Big data

Subject: 805-182 Computer: The Internet and Society 3(3-0-6)

Semester 1/63

Dr. Norrathep Rattanavipanon

Faculty of International Studies, PSU Phuket.

Courtesy: Dr. Jirawat Thaenthong

Reminder

1. Your midterm score is out! I also published the midterm solution.
2. We have a group presentation on this Thursday
3. Progress report is due Sept 4th , the end of the day!

Contents

1. What is data?
2. What is information?
3. What is Big Data?
4. Evolution of Big Data
5. Characteristics of Big Data
6. Big Data Analytics
7. Case Study of Big Data
8. Summary

What is Data?

The quantities, characters, or symbols on which operations are performed by a computer, which may be transmitted in the form of digital electrical signals and recorded on magnetic, optical, or mechanical recording media.

Is *a student name* data?

Is *a student address* data?

What is information?

The result of applying data processing to data, giving it context and meaning.

1234567.89 is data.
"Your bank balance has jumped 8087% to $1234567.89" is information.

What is big data?

Forbes reports that every minute, users watch *4.15 million YouTube videos*, send *456,000 tweets*, post *46,740 photos* on Instagram, and there are *510,000 comments* posted and *293,000 statuses* updated on Facebook!

500+terabytes of new data => **Facebook’s database**, every day

Evolution of big data

When was the last time you guys remember using a floppy or a CD to store your data?

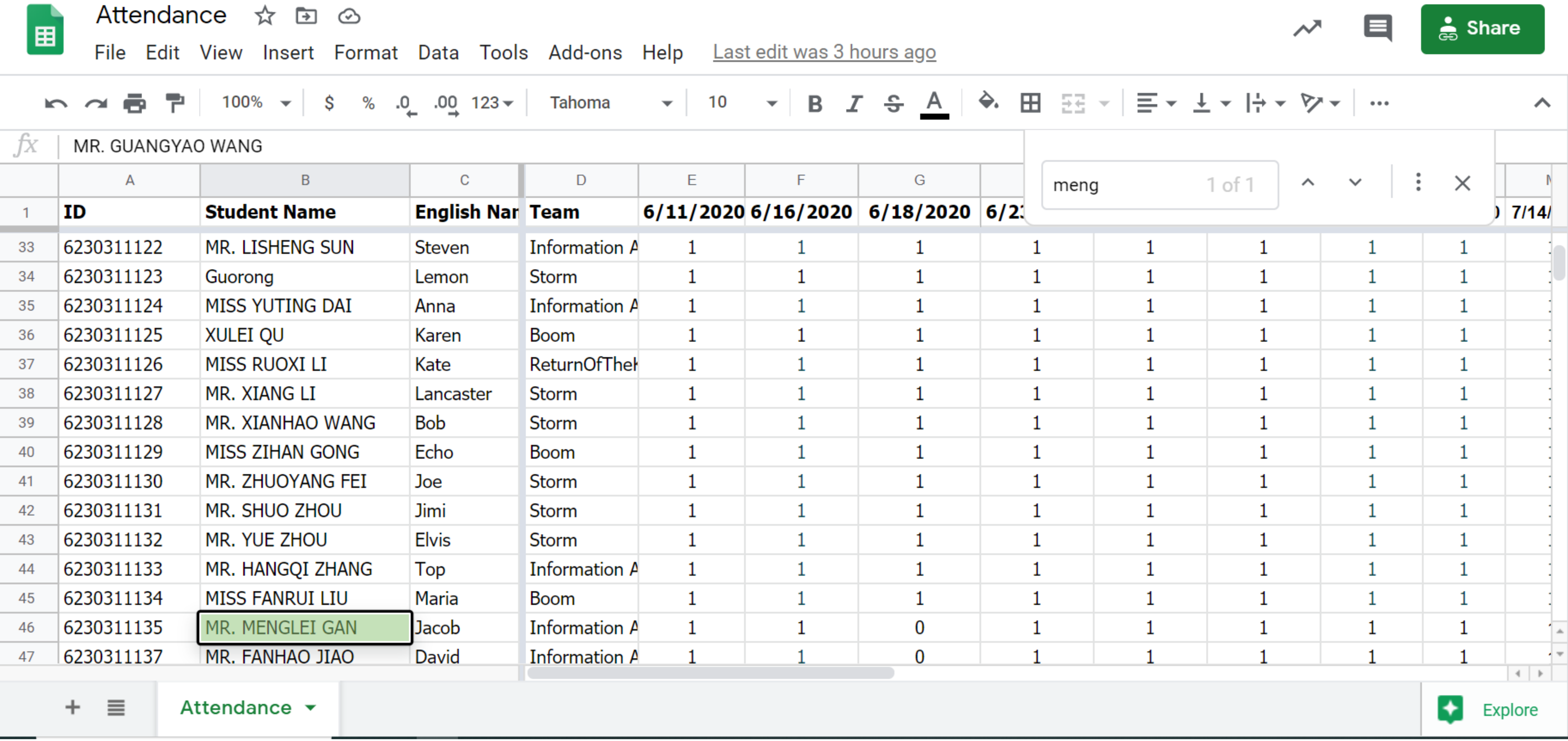
History of recorded data

1. Paper
2. Floppy disc
3. CD-ROM disc
4. Database in internal storage
5. External storage

Exponential Growth of Data

Storing their data in database systems
is insufficient with
the introduction of the internet, internet of things, and mobile technologies.

