SESSION 2016-2017

B.TECH (CSE) YEAR: IV SEMESTER: VIII

BIG DATA AND ANALYTICS (CSE448)

MODULE 1 (L2)

Presented By

Dilip Kumar Sharma , Rahul Pradhan, Vivek Kumar, Yogesh Gupta

Dept of Computer Engineering & Applications

GLA University India

Quote

“Data is widely available. What is scarce is the ability to extract wisdom from it.”

Hal Varian, Google’s Chief Economist, 2010

Objectives and Outcomes

Learning Objectives	Learning Outcomes
Introduction to big data 1. Definition of big data. 2. Challenges of big data. 3. Why big data? 4. Traditional Business Intelligence versus big data.	a) To understand the significance of big data. b) To understand the other characteristics of data that are not definitional characteristics of big data. c) To understand the challenges of big data and how to deal with the same. d) To understand what is new today.

Agenda

- Definition of Big Data
  - Volume
  - Velocity
  - Variety
- Challenges of Big Data
- Other Characteristics of Data Which are Not Definitional Traits of Big Data
- Why Big Data?
- Traditional Business Intelligence (BI) versus Big Data
  - A Typical Data Warehouse Environment
  - A Typical Hadoop Environment
  - Coexistence of Big Data and Data Warehouse

CHARACTERISTICS OF DATA

□ **Composition:** The composition of data deals with the structure of data, that is, the sources of data the granularity, the types, and the nature of data as to whether it is static or real-time streaming.

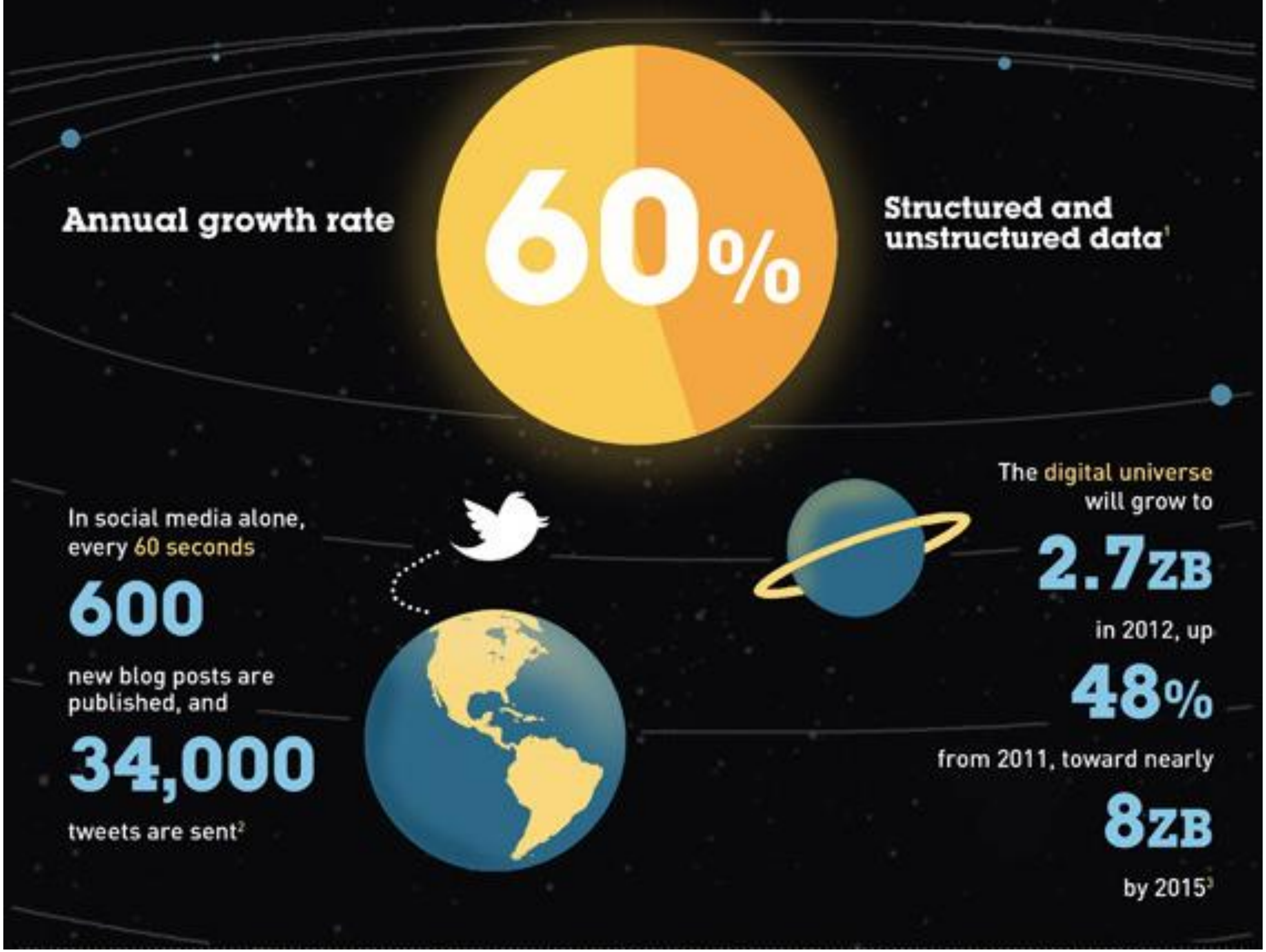
□ **Condition:** The condition of data deals with the state of data, that is, “Can one use this data as is for analysis?” or “Does it require cleansing for further enhancement and enrichment?”

