CARLSON SCHOOL OF MANAGEMENT
UNIVERSITY OF MINNESOTA

MSBA 6330
Big Data Analytics

Professor De Liu

1

Carlson School of Management

Agenda

- Instructor
- Introduction to Big Data
- Syllabus

2

Carlson School of Management

Course introduction

INSTRUCTOR

3

Carlson School of Management

About the Instructor

- Dr. De Liu (刘德)
- Originally from Shandong Province, China
- Associate Professor & 3M Fellow in Business Analytics
- Research interests:

4

Carlson School of Management

Course introduction

START FROM A USE CASE

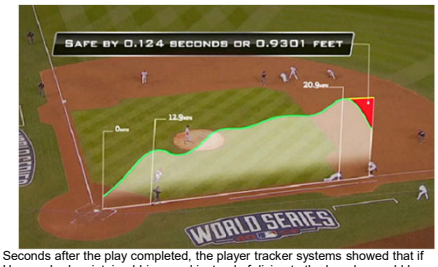
5

Carlson School of Management

How big data helps Major League Baseball (MLB)?

Carlson School of Management

Player Tracking Systems Powered by Big Data



Seconds after the play completed, the player tracker systems showed that if Hosmer had maintained his speed instead of diving to the bag, he would have been safe by about a foot

Carlson School of Management

Behind the scene

- Data capturing
  - A Doppler radar system sits behind home plate, sampling ball position 2000 times a second.
  - Two stereoscopic imaging devices, sampling positions of players on the field 30 times a second.
  - Brief written notes of each play entered by personnel on the field after the action is over
  - ~ 30 JSON docs per second per game, 7TB per game.
- Data transmission
  - Seconds after a player is completed, data is transmitted from stadium to cloud servers

Carlson School of Management

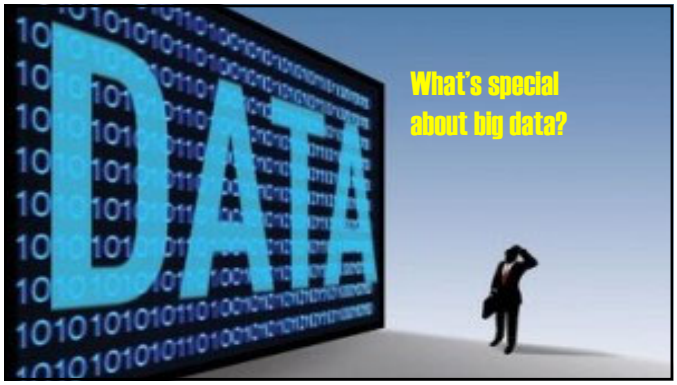
Behind the scene (cont.)

- Data analytics
  - within milliseconds of data transmission, parallel processing of data began
  - e.g. measuring player speed, forecasting/what if analysis, visualization
- Delivering results
  - Results of analysis are delivered to the Internet destinations
  - e.g. customers' mobile phones and broadcaster's monitors

Carlson School of Management

More Use Cases of Big Data

- United Healthcare mines customer calls
  - Turn voice data into text, then analyze it with Natural Language Processing (NLP) software to detect consumer attitudes, using Hadoop and NoSQL.
- Medtronic: Using Hadoop + Spark + R to achieve 50+ speed up than SQL Server in analyzing billions of clinical observations to predict heart failure
- NeuroID: Loan fraud detection by analyzing real-time mouse-movements data
  - Analyzing mouse trajectory when people fill out loan application forms online to flag fraudulent cases (Hibbeln et al 2014). Data is streamed and analyzed on Amazon cloud.



Carlson School of Management

Course introduction

BIG DATA: CONCEPT AND OPPORTUNITIES

12

Carlson School of Management

What is big data?

Big Data: "large volumes of high velocity, complex, and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management and analysis of the information

-- from a US Congress Report in 2012

What are the characteristics of big data?

- Volume
- Velocity
- Variety

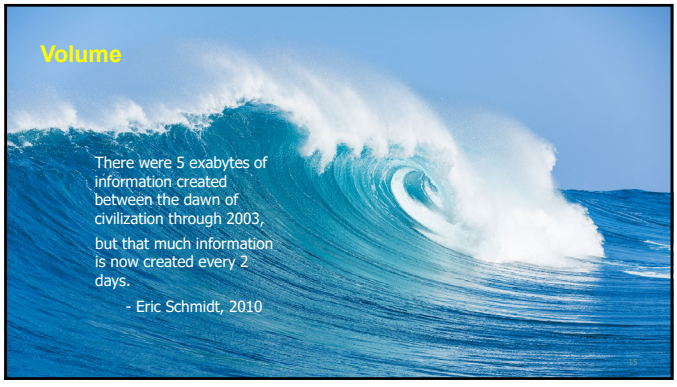
Doug Laney, Gartner (2001)

- Veracity
- Value

Bernad Marr, "Big Data" (2015)

D. Laney, 3D data management: controlling data volume, velocity and variety, META Group Res. Note 6 (2001) 70.
B. Marr, Big Data: Using SMART Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance, John Wiley & Sons, 2015.