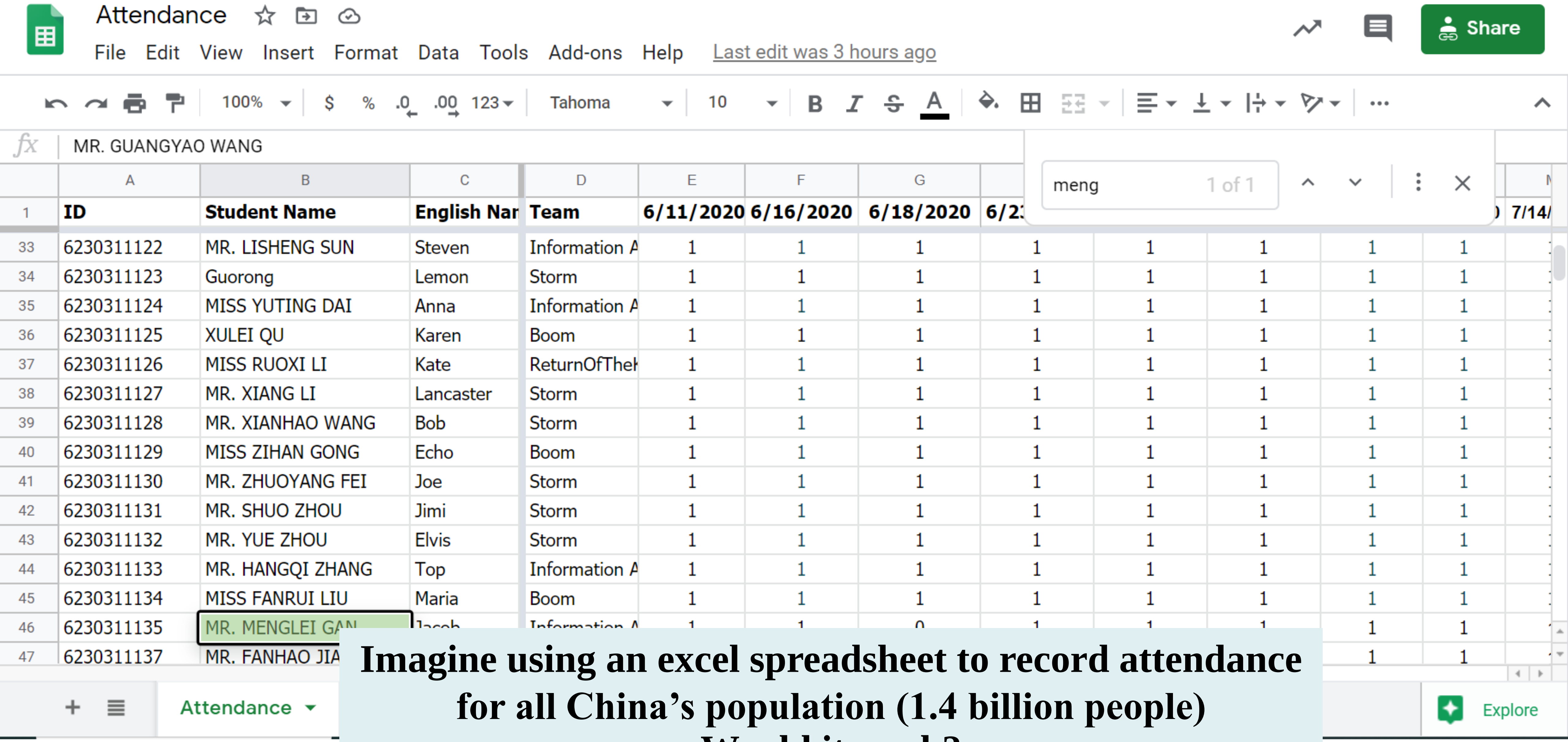
These technologies impact the generation of massive data.
It has become Big Data.



fx MR. GUANGYAO WANG

meng 1 of 1

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	ID	Student Name	English Nar	Team	6/11/2020	6/16/2020	6/18/2020	6/2					7/14/
33	6230311122	MR. LISHENG SUN	Steven	Information A	1	1	1	1	1	1	1	1	
34	6230311123	Guorong	Lemon	Storm	1	1	1	1	1	1	1	1	
35	6230311124	MISS YUTING DAI	Anna	Information A	1	1	1	1	1	1	1	1	
36	6230311125	XULEI QU	Karen	Boom	1	1	1	1	1	1	1	1	
37	6230311126	MISS RUOXI LI	Kate	ReturnOfTheK	1	1	1	1	1	1	1	1	
38	6230311127	MR. XIANG LI	Lancaster	Storm	1	1	1	1	1	1	1	1	
39	6230311128	MR. XIANHAO WANG	Bob	Storm	1	1	1	1	1	1	1	1	
40	6230311129	MISS ZIHAN GONG	Echo	Boom	1	1	1	1	1	1	1	1	
41	6230311130	MR. ZHUOYANG FEI	Joe	Storm	1	1	1	1	1	1	1	1	
42	6230311131	MR. SHUO ZHOU	Jimi	Storm	1	1	1	1	1	1	1	1	
43	6230311132	MR. YUE ZHOU	Elvis	Storm	1	1	1	1	1	1	1	1	
44	6230311133	MR. HANGQI ZHANG	Top	Information A	1	1	1	1	1	1	1	1	
45	6230311134	MISS FANRUI LIU	Maria	Boom	1	1	1	1	1	1	1	1	
46	6230311135	MR. MENGLEI GAN	Jacob	Information A	1	1	0	1	1	1	1	1	
47	6230311137	MR. FANHAO JIAO	David	Information A	1	1	0	1	1	1	1	1	