□ **Context:** The context of data deals with “Where has this data been generated?” “Why was this data generated?” “How sensitive is this data?” “What are the events associated with this data?” and so on.

EVOLUTION OF BIG DATA

- 1970s and before was the era of mainframes. The data was essentially primitive and structured. Relational databases evolved in 1980s and 1990s.
- The era was of data intensive applications. The World Wide Web (WWW) and the Internet of Things (IoT) have led to an onslaught of structured, unstructured, and multimedia data.

Big Data is growing fast



Annual growth rate **60%** Structured and unstructured data[1]

In social media alone, every 60 seconds

600

new blog posts are published, and

34,000

tweets are sent[2]

The digital universe will grow to

2.7ZB

in 2012, up

48%

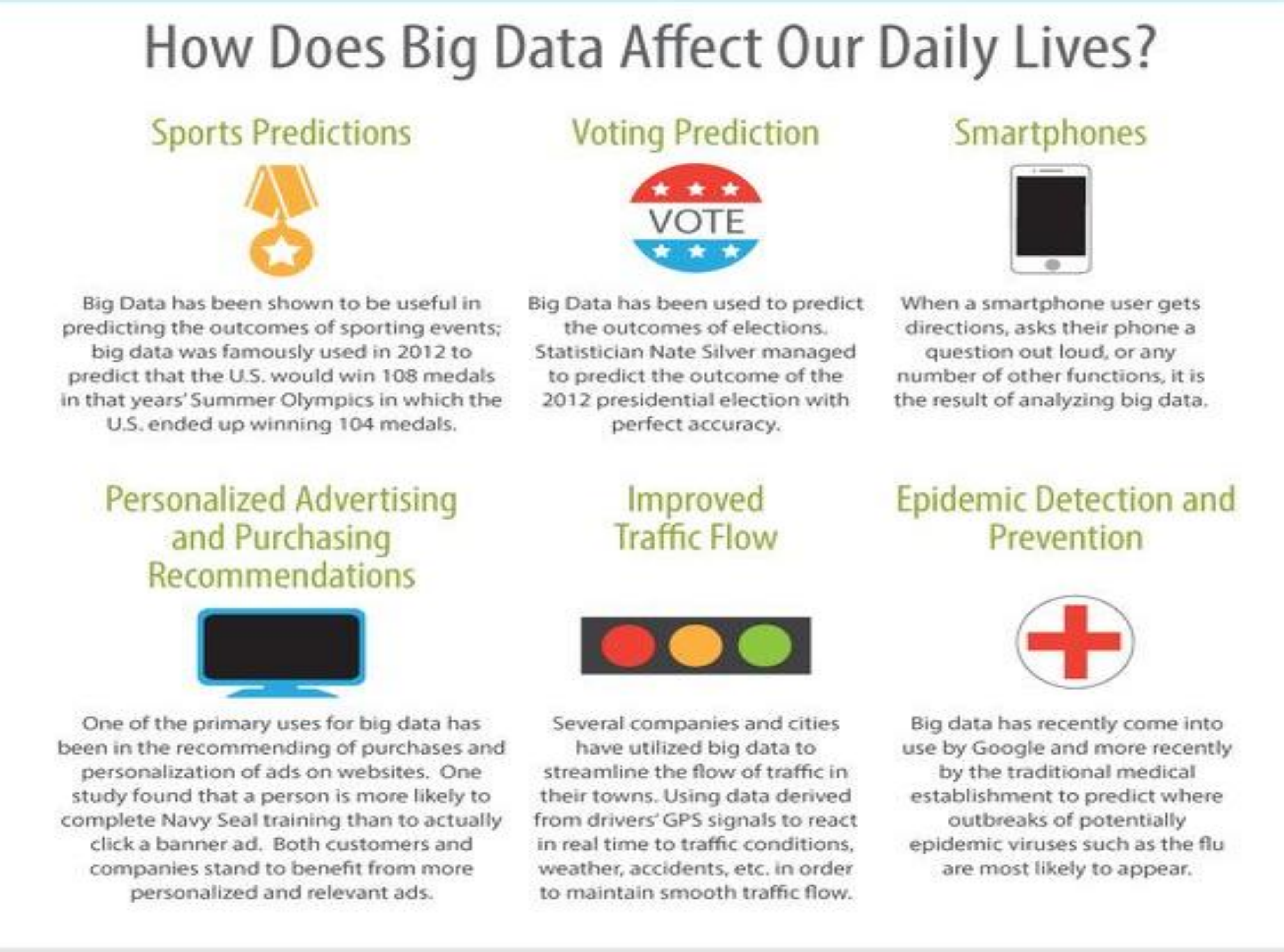
from 2011, toward nearly

8ZB

by 2015[3]

EVOLUTION OF BIG DATA

	Data Generation and Storage	**Data Utilization**	**Data Driven**
Complex and Unstructured			Structured data, Unstructured data, Multimedia data
Complex and Relational		Relational databases: Data-intensive applications	
Primitive and Structured	Mainframes: Basic data storage		
Existance	**1970s and before**	**Relational(1980s and 1990s)**	**2000s and beyond**



How Does Big Data Affect Our Daily Lives?

Sports Predictions

