

## 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 1024 (SI or bytes B)			
Decimal		Binary	
Value	SI	IEC	IEC
1000	k	kibi	kibi
1000 <sup>2</sup>	M	mebi	mebi
1000 <sup>3</sup>	G	gibi	gibi
1000 <sup>4</sup>	T	tebi	tebi
1000 <sup>5</sup>	P	pebi	pebi
1000 <sup>6</sup>	E	exbi	exbi
1000 <sup>7</sup>	Z	zebi	zebi
1000 <sup>8</sup>	Y	yobi	yobi
1024	K	Ki	Ki
1024 <sup>2</sup>	M	Mi	Mi
1024 <sup>3</sup>	G	Gi	Gi
1024 <sup>4</sup>		Ti	Ti
1024 <sup>5</sup>		Pi	Pi
1024 <sup>6</sup>		Ei	Ei
1024 <sup>7</sup>		Zi	Zi
1024 <sup>8</sup>		Yi	Yi

6

### The Three V's of Big Data

- Volume
- Velocity
- Variety

META Group (now Gartner) analyst, Doug Laney

### Introductory Applied Course

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

Only **one** course (De Veaux)

### Volume

- Massive Data Sets
- Practice Significance
- Visualization

### Big Data Sets

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

### 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

### 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

### 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

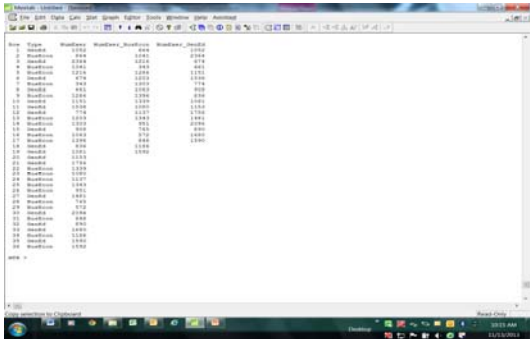
15

### Variety (structure)

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


16

### Variety: Two Sample Data



17

### Text Data: Word Cloud



18



## 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