fx MR. GUANGYAO WANG

	A	B	C	D	E	F	G	H				
1	ID	Student Name	English Nar	Team	6/11/2020	6/16/2020	6/18/2020	6/2				)
33	6230311122	MR. LISHENG SUN	Steven	Information A	1	1	1	1	1	1	1	1
34	6230311123	Guorong	Lemon	Storm	1	1	1	1	1	1	1	1
35	6230311124	MISS YUTING DAI	Anna	Information A	1	1	1	1	1	1	1	1
36	6230311125	XULEI QU	Karen	Boom	1	1	1	1	1	1	1	1
37	6230311126	MISS RUOXI LI	Kate	ReturnOfTheK	1	1	1	1	1	1	1	1
38	6230311127	MR. XIANG LI	Lancaster	Storm	1	1	1	1	1	1	1	1
39	6230311128	MR. XIANHAO WANG	Bob	Storm	1	1	1	1	1	1	1	1
40	6230311129	MISS ZIHAN GONG	Echo	Boom	1	1	1	1	1	1	1	1
41	6230311130	MR. ZHUOYANG FEI	Joe	Storm	1	1	1	1	1	1	1	1
42	6230311131	MR. SHUO ZHOU	Jimi	Storm	1	1	1	1	1	1	1	1
43	6230311132	MR. YUE ZHOU	Elvis	Storm	1	1	1	1	1	1	1	1
44	6230311133	MR. HANGQI ZHANG	Top	Information A	1	1	1	1	1	1	1	1
45	6230311134	MISS FANRUI LIU	Maria	Boom	1	1	1	1	1	1	1	1
46	6230311135	MR. MENGLEI GAN	Jacob	Information A	1	1	0	1	1	1	1	1
47	6230311137	MR. FANHAO JIA									1	1

Imagine using an excel spreadsheet to record attendance for all China's population (1.4 billion people) Would it work?

To make clear, What is big DATA?

Big Data is also data but with a huge size.

Big Data is a term used to describe a collection of data that is huge in size and yet growing exponentially with time.

Problem: such data is so large and complex that the traditional data management tools cannot handle it efficiently.

Example of traditional data management tools

- Microsoft Excel
- Big data -> big Excel file (e.g., 1TB file)
- **Can you open this large file in Excel?**
- Data Size:

○ 1 MB = 1,000,000 Bytes

○ 1 GB = 1000 MB

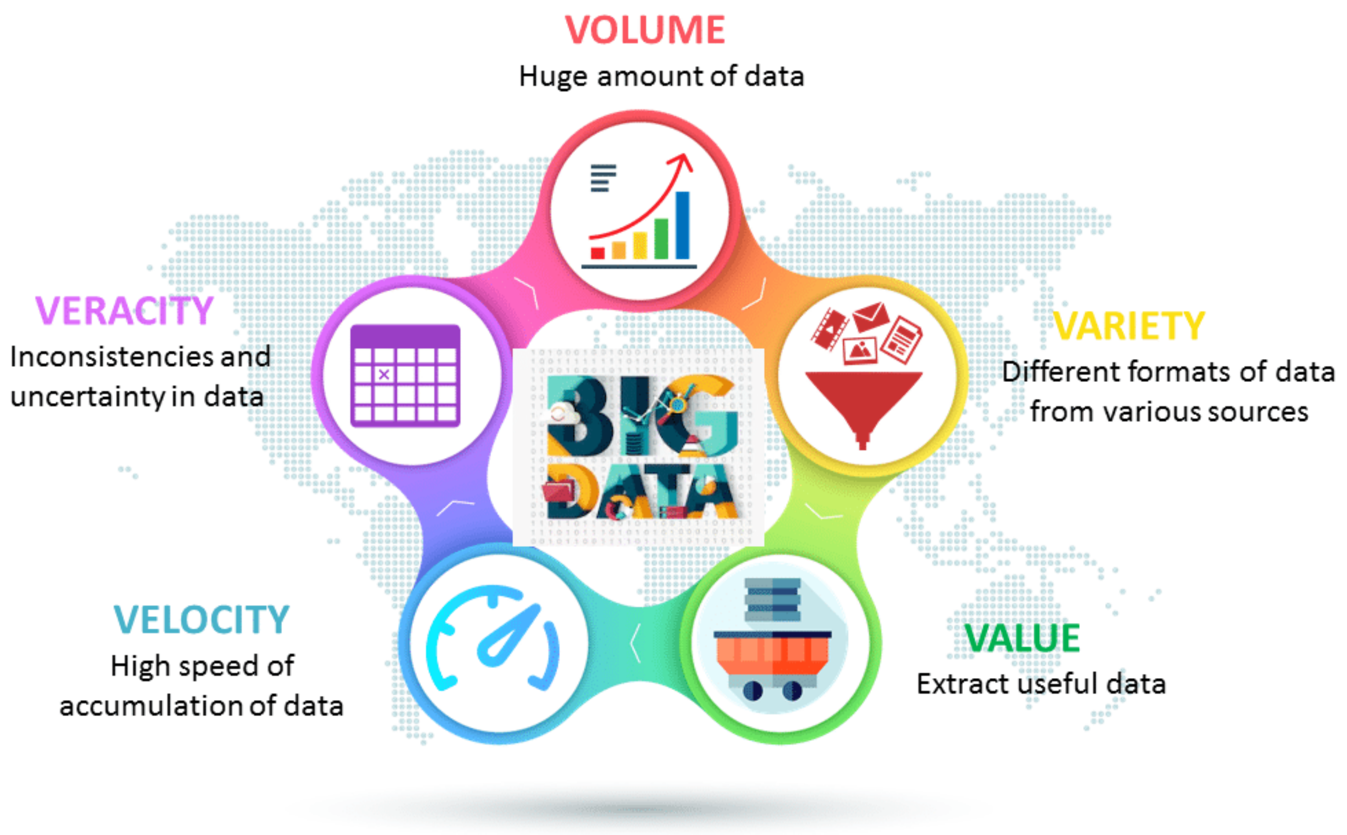
○ 1 TB = 1000 GB

○ 1 Petabyte = 1 PB = 1000 TB

Task: (10 minutes)

Give activities that generate Big Data in real-life?