Volume



There were 5 exabytes of information created between the dawn of civilization through 2003,

but that much information is now created every 2 days.

- Eric Schmidt, 2010

Carlson School of Management

Some Examples of Big Data

- Every day…
  - Over 2.25 billion shares are traded on the New York Stock Exchange
  - Facebook stores 4.5 billion "Likes"
  - Google processes about 24 petabytes of data
- Every minute…
  - Facebook users share nearly 2.5 million pieces of content
  - Email users send 204,000,000 messages

24 petabytes =  *X 24,576*

16

Carlson School of Management

How big is "big"?

- 50% consider datasets between Terabyte and Petabye to be big.
- Whatever is considered "high volume" today will be even higher tomorrow.

Specific units of IEC 60027-2 A.2 and ISO/IEC 80000

IEC prefix		Representations				Customary prefix	
Name	Symbol	Base 2	Base 1024	Value	Base 10	Name	Symbol
kibi	Ki	2^{10}	1024^1	1024	$\approx 1.02 \times 10^3$	kilo	k[13] or K
mebi	Mi	2^{20}	1024^2	1 048 576	$\approx 1.05 \times 10^6$	mega	M
gibi	Gi	2^{30}	1024^3	1 073 741 824	$\approx 1.07 \times 10^9$	giga	G
tebi	Ti	2^{40}	1024^4	1 099 511 627 776	$\approx 1.10 \times 10^{12}$	tera	T
pebi	Pi	2^{50}	1024^5	1 125 899 906 842 624	$\approx 1.13 \times 10^{15}$	peta	P
exbi	Ei	2^{60}	1024^6	1 152 921 504 606 846 976	$\approx 1.15 \times 10^{18}$	exa	E
zebi	Zi	2^{70}	1024^7	1 180 591 620 717 411 303 424	$\approx 1.18 \times 10^{21}$	zetta	Z
yobi	Yi	2^{80}	1024^8	1 208 925 819 614 629 174 706 176	$\approx 1.21 \times 10^{24}$	yotta	Y

17

Carlson School of Management

How much data?

- There are about 5,000,000 articles in the English Wikipedia 2015. How much data is that
  - if the articles are stored in plain text (compressed)? *11.5 GB*
  - If the articles and edit histories are stored in XML text (compressed)? *100 GB*
  - If the articles and edit histories are stored in XML text (uncompressed)? *10 TB*

Source: https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia

18

Carlson School of Management

Opportunities of Big Volume

"It's not who has the best algorithm that wins, it's who has the most data" (Andrew Ng)

20

Carlson School of Management

The Challenge of Big Volume

- Question: How much time does it take to read one Terabyte of data from hard disk into memory?

1 TB = 1024 GB = 1024*1024 MB = 1,048,576 MB

1,048,576/100/3600 = 2.91 hour

21

Carlson School of Management

Velocity

- Velocity: Data in motion
  - The speed at which data is created processed and analyzed continues to accelerate.
- Examples of high velocity data:
  - Twitter processes 340 million messages / per day
  - Trend Micro processes 6 TB of data/day to identify new security threats
  - Financial institutions process more than 10,000 credit card transactions/second
  - Amazon Web Services fields more than 650,000 requests / second
  - Large Hadron Collider produces 572 terabytes of data per second
  - MLB game generates 2.5 Terabytes / hour.

22

Carlson School of Management

Fantastic velocity and where to find them

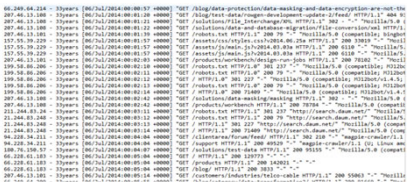
- Can you think of an every-day example of high velocity data around you?

23

Carlson School of Management

An example of high velocity data

- Imagine you work for an e-commerce company that has 5 TB of web log data per day



- How do you analyze such data?
  - Assume a Gigabit network

Why are we interested in processing log data?

25

Carlson School of Management

Let's Do the Math

Assuming a gigabit network, 1024 Mbps = 1Gbps

Sending 1 GB requires 8 seconds

Sending 1 TB requires 1024*8 /3600 = 2.27 Hours

Sending 5 TB requires 5*1024*8/3600 = 12 Hours!

26

Carlson School of Management

Variety

- **Variety**: the complexity of multiple data types, including structured, semi-structured and unstructured data.
- They are also different forms:
  - **Structured**: transactional
  - **Semi-structured**: sensor data, logs, RFID
  - **Unstructured**: reviews, images, tweets, audio, video
- Inside or outside of enterprises
  - **Internal**: transactional systems, server logs, emails, chats, etc
  - **External**: social media, sensor networks, weather data, geographic data, census, macroeconomic data, third party data providers

27

Carlson School of Management

Fantastic variety and where to find them

- Can you think of an every-day example of high variety data around you?

