Introducing Big Data in Stat 101 with Small Changes

John D. McKenzie, Jr.
Babson College
Babson Park, MA 02457-0310
mckenzie@babson.edu

DSI
Baltimore, MD
2013 November 18

1

Abstract

Today's technology produces massive amounts of data from a variety of sources such as social networking activities, financial transactions, genetic sequences, and astronomical transmissions. Very few introductory applied statistics courses consider such 'Big Data', for which many standard descriptive and inferential methods fail. This presentation will consider a number of ways that students can be easily exposed to the three V's of 'Big Data' (Volume, Velocity, and Variety) in such courses.

2

Agenda

- Big Data and its Three+ V's
- Standard Introductory Applied Course
- Big Data Sets
- Volume
- Velocity
- Varieties

3

2012 Mathematics Awareness Month

http://www.maa.org/mathematics-awareness-month-2012

4

Big Data in the News

- OSTP's Big Data Initiative (US$200,000,000)
  (nsf.gov – search on Big Data)
- McKinsey Global Institute Report (a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions)
- Big Data Special Issue of Significance Magazine (August 2012)
- NSA Disclosures,…

5

Bits and Bytes

Prefixes for multiples of bits (b) or **bytes** (B)				
Decimal				
Value	Metric			
1000	k	kilo		
1000^2	M	mega		
1000^3	G	giga		
1000^4	T	tera		
1000^5	P	peta		
1000^6	E	exa		
1000^7	Z	zetta		
1000^8	Y	yotta		
Binary				
Value	JEDEC		IEC	
1024	K	kilo	Ki	kibi
1024^2	M	mega	Mi	mebi
1024^3	G	giga	Gi	gibi
1024^4			Ti	tebi
1024^5			Pi	pebi
1024^6			Ei	exbi
1024^7			Zi	zebi
1024^8			Yi	yo

6

The Three V's of Big Data

- Volume
- Velocity
- Variety

META Group (now Gartner) analyst, Doug Laney

7

Introductory Applied Course

Terminology and Sampling Methods
Descriptive Statistics (graphs and numeric measures)
Basic Probability
Fundamental Inference
Advanced Topics

Only one course (De Veaux)

8

Volume

- Massive Data Sets
- Practice Significance
- Visualization

9

Big Data Sets

http://www.kdnuggets.com/datasets/
Over 60 Data Repositories
and
growing
Data Mining Competitions
KDD Cup Results Summary

10

Practical Significance

p-value > .05 from one-sample z-test and
versus
p-value = .000 from one-sample z-test with same sample mean and standard deviation but a 1000 times the sample size

Doane and Steward (2009), Applied Statistics in Business & Economics
pp. 364, 371, 374, 404, and 594 reinforcement

11

Practical Significance 2

Chi-Square Test of Independence

100	60
90	70

with p-value of .255 to
a p-value of .000 for

1000	600
900	700

12

Data Visualization

A visualization created by IBM of Wikipedia edits. At multiple terabytes in size, the text and images of Wikipedia are a classic example of big data

13

Data Visualization

Twitter Mentions

14

Velocity

- Time Series Data
- Process Data

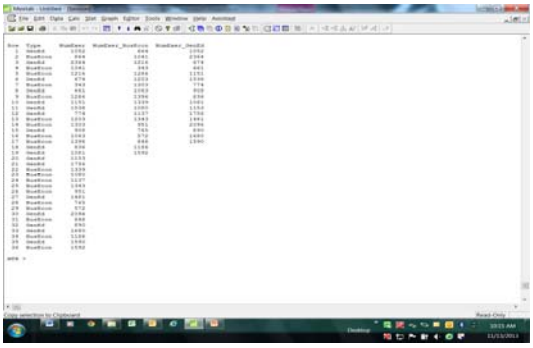
15

Variety (structure)

- Two Sample Data
- Missing Data
- Messy Data
- Text Data
- Date and Time Data

16

Variety: Two Sample Data



Text Data: Word Cloud

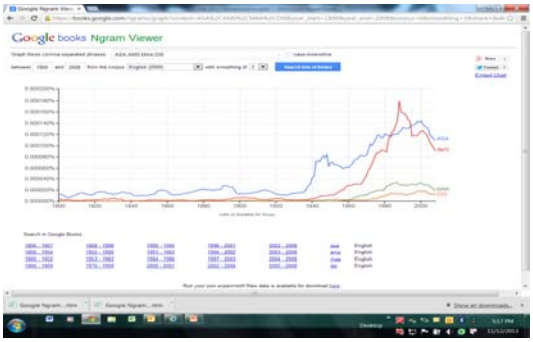
18

Text Data: Word Cloud

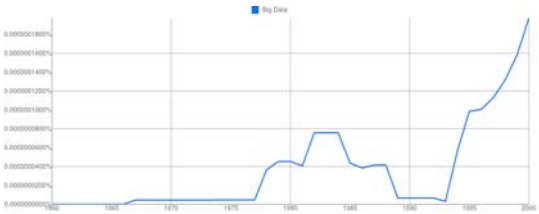
DSI Constitution and By-Laws

20

Text Data: N-Gram



Big Data, Business Analytics, Predictive Analytics , …, Data Science



22

Variety (sources)

http://www.amstat.org/publications/jse/jse_data_archive.htm: JSE Data Archive
http://www.causeweb.org/cwis/SPT--BrowseResources.php?ParentId=5: CAUSE Data Sets
http://stat-computing.org/dataexpo : ASA Statistical Computing and Statistical Graphics Bi-Annual Data Exposition
http://www.kdnuggets.com/datasets : Datasets for Data Mining
http://www.data.gov : U.S. Government Data
http://data.worldbank.org/ : The World Bank Data
http://bitly.com/bundles/hmason/1 : Research-Quality Data Sets
http://aws.amazon.com/big-data : Big Data on Amazon Web Services
http://www.bigdata-startups/public-data: 14 Sources of Public Data Sets
http://es.slideshare.net/CengageLearning/mark-frydenberg-drinking-from-the-fire-hose : "Big Data: What It Is and How You Can Use It" slide show
https://developers.google.com/fusiontables/ : Google experimental application that lets you store, share, query, and visualize data tables
https://developers.google.com/bigquery/ : Google site to interactively analyze massive datasets
http://citizen-statistician.org/ : Learning to Swim in the Data Deluge Blog
http://www.williams.edu/feature-stories/visualizing-the-liberal-arts/ : Williams College Majors and Employment

23

Future Introductory Course

Math Common Core State Standards
will result in
Remedial Sections?
Today's Course with More Topics?
Today's Second Core?
Big Data Analytics Course?
or ?

24

Two Current Examples of Analytics

Sharpe, De Veaux, and Velleman (2012), Business Statistics, Second Edition, Chapter 25, Introduction to Data Mining (Paralyzed Veterans of America)

Berenson, Levine, and Krehbiel (2012), Basic Business Statistics, Twelfth Edition, Online Topic: Analytics and Data Mining

2015?

25