- The New York Stock exchange
- Online Shopping Customer Data
- Weather Forecast
- Inventory Ordering
- Game user data
- Autonomous vehicles



VOLUME
Huge amount of data

VARIETY
Different formats of data from various sources

VALUE
Extract useful data

VELOCITY
High speed of accumulation of data

VERACITY
Inconsistencies and uncertainty in data

Characteristics of Big data

These are the following characteristics associated with Big Data:

1) Volume

(i) Volume – Data Size

The bigger it is, the harder to process/extract value out of it.

2) VARIETY

***(ii) Variety* –** Different formats of data from various sources.

Source of data:

- Past: spreadsheets and databases
- Present: **everything** (emails, photos, videos, IoT devices, etc.)

3) Velocity

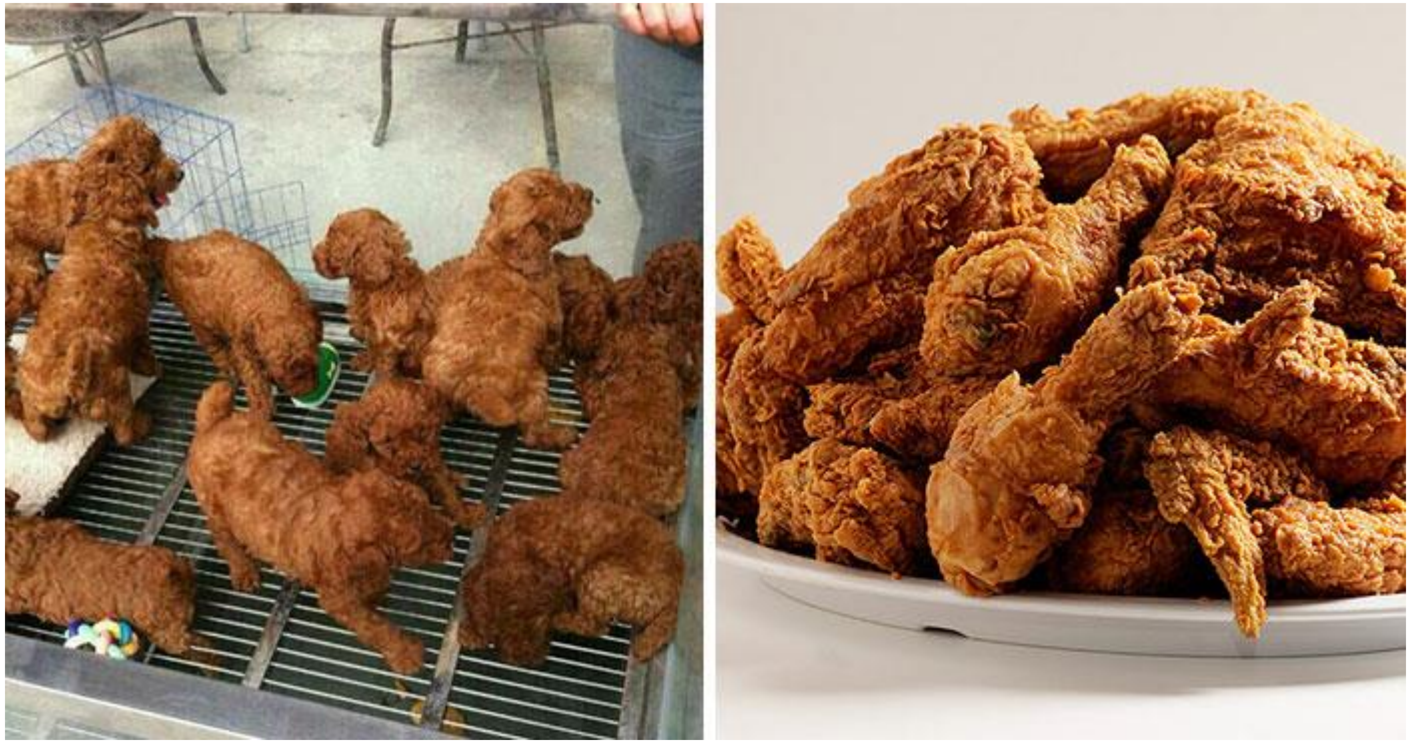
(iii) Velocity – speed of data generation

How fast the data is generated and processed to meet the demands.

For example: intersection traffic light data

4) Veracity

(iv) Veracity – Inconsistencies and uncertainty in data



Which photo contains fried chicken?

sciforce

Sources of Data Veracity

- Statistical biases
- Lack of data lineage
- Software bugs
- Noise
- Abnormalities
- Information Security
- Untrustworthy data sources
- Falsification
- Uncertainty and ambiguity of data
- Duplication of data
- Out of date and obsolete data
- Human error

5) Value

(v) Value – Value of data

Remember: we can access big data, but it will only be useful if we can turn it into something useful.

Big data Applications

Domains where Big Data Applications have been revolutionized:

1. Entertainment
2. Social Networking
3. Online Shopping – use Big Data to provide recommendation to users
4. Internet-of-thing
5. Healthcare
6. Travel
7. Bank industry

Etc.

