

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

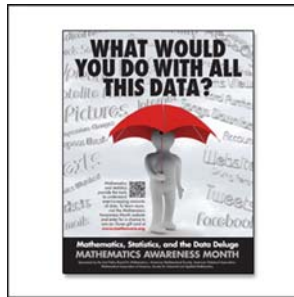
Agenda

- 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	Metric	IEC	IEC
1000	k	kibi	kibi
1000 ²	M	mebi	mebi
1000 ³	G	gibi	gibi
1000 ⁴	T	tebi	tebi
1000 ⁵	P	pebi	pebi
1000 ⁶	E	exbi	exbi
1000 ⁷	Z	zebi	zebi
1000 ⁸	Y	yobi	yobi
1024	K	Ki	Ki
1024 ²	M	Mi	Mi
1024 ³	G	Gi	Gi
1024 ⁴		Ti	Ti
1024 ⁵		Pi	Pi
1024 ⁶		Ei	Ei
1024 ⁷		Zi	Zi
1024 ⁸		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

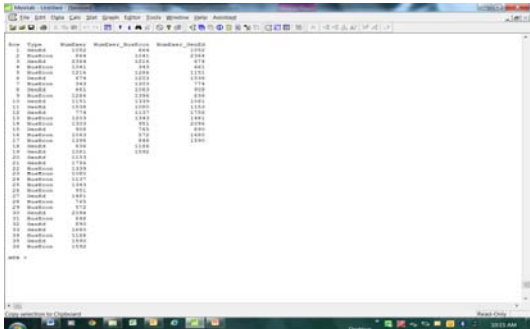
15

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

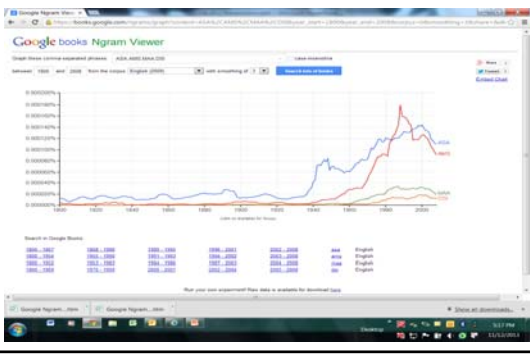
Text Data: Word Cloud



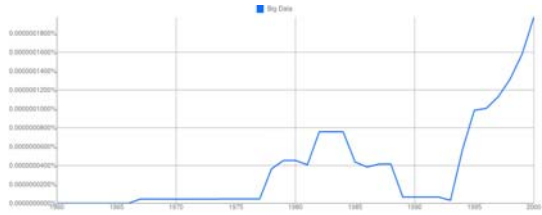
DSI Constitution and By-Laws



Text Data: N-Gram



Big Data, Business Analytics, Predictive Analytics , ..., Data Science



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

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 ?

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

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

Abstract

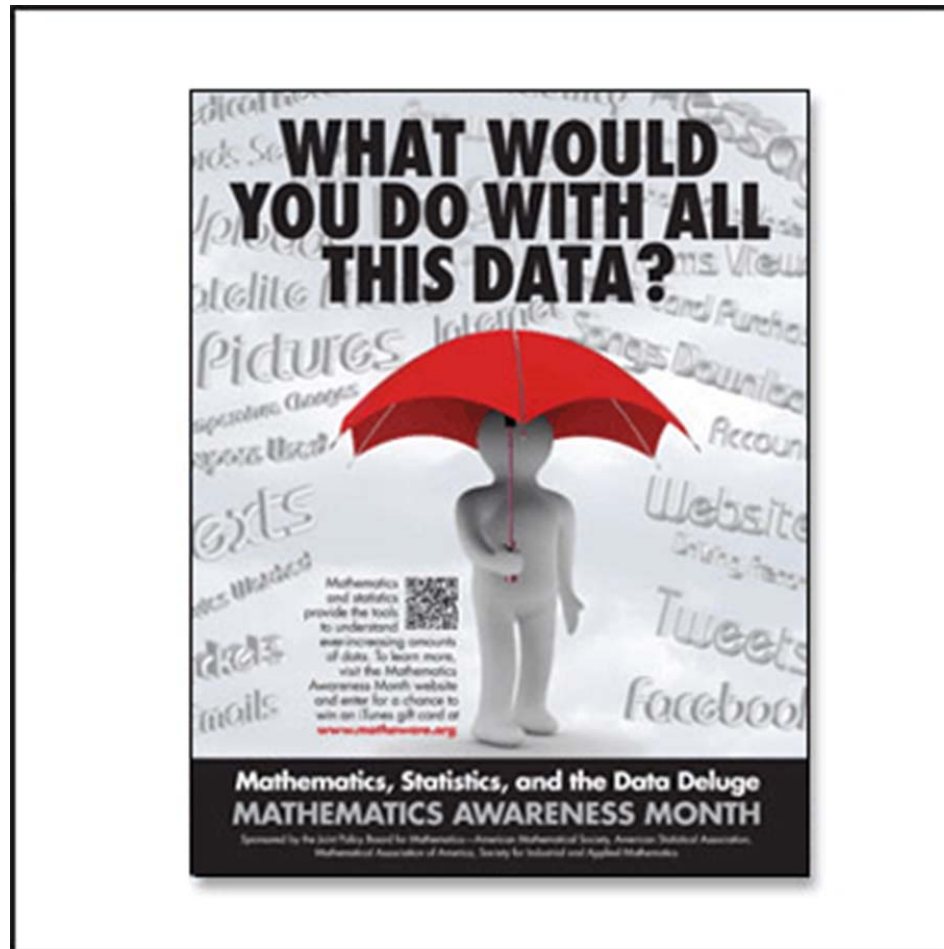
Today's technology produces massive amounts of data from a variety of sources such as social networking activities, financial transactions, genetic sequences, and astronomical transmissions. Very few introductory applied statistics courses consider such 'Big Data', for which many standard descriptive 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.