28

Carlson School of Management

Data Variety in Health Care

- 80% of information in healthcare industry is unstructured data, e.g.
  - Outputs from medical devices
  - Doctor's notes
  - Lab results
  - Medical imaging
  - Medical correspondence
  - Clinical data
  - Patient behavior and sentiment data
  - Genomic data

Carlson School of Management

Opportunity of high variety data

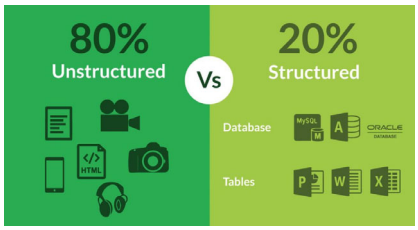
- Half of the battle is to get quality signals
- Big data provides a way to capture and analyze novel data sources (e.g. social media, click stream, imagery, sensor data)

31

Carlson School of Management

Challenges associated with variety

- Many earlier data technologies are not flexible enough to deal with large variety of semi- or un-structured data



32

Carlson School of Management

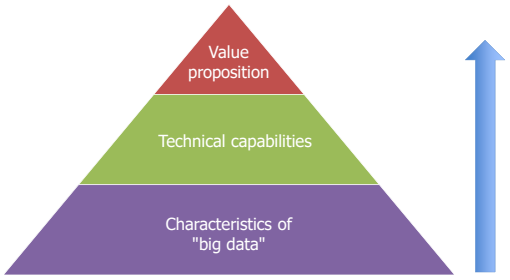
What about 2 other V's

- **Verasity** refers to data uncertainty
  - Large volumes of disparate data being ingested at high speed are only useful if the information is correct. Incorrectly indexed data or spelling mistakes could make complete datasets useless and thus the veracity is important.
- **Value**: Big data has many valuable applications
  - Value is a multifaceted property of big data. As the volume of data grows the incremental value of each data point begins to decrease. As the variety of data available increases, not all the data may aid in product development, sales, or system management. Big data is not the retention of all data; some data needs to remain volatile.

33

Carlson School of Management

Organizational Implications of 5 V's



34

Carlson School of Management

Course introduction

BIG DATA: HOW ARE COMPANIES USING BIG DATA?

35

36

Carlson School of Management

How Are Companies Using Big Data?

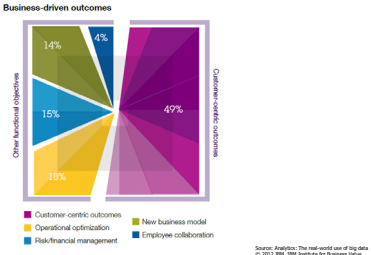
- Big data has reached a point of **mainstream adoption** within Fortune 1000 firms
  - In 2016, 62.5% have at least one instance of big data in production. In 2018, 97.2% are investing in building or launching big data and AI initiatives
- **Chief Data Officer (CDO)** is well established
  - 54% named a CDO in 2016, compared to 12% in 2012.
  - 62.5% named a CDO in 2018, compared to 12% in 2012
  - "Data is essentially the new oil, and the CDO is beginning to be recognized as the linchpin for tackling one of the most important problems in enterprises today: driving value from data"

Sources: NewVantage. 2018. "Big Data Executive Survey 2018," *NewVantage Partner Report.*

Carlson School of Management

How Are Companies Using Big Data? (continue)

- Customer-centric activities are the top priority



Source: Analytics: The real-world use of big data
© 2012 IBM, IBM Institute for Business Value

38

Carlson School of Management

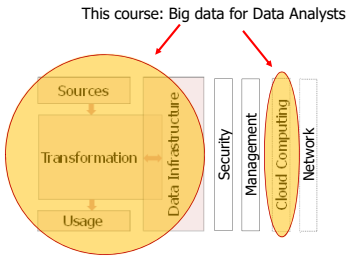
Course introduction

THE FOCUS OF THIS COURSE

39

Carlson School of Management

Where does this course fit?



Source: Geoffrey Fox and Wo Chang, CCL Meeting Presentation, 2013

40

Carlson School of Management

Course Topics

- Learn through hands-on examples
  - Hadoop: MapReduce/HDFS/YARN
  - Data ingestion: Scoop
  - Data analysis / ETL: Hive
  - Spark: Core Spark, Spark SQL,
  - Machine Learning: Spark MLlib
  - Streaming: Spark Streaming
  - Cloud computing: Amazon AWS

41

Carlson School of Management

Course Objectives

- Develop an understanding of the big data ecosystem, the kinds of problems it aims to solve, the characteristics of big data technologies, and their key advantages and disadvantages.
- Develop core competencies in using a variety of essentially big data tools (such as Scoop, Hive, Spark, and Cloud computing) and processes to solve data science problems at scale.

42