Other domains?

Big data Analytics

Case Study:

Starbucks with Big Data

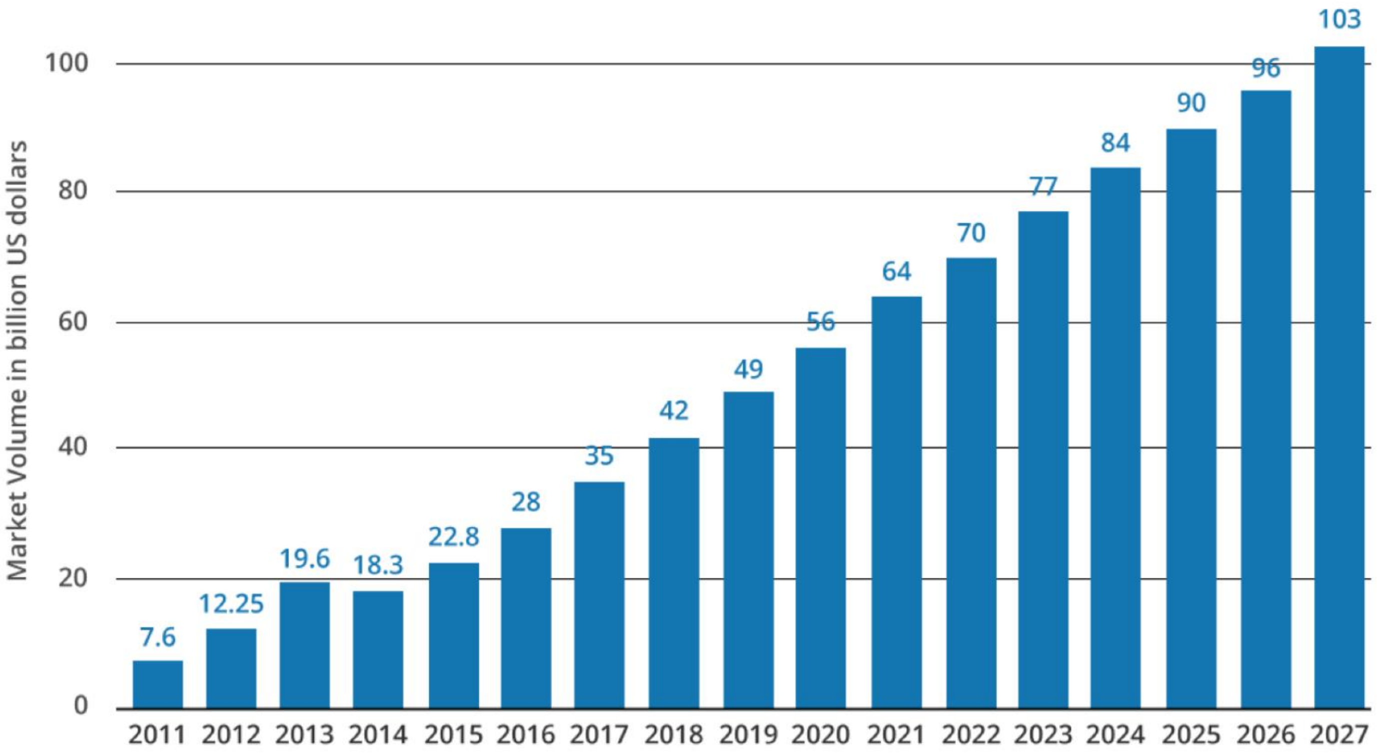
Premium service for customers

1. Data: coffee buying habits + time of day

2. Analyze data to give the barista their preferred order.

3. App will also suggest new products.

The growing market revenue of Big Data in billion U.S. dollars from the year 2011 to 2027



