simdataset(n = n.size, Pi = Q$Pi, Mu = Q$Mu, S = Q$S)
X11()
colors <- c("red", "green", "blue", "brown", "magenta")
par(mar = c(0.1, 0.1, 0.1, 0.1))
plot(A$X, col = colors[A$id], pch = 19, cex = 0.8)
B <- kmeans(A$X,5)
X11()
plot(A$X, col = colors[B$cluster], pch = 19, cex = 0.8)
```

In summary, simulation can get quite large and still be within the RAM memory on a computer. Writing to a .csv file gives a clear view of how large the data file is in KB or MB.

The plot() function is very slow. There is a clear need for an alternative method of visualizing the data. Maybe hexbin() or smoothScatter() could be introduced.

6.2 NYC taxi data.

Read online about the FOIL request that Chris Whong made and the efforts that followed to decode the data, Whong (2014). Two *Large Data* sets, or maybe *Big Data* sets, Trip Data (11.0 GB) and Fare Data (7.7GB) were made available as a result of his FOIL request.

Here are some ideas for students to work with these data.

1. Split the files so the parts can be loaded quickly into memory, within `R`.
2. Access the same data in a database, from within `R`.
3. Propose a sampling procedure of the *Large Data* set to produce an appropriate random sample from the overall *Large Data* set.

Here are some ideas for students to analyze these data.

1. Propose various forms of analysis and perform them on the sample.
2. Propose a overall analysis and perform it.
3. Evaluate the analysis.
4. Communicate the results clearly.

These datasets give an opportunity for students to *merge* data. Trying to do the merge using SQL in a `mysql` database my be challenging. Sampling when the full datafile cannot be read into `R`. How to proceed?

Some further questions for students to consider.

1. These *Large Data* set includes GPS data for pick up and drop off. How can the GPS information be used.
2. This *Large Data* set include a variable that has time and date same field. How to split this field?
3. Consider *Linear Regression* and *Logistic Regression*.
4. Read of the other blog posts that followed.

This was a very nice *Big Data* example that can be used with students directly and it includes many other issues related to the topic of *Big Data*, such as data security.

The next three examples give only general suggestions about sources of data that can be used with students to develop skills for *data science* and working with *Big Data*.

6.3 Airline on-time performance - Data expo 2009.

Read online about the data and the original student data competition sponsored by the ASA Section on Computing and Statistical Graphics, Data expo, Airline on-time performance, (2009). The website gives very good guidance about how to start to work with these data and some of the questions that can be answered.

Here are some ideas for students to work with these data.

1. Split each file so the parts can be loaded quickly into memory, within `R`.
2. Access the same data in a database, from within `R`.

3. Propose a sampling procedure of the *Large Data* set to produce an appropriate random sample from the overall *Large Data* set.

Here are some ideas for students to analyze these data.

1. Propose various forms of analysis and perform them on the sample.
2. Propose a overall analysis and perform it.
3. Evaluate the analysis.
4. Communicate the results clearly.

6.4 kaggle Titanic.

There is an excellent dataset provided on the kaggle.com (2014) related to data from the Titanic. The example focuses on building a model to predict survival on the Titanic.

Beyond this example, students should be encouraged to read the website to learn about the current and past competitions.

This is an excellent introduction to how these competition websites works. It should also be noted that kaggle now refers to itself as the home of Data Science.

An excellent alternative website with similar data examples is leada (2014).

6.5 Airline data again.

Another way to work with the airline data, ASA Section on Computing and Statistical Graphics, Data expo, Airline on-time performance, (2009), could be with Revolution Analytics `R` (2014) using the rxLogit function provided in their version of `R`. See the YouTube video (https://www.youtube.com/watch?v=KZHioV-DOD8). It shows a *Large Data* analysis.

It should be possible for students to replicate the data analysis if they have a powerful enough computer or they can give it a try for free on amazon EC2.

6.6 For introductory Statistics at the Freshman level.

There is very little experience students of Statistics receive related to databases. One additional topic that would be very easy to add into basic Statistics classes, while discussing introductory probability, is password security and password management software.