Big Data has been shown to be useful in predicting the outcomes of sporting events; big data was famously used in 2012 to predict that the U.S. would win 108 medals in that years' Summer Olympics in which the U.S. ended up winning 104 medals.

Voting Prediction

Big Data has been used to predict the outcomes of elections. Statistician Nate Silver managed to predict the outcome of the 2012 presidential election with perfect accuracy.

Smartphones

When a smartphone user gets directions, asks their phone a question out loud, or any number of other functions, it is the result of analyzing big data.

Personalized Advertising and Purchasing Recommendations

One of the primary uses for big data has been in the recommending of purchases and personalization of ads on websites. One study found that a person is more likely to complete Navy Seal training than to actually click a banner ad. Both customers and companies stand to benefit from more personalized and relevant ads.

Improved Traffic Flow

Several companies and cities have utilized big data to streamline the flow of traffic in their towns. Using data derived from drivers' GPS signals to react in real time to traffic conditions, weather, accidents, etc. in order to maintain smooth traffic flow.

Epidemic Detection and Prevention

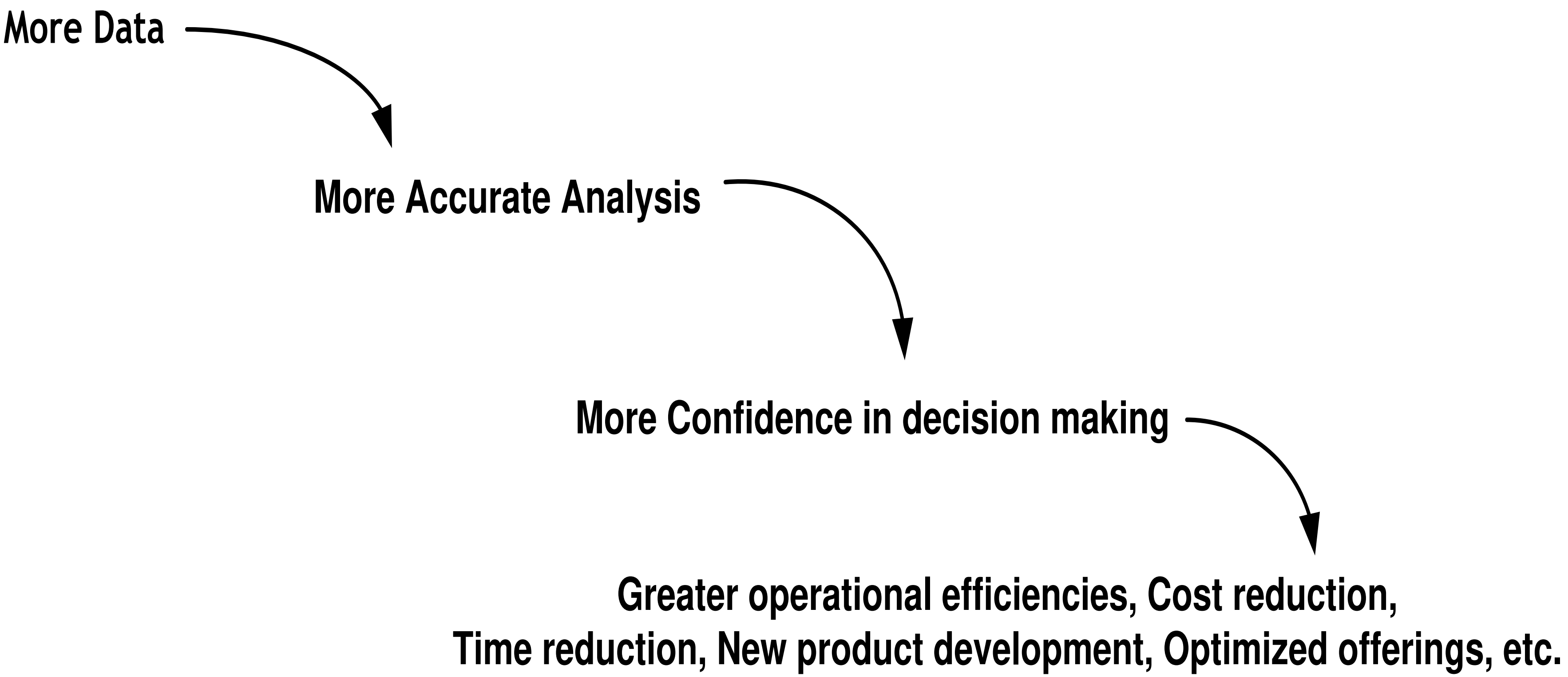
Big data has recently come into use by Google and more recently by the traditional medical establishment to predict where outbreaks of potentially epidemic viruses such as the flu are most likely to appear.