Credit: statista.com

Case Study of Big Data

Wait times in a theme park

Orange County, California, USA

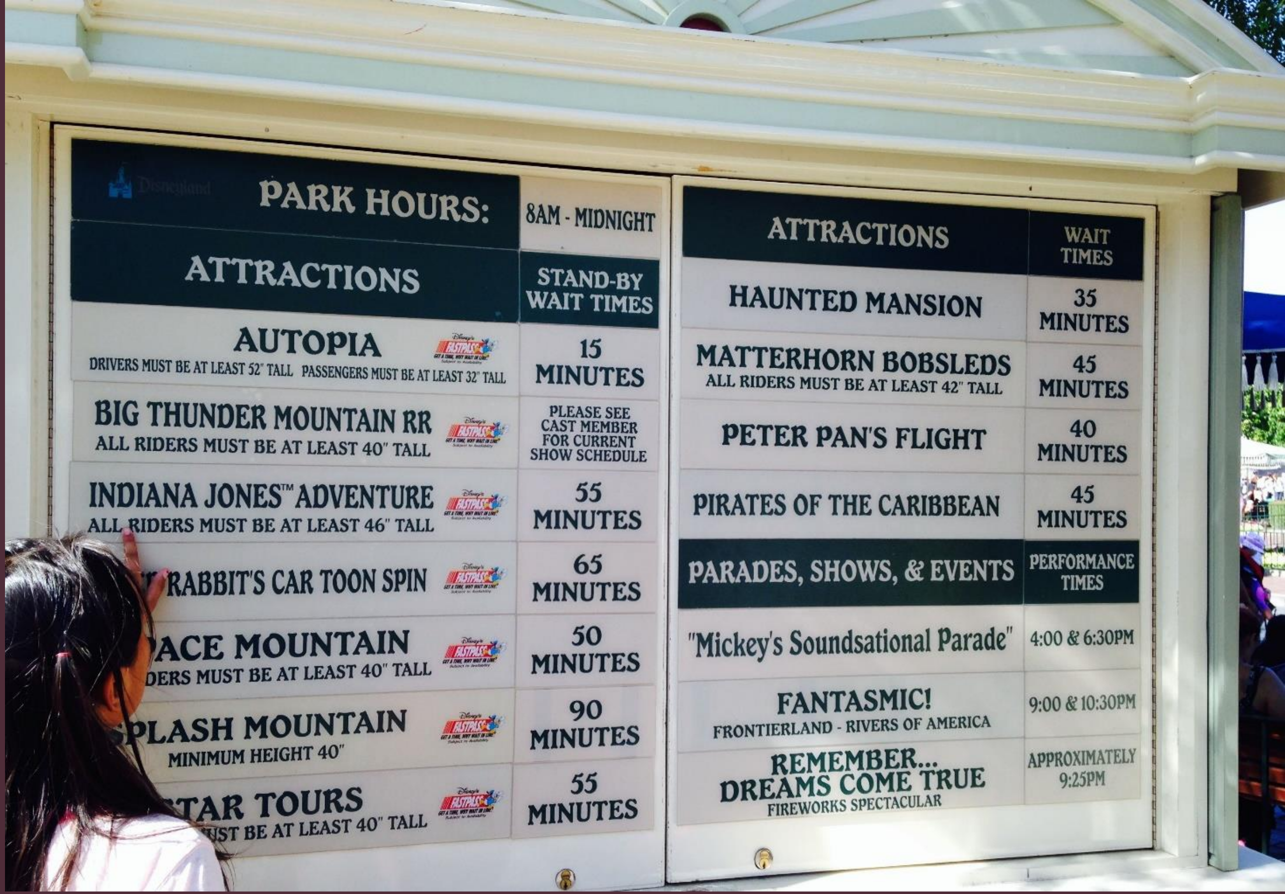
8/20/2020
DR. NORRATHEP RATTANAVIPANON

Disney Park Locations

- Orange County
- Florida
- Paris
- Hong Kong
- Shanghai
- Tokyo

Disney Parks: the good

8/20/2020
DR. NORRATHEP RATTANAVIPANON



PARK HOURS:	8AM - MIDNIGHT
ATTRACTIONS	STAND-BY WAIT TIMES
AUTOPIA DRIVERS MUST BE AT LEAST 52" TALL PASSENGERS MUST BE AT LEAST 32" TALL	15 MINUTES
BIG THUNDER MOUNTAIN RR ALL RIDERS MUST BE AT LEAST 40" TALL	PLEASE SEE CAST MEMBER FOR CURRENT SHOW SCHEDULE
INDIANA JONES™ ADVENTURE ALL RIDERS MUST BE AT LEAST 46" TALL	55 MINUTES
RABBIT'S CAR TOON SPIN	65 MINUTES
ACE MOUNTAIN DERS MUST BE AT LEAST 40" TALL	50 MINUTES
PLASH MOUNTAIN MINIMUM HEIGHT 40"	90 MINUTES
TAR TOURS UST BE AT LEAST 40" TALL	55 MINUTES

ATTRACTIONS	WAIT TIMES
HAUNTED MANSION	35 MINUTES
MATTERHORN BOBSLEDS ALL RIDERS MUST BE AT LEAST 42" TALL	45 MINUTES
PETER PAN'S FLIGHT	40 MINUTES
PIRATES OF THE CARIBBEAN	45 MINUTES
PARADES, SHOWS, & EVENTS	PERFORMANCE TIMES
"Mickey's Soundsational Parade"	4:00 & 6:30PM
FANTASMIC! FRONTIERLAND - RIVERS OF AMERICA	9:00 & 10:30PM
REMEMBER... DREAMS COME TRUE FIREWORKS SPECTACULAR	APPROXIMATELY 9:25PM

Disney Parks : the bad

How to avoid long wait times?

- Pay for a FASTPASS: $20 extra
- The FASTPASS line could be still long
- Maybe not worth it? 5 people = $100 extra on top of ticket price
- Better options?

Use Big Data to solve this problem

Big Data: Touringplans has recorded wait times for all Disney World's rides from 2012 to present

5V: Volume, Variety, Velocity, Veracity and Value of this Big Data

Volume of this Big Data

~40 million visitors each year for Disney World and 46 rides in Disney World

Challenge: Cannot store all data on a single Excel spreadsheet. Other ways to store it?

Big data database (e.g., Cassandra)

Variety of this Big Data

This Big Data could be generated by:

1. Images from cameras
2. Humans by manual counting number of people in line
3. Measuring the time it takes for each visitor to play the next ride

Challenge: How to store and process different variety of this Big Data?

Velocity of this Big Data

This big data is generated all the time when the park is open.