Agenda

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

2012 Mathematics Awareness Month

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



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,...

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

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



Data Visualization

Twitter Mentions



Velocity

- Time Series Data
- Process Data

Variety (structure)

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

Variety: Two Sample Data

Minitab - Untitled - [Session]

File Edit Data Calc Stat Graph Editor Tools Window Help Assistant

Row	Type	NumExer	NumExer_BusEcon	NumExer_GenEd
1	GenEd	1052	864	1052
2	BusEcon	864	1041	2364
3	GenEd	2364	1216	674
4	BusEcon	1041	343	661
5	BusEcon	1216	1286	1151
6	GenEd	674	1203	1538
7	BusEcon	343	1303	774
8	GenEd	661	1063	908
9	BusEcon	1286	1396	836
10	GenEd	1151	1339	1081
11	GenEd	1538	1080	1153
12	GenEd	774	1137	1756
13	BusEcon	1203	1343	1481
14	BusEcon	1303	951	2096
15	GenEd	908	765	890
16	BusEcon	1063	572	1680
17	BusEcon	1396	848	1590
18	GenEd	836	1186	
19	GenEd	1081	1592	
20	GenEd	1153		
21	GenEd	1756		
22	BusEcon	1339		
23	BusEcon	1080		
24	BusEcon	1137		
25	BusEcon	1343		
26	BusEcon	951		
27	GenEd	1481		
28	BusEcon	765		
29	BusEcon	572		
30	GenEd	2096		
31	BusEcon	848		
32	GenEd	890		
33	GenEd	1680		
34	BusEcon	1186		
35	GenEd	1590		
36	BusEcon	1592		