A discussion of these topics could start with the following question. What do your passwords look like? Do they look like this?

```
5′+M{bh6U7VGsT!2T&Zr}zv&HDSvi2M
```

The usual answer is some short collection of words and symbols. Password security is one modern reason for studying probability and understanding equally likely outcomes these days. Most software that generates passwords uses some form of equally likely probabilities on the list of keyboard symbols that can be included in the password.

Here are some ideas for students to work with passwords.

1. How to generate random passwords?
2. How to store passwords in a database?

One open source software package for password management is KeePass (2014). Students could be encouraged to try the software and to see how randomly generated long passwords can be used and stored in the software database. This is an excellent opportunity to learn about data security using more secure passwords and to learn about how an encrypted database can be used.

7. Next Steps.

Further efforts need to be made to develop Statistics faculty experience with MapReduce, Hadoop, Hive, Pig, etc.

Other software beyond `R` and `SAS` needs to become part of the usual Statistics curriculum.

Some suggestions:

1. Hortonworks (2014)
2. Oracle `R` Distribution (2014)

Also, further efforts need to be made to develop the awareness by Statistics faculty of the availability of data on the internet that can be used in classes. One such website to find example data.

1. NYC Open Data (2014)

Finally, we need to determine connections to courses offered in Business and in Computer Science.

8. Conclusions.

Having recently learned of Oracle Big Data Lite (2014) effort, we see that we have been trying to do what Oracle has produced using `R` on linux. Our next step is to get access to an install or VM of Oracle Linux.

Other software that we plan to investigate are SAS University Edition (2014) and Teradata University Network (2014). Both of these new free software platforms look promising.

Learning about *Big Data* and the computing software related to accessing and analyzing *Large Data* sets is now possible within the usual Statistics curriculum.

Statistics education needs to find a place for introducing these ideas to the students.

REFERENCES

Data expo, Airline on-time performance, (2009). ASA Section on Computing and Statistical Graphics, http://stat-computing.org/dataexpo/2009/.

Cleveland, W. S. (2001), "Data Science: an Action Plan for Expanding the Technical Areas of the Field of Statistics," *International Statistical Review*, 69, 2126.

Heritage Provider Network Health Prize (2012), https://www.heritagehealthprize.com/c/hhp.

Hey, T., Tansley, S., and Tolle, K. (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Redmond, WA: Microsoft Research.

Hortonworks (2014), http://hortonworks.com.

kaggle.com (2014), "Titanic: Machine Learning from Disaster," https://www.kaggle.com/c/titanic-gettingStarted.

KeePass (2014), http://keepass.info/.

Leada (2014), http://www.teamleada.com/.

Melnykov, V., Chen, W. and Maitra, R. (2012). "MixSim: An R Package for Simulating Data to Study Performance of Clustering Algorithms," Journal of Statistical Software, 51, http://www.jstatsoft.org/v51/i12/.

NetFlix Prize (2009), http://www.netflixprize.com/.

Nolan, D. and Temple Lang, D. (2010), Computing in the Statistics Curricula, *The American Statistician*, 64, 97-107.

NYC Open Data (2014), https://nycopendata.socrata.com/.

Oracle R Distribution (2014), http://www.oracle.com/technetwork/database/database-technologies/r/r-distribution/overview/index.html.

Oracle Big Data Lite (2014), http://www.oracle.com/technetwork/database/bigdata-appliance/oracle-bigdatalite-2104726.html.

Rattle (2014), http://rattle.togaware.com/.

Revolution Analytics (2014), http://www.revolutionanalytics.com/.

SAS University Edition (2014), http://www.sas.com/en_us/software/university-edition.html.

tableau (2014), http://www.tableausoftware.com/.

Teradata University Network (2014), http://www.teradatauniversitynetwork.com/.

Whong, C. (2014), FOILing NYC's Taxi Trip Data, http://chriswhong.com/open-data/foil_nyc_taxi/.