Challenge: Need to have some ways to provide ***real-time*** monitoring of people in line

Veracity of this Big Data

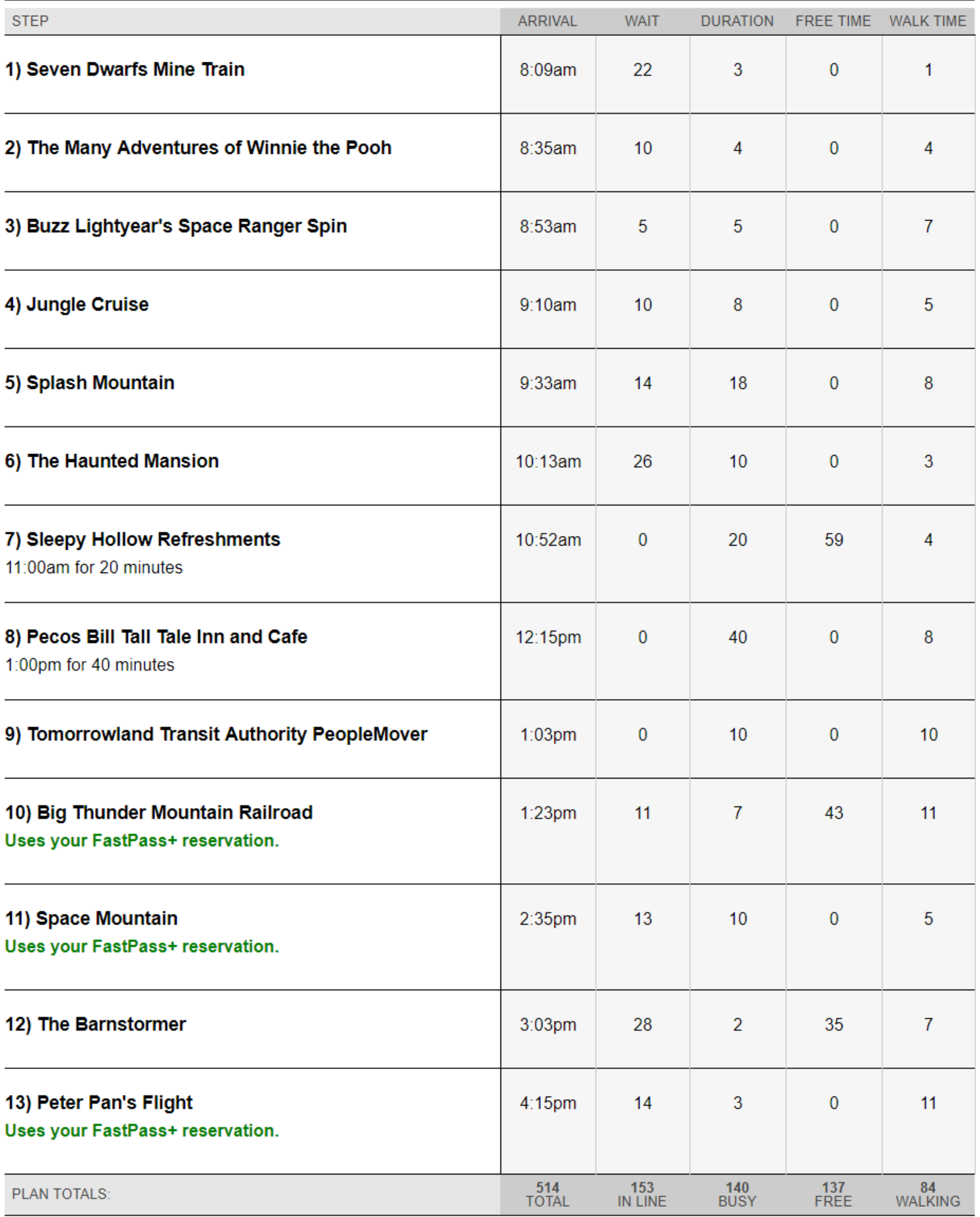
Due to COVID-19, all Disney parks are closed ☹

Challenge: How to automatically detect and exclude data from this scenario during data analytics?

Value of this Big Data

- Using big data, Touringplans can predict wait times of each ride in the future.
- Visitors can use Touringplans' service to optimize wait times at Disney World
- This service costs $16 per year. *Pay once for your group vs pay $20 per person for FASTPASS*

https://touringplans.com/blog/2018/06/25/disney-world-wait-times-available-for-data-science-and-machine-learning/