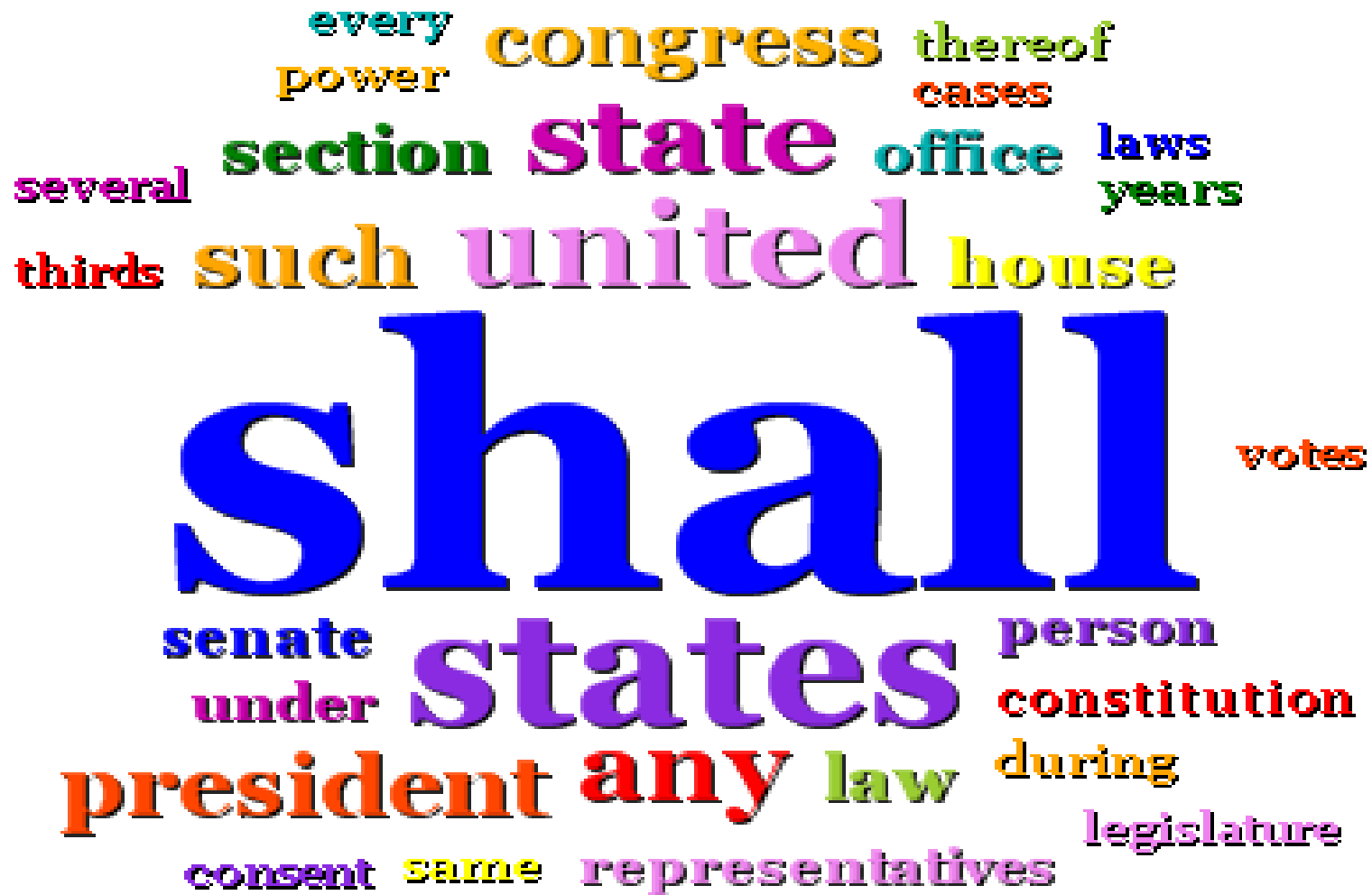
MTB >

Copy selection to Clipboard

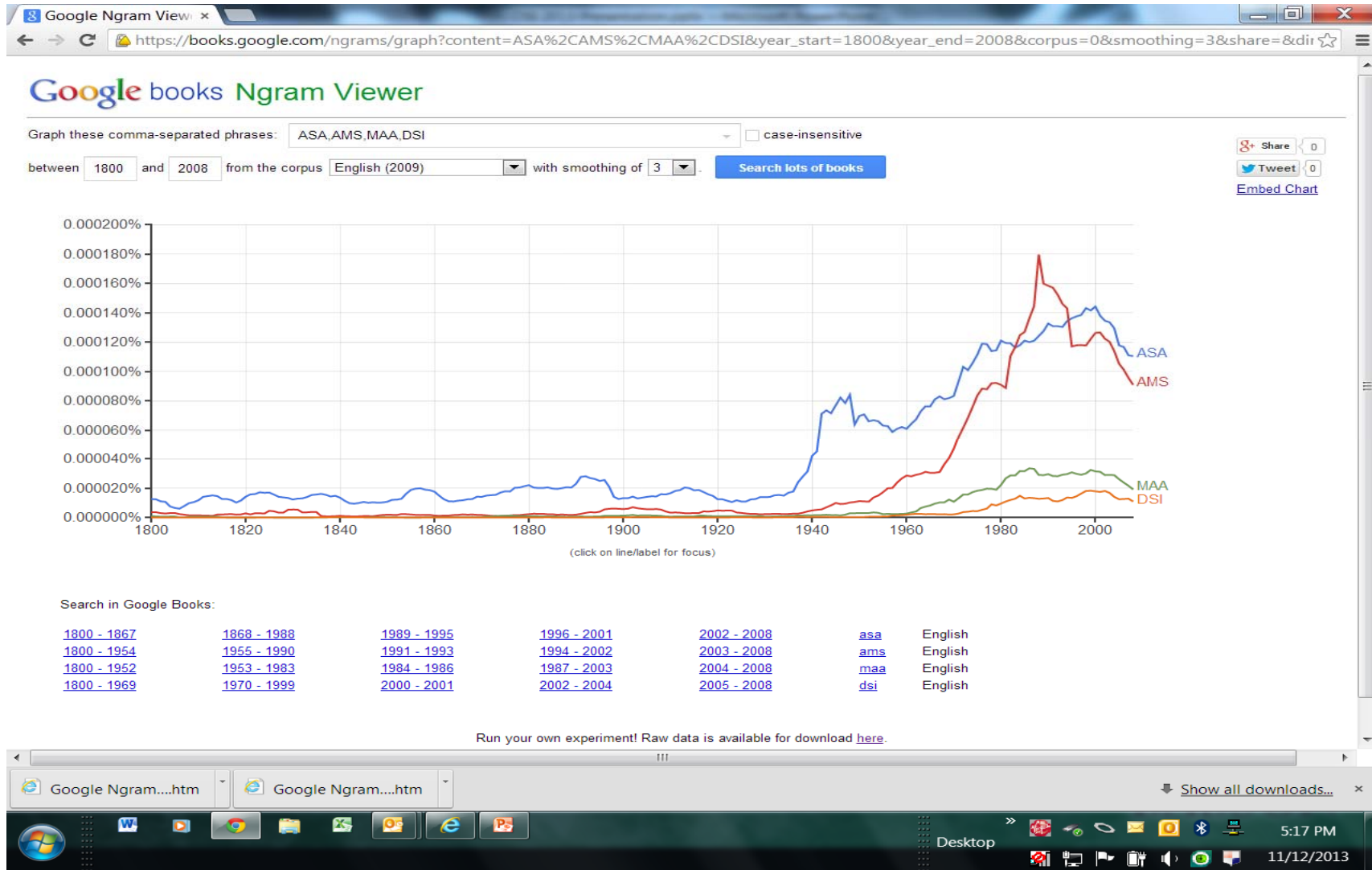
Read-Only

Desktop 10:15 AM 11/13/2013

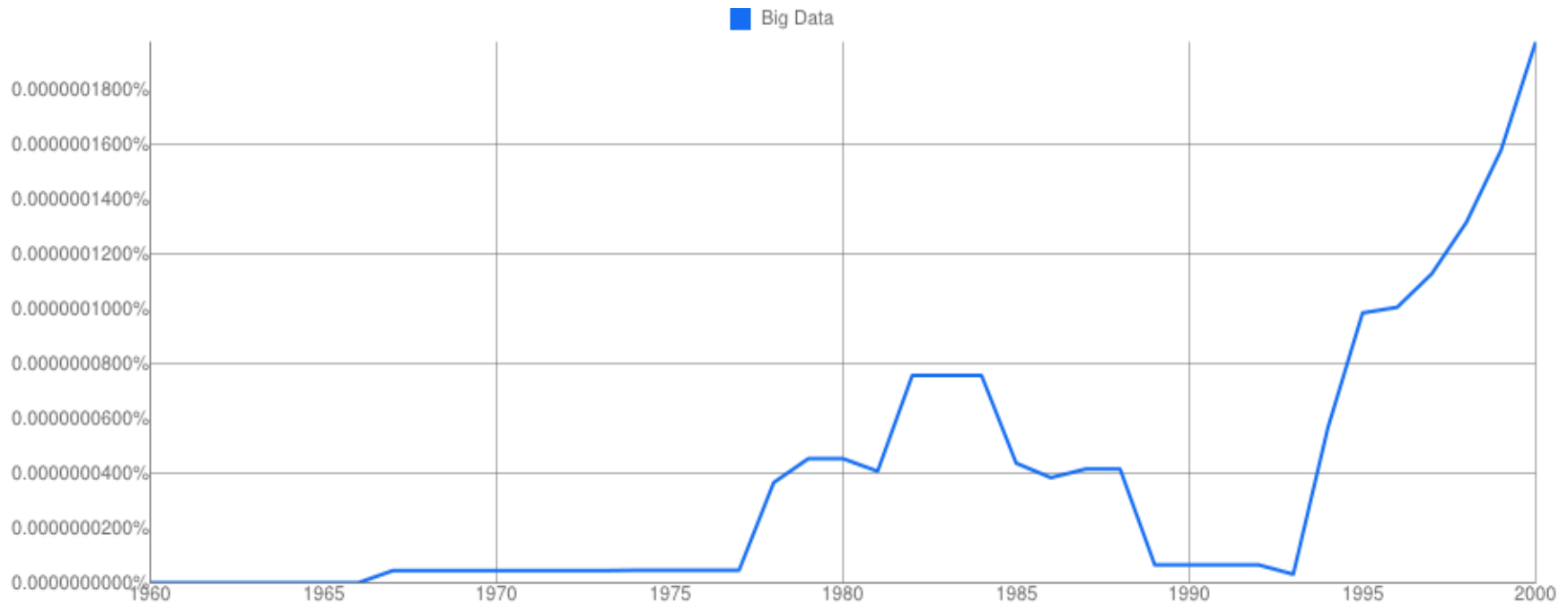
Text Data: Word Cloud



Text Data: N-Gram



Big Data, Business Analytics, Predictive Analytics , ..., Data Science



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

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 ?

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?