Why Big Data?

- The more data we have for analysis, the greater will be the analytical accuracy and also the greater would be theconfidence in our decisions based on these analytical findings.

  More data —» More accurate analysis —» Greater confidence in decision making —» Greater operational efficiencies, cost reduction, time reduction, new product development, and optimized offerings, etc.

Why Big Data?

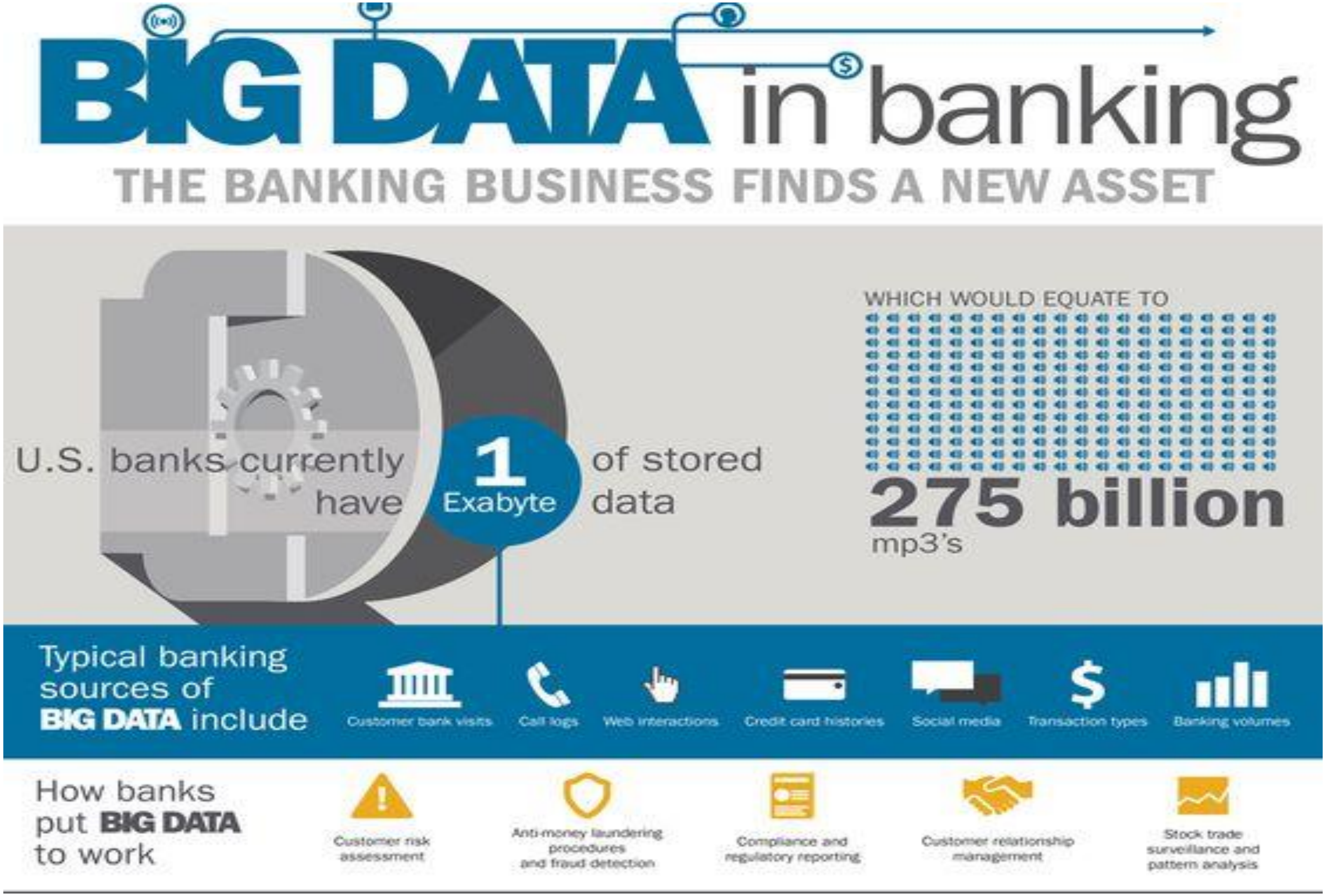


More Data

More Accurate Analysis

More Confidence in decision making

Greater operational efficiencies, Cost reduction,
Time reduction, New product development, Optimized offerings, etc.



BIG DATA in banking

THE BANKING BUSINESS FINDS A NEW ASSET

U.S. banks currently have 1 Exabyte of stored data

WHICH WOULD EQUATE TO 275 billion mp3's

Typical banking sources of BIG DATA include

- Customer bank visits
- Call logs
- Web interactions
- Credit card histories
- Social media
- Transaction types
- Banking volumes

How banks put BIG DATA to work

- Customer risk assessment
- Anti-money laundering procedures and fraud detection
- Compliance and regulatory reporting
- Customer relationship management
- Stock trade surveillance and pattern analysis

BIG DATA problem solving for financial institutions

Preventing Customer Churn

In 2012

50%

Of coustomers changed banks or were planning to change banks

SOLUTION

A customer churn prediction model based on big data analysis of social media customer sentiment and purchasing power helps identify customers at risk of leaving