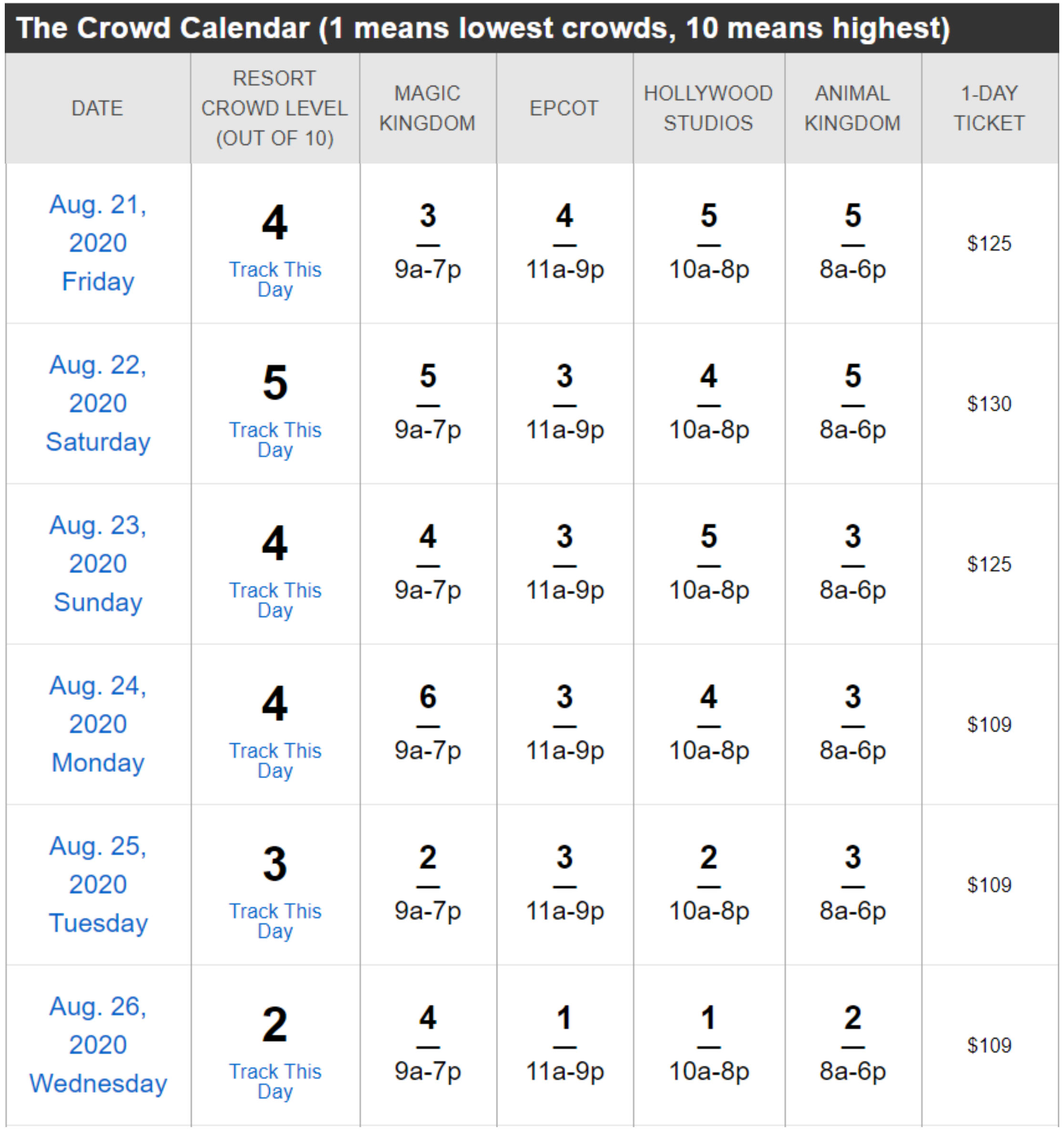


STEP	ARRIVAL	WAIT	DURATION	FREE TIME	WALK TIME
1) Seven Dwarfs Mine Train	8:09am	22	3	0	1
2) The Many Adventures of Winnie the Pooh	8:35am	10	4	0	4
3) Buzz Lightyear's Space Ranger Spin	8:53am	5	5	0	7
4) Jungle Cruise	9:10am	10	8	0	5
5) Splash Mountain	9:33am	14	18	0	8
6) The Haunted Mansion	10:13am	26	10	0	3
7) Sleepy Hollow Refreshments 11:00am for 20 minutes	10:52am	0	20	59	4
8) Pecos Bill Tall Tale Inn and Cafe 1:00pm for 40 minutes	12:15pm	0	40	0	8
9) Tomorrowland Transit Authority PeopleMover	1:03pm	0	10	0	10
10) Big Thunder Mountain Railroad **Uses your FastPass+ reservation.**	1:23pm	11	7	43	11
11) Space Mountain **Uses your FastPass+ reservation.**	2:35pm	13	10	0	5
12) The Barnstormer	3:03pm	28	2	35	7
13) Peter Pan's Flight **Uses your FastPass+ reservation.**	4:15pm	14	3	0	11
PLAN TOTALS:	514 TOTAL	153 IN LINE	140 BUSY	137 FREE	84 WALKING

Value of this Big Data

- Use the same Big Data to determine crowd size and ticket price

- Visitors can plan to buy cheap tickets and avoid the crowd



The Crowd Calendar (1 means lowest crowds, 10 means highest)

DATE	RESORT CROWD LEVEL (OUT OF 10)	MAGIC KINGDOM	EPCOT	HOLLYWOOD STUDIOS	ANIMAL KINGDOM	1-DAY TICKET
Aug. 21, 2020 Friday	4 Track This Day	3 9a-7p	4 11a-9p	5 10a-8p	5 8a-6p	$125
Aug. 22, 2020 Saturday	5 Track This Day	5 9a-7p	3 11a-9p	4 10a-8p	5 8a-6p	$130
Aug. 23, 2020 Sunday	4 Track This Day	4 9a-7p	3 11a-9p	5 10a-8p	3 8a-6p	$125
Aug. 24, 2020 Monday	4 Track This Day	6 9a-7p	3 11a-9p	4 10a-8p	3 8a-6p	$109
Aug. 25, 2020 Tuesday	3 Track This Day	2 9a-7p	3 11a-9p	2 10a-8p	3 8a-6p	$109
Aug. 26, 2020 Wednesday	2 Track This Day	4 9a-7p	1 11a-9p	1 10a-8p	2 8a-6p	$109

Summary

Big Data = data huge in size + growing rapidly in time.

Examples of Big Data: stock exchanges, social media sites, etc.

5V: Volume, Variety, Velocity, Veracity, and Value

Case study of Big Data

Q/A

Group assignment

Do the research and give a case study that benefits from using Big Data

Present in class next week

Suggested Topic:

- What is the problem that this case study is trying to solve?
- Describe how this case study uses Big Data to solve the problem in details
- Explain 5V's (Volume, Variety, Velocity, Veracity, Value) in this case study's Big Data
- Explain the benefits from using Big Data in this case study

Next Week's Group Assignment

Do the research and find a website or an application that uses Machine Learning (or Artificial Intelligence)

Suggested Topics:

- Explain the details of this website/application
- Explain how Machine Learning is used in that application/website (e.g., what kind of information it is trying to predict? How does it get input data to train machine learning model?)
- Explain how the website/application can gain benefits from using Machine Learning (e.g., increase in revenue)

Present in class next week (5-10 minute)