Setting Effective Staffing Levels

Staffing costs account for

66%

of a branch bank's costs

SOLUTION

Staffing models based on transaction times and account holder traffic patterns data make annual resource planning easier and more effective

Understanding Customer Needs

Customers face a bewildering number of banking "journeys" through websites, call centers, branch bank personnel and more

SOLUTION

Big data from customer waiting and assist periods and online banking habits helps track the paths customers follow and how they affect purchasing decisions

Managing Rising Security Costs

More stringent federal security regulations require banks to have safeguards for anti-money laundering in place

SOLUTION

Big data from cross channel security alerts and international money transactions creates compliance analytics for performance models that reduce alerts while delivering ongoing monitoring

Insights for Product Development

The costs of developing new products and services can be staggering

SOLUTION

Customer transactional data such as timing of visits and duration of teller transactions can be analyzed to find gaps in product offerings

Scoring Credit Risks

Banks need to cut lending risks while improving customer marketing

SOLUTION

Consumer payment patterns and law enforcement databases supply the data

credit bureaus store over

800 billion records

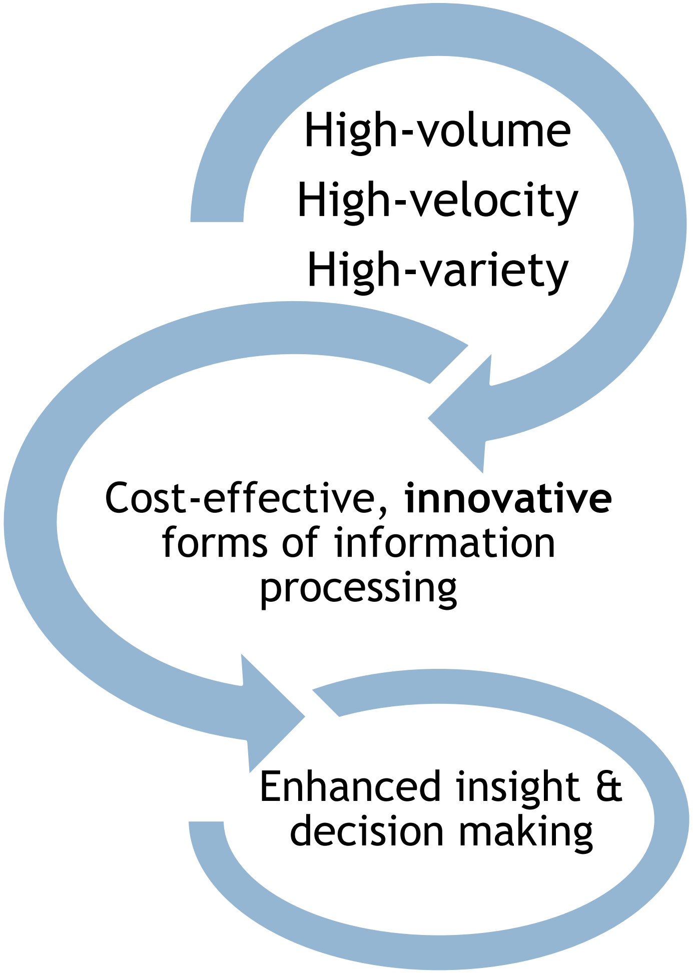
to be sliced, diced and analyzed for more accurate credit risk scores.

For comparison, the FBI's Investigative Data Warehouse has only 1.5 billion documents.

Definition of Big Data

Big Data is high-volume, high-velocity, and high-variety information assets that demand cost effective, innovative forms of information processing for enhanced insight and decision making.

Source: Gartner IT Glossary


High-volume
High-velocity
High-variety
Cost-effective, **innovative** forms of information processing
Enhanced insight & decision making

Other Definitions of Big Data

- "Big data is high-volume, high-velocity, and high-variety information assets" talks about voluminous data (humongous data) that may have great variety (a good mix of structured, semi-structured, and unstructured data) and will require a good speed/pace for storage, preparation, pro-cessing, and analysis.
- "Cost effective, innovative forms of information processing" talks about embracing new techniques and technologies to capture (ingest), store, process, persist, integrate, and visualize the high-volume, high-velocity, and high-variety data.

Other Definitions of Big Data cont…

- “Enhanced insight and decision making” talks about deriving deeper, richer, and meaningful insights and then using these insights to make faster and better decisions to gain business value and thus a competitive edge.

Data —» Information —» Actionable intelligence —» Better decisions —» Enhanced business value

Volume - A Mountain of Data

1 Kilobyte (KB) = 1000 bytes

1 Megabyte (MB) = 1,000,000 bytes

1 Gigabyte (GB) = 1,000,000,000 bytes

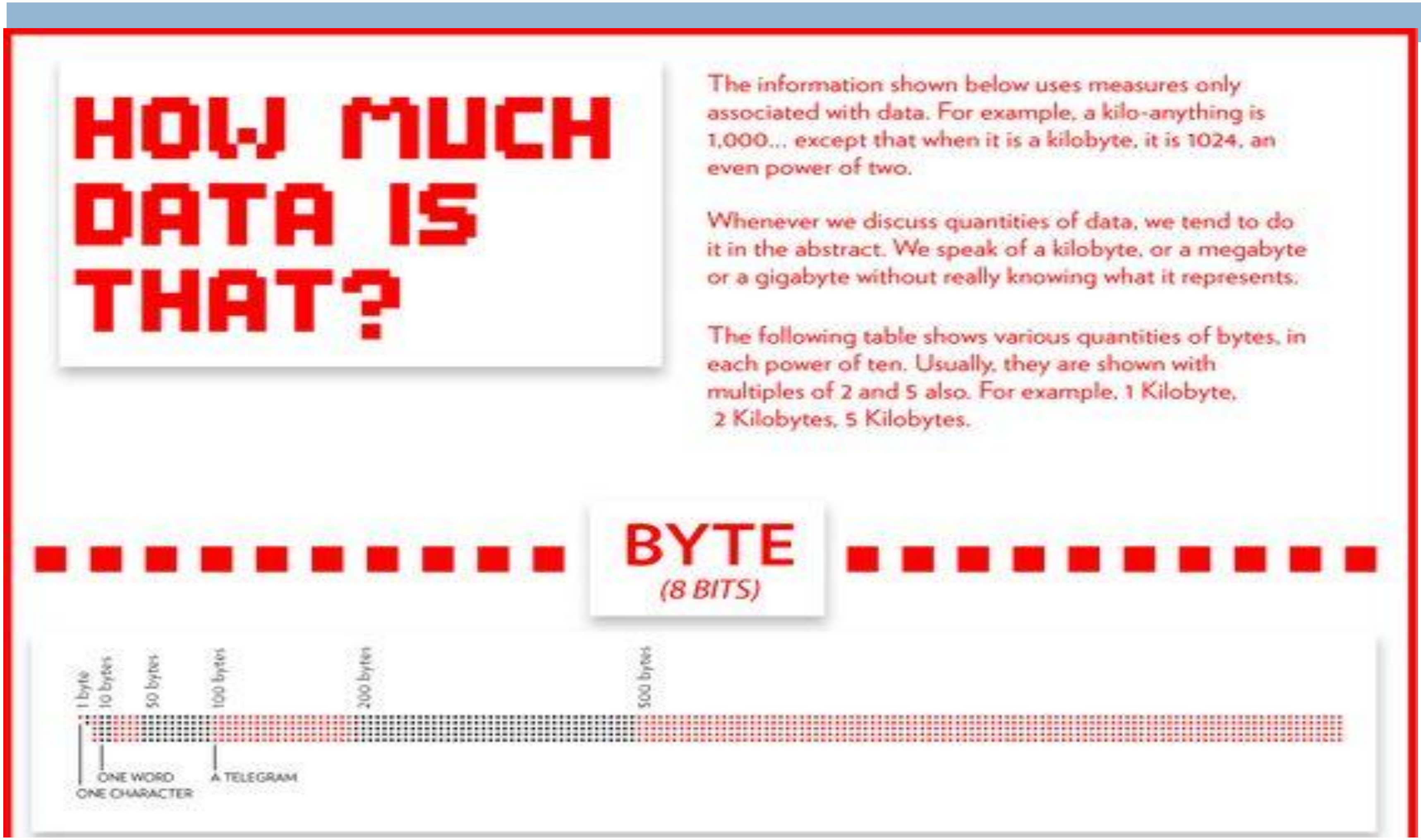
1 Terabyte (TB) = 1,000,000,000,000 bytes

1 Petabyte (PB) = 1,000,000,000,000,000 bytes

1 Exabyte (EB) = 1,000,000,000,000,000,000 bytes

1 Zettabyte (ZB) = 1,000,000,000,000,000,000,000 bytes

1 Yottabyte (YB) = 1,000,000,000,000,000,000,000,000 bytes



HOW MUCH DATA IS THAT?

The information shown below uses measures only associated with data. For example, a kilo-anything is 1,000... except that when it is a kilobyte, it is 1024, an even power of two.

Whenever we discuss quantities of data, we tend to do it in the abstract. We speak of a kilobyte, or a megabyte or a gigabyte without really knowing what it represents.

The following table shows various quantities of bytes, in each power of ten. Usually, they are shown with multiples of 2 and 5 also. For example, 1 Kilobyte, 2 Kilobytes, 5 Kilobytes.

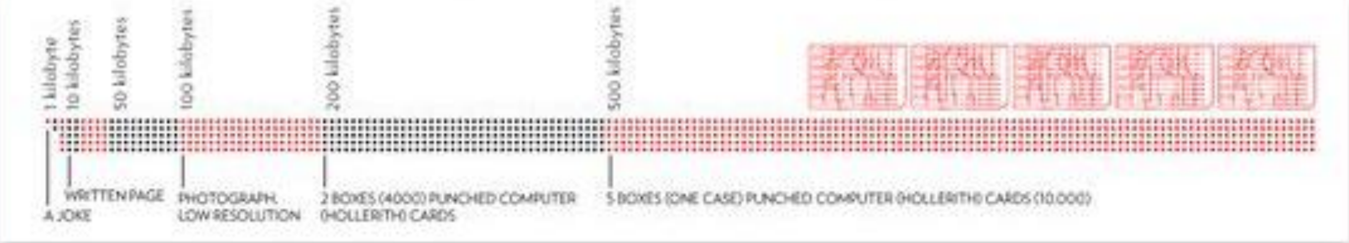
BYTE

(8 BITS)

KILOBYTE

(1,024 BYTES; 2^{10})

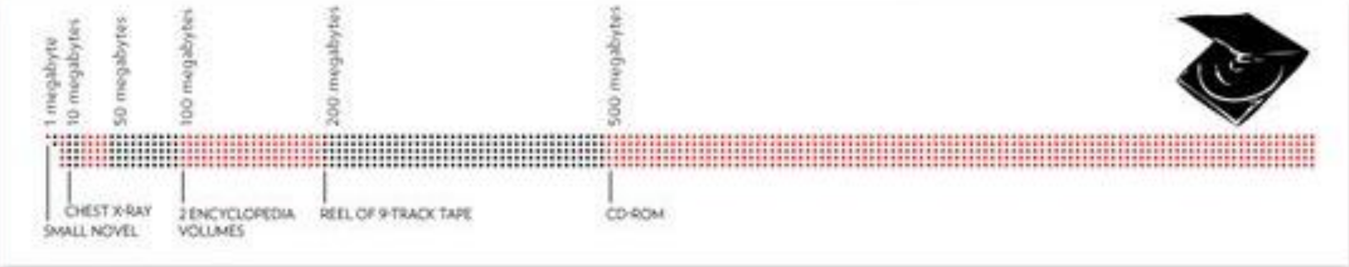
approx. 1,000 or 10^3


1 kilobyte
10 kilobytes
50 kilobytes
100 kilobytes
200 kilobytes
500 kilobytes
A JOKE
WRITTEN PAGE
PHOTOGRAPH, LOW RESOLUTION
2 BOXES (4000) PUNCHED COMPUTER (HOLLERITH) CARDS
5 BOXES (ONE CASE) PUNCHED COMPUTER (HOLLERITH) CARDS (10,000)

MEGABYTE

(1,048,576 BYTES; 2^{20})

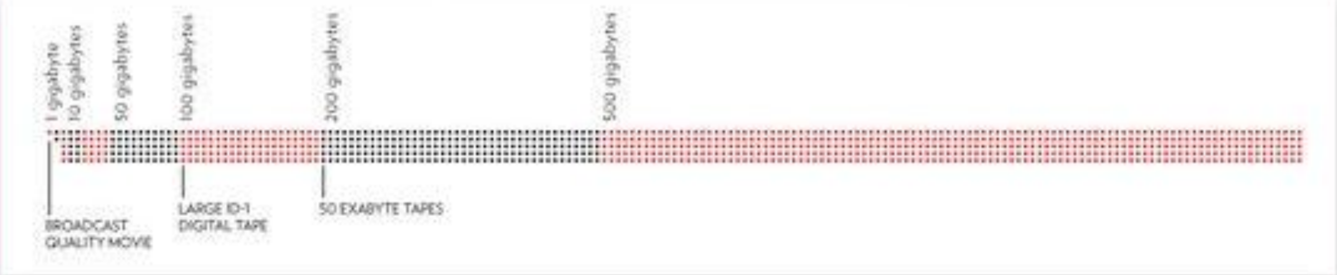
approx. 1,000,000 or 10^6



GIGABYTE

(1,073,741,824 BYTES; 2^{30})

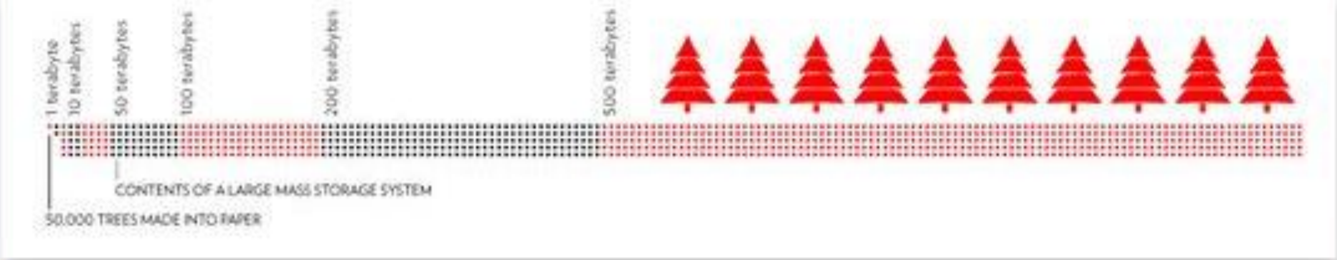
approx. 1,000,000,000 or 10^{9}



TERABYTE

(1,099,511,627,776 BYTES; 2^{40})

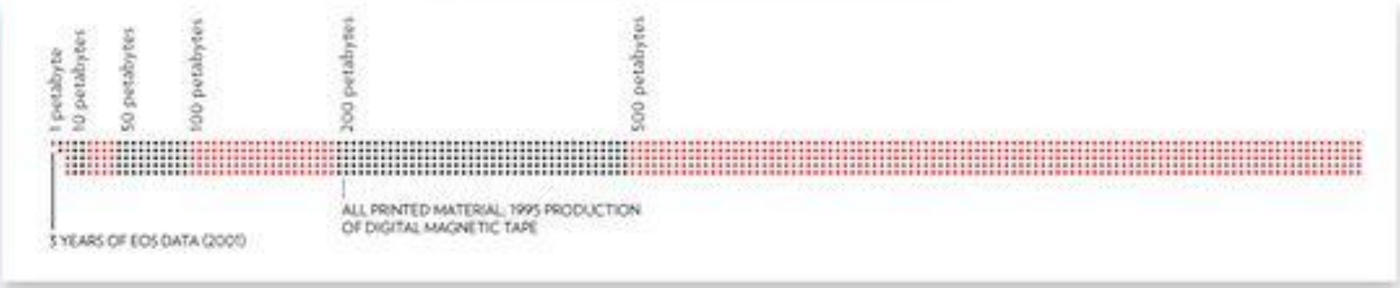
approx. 1,000,000,000,000 or 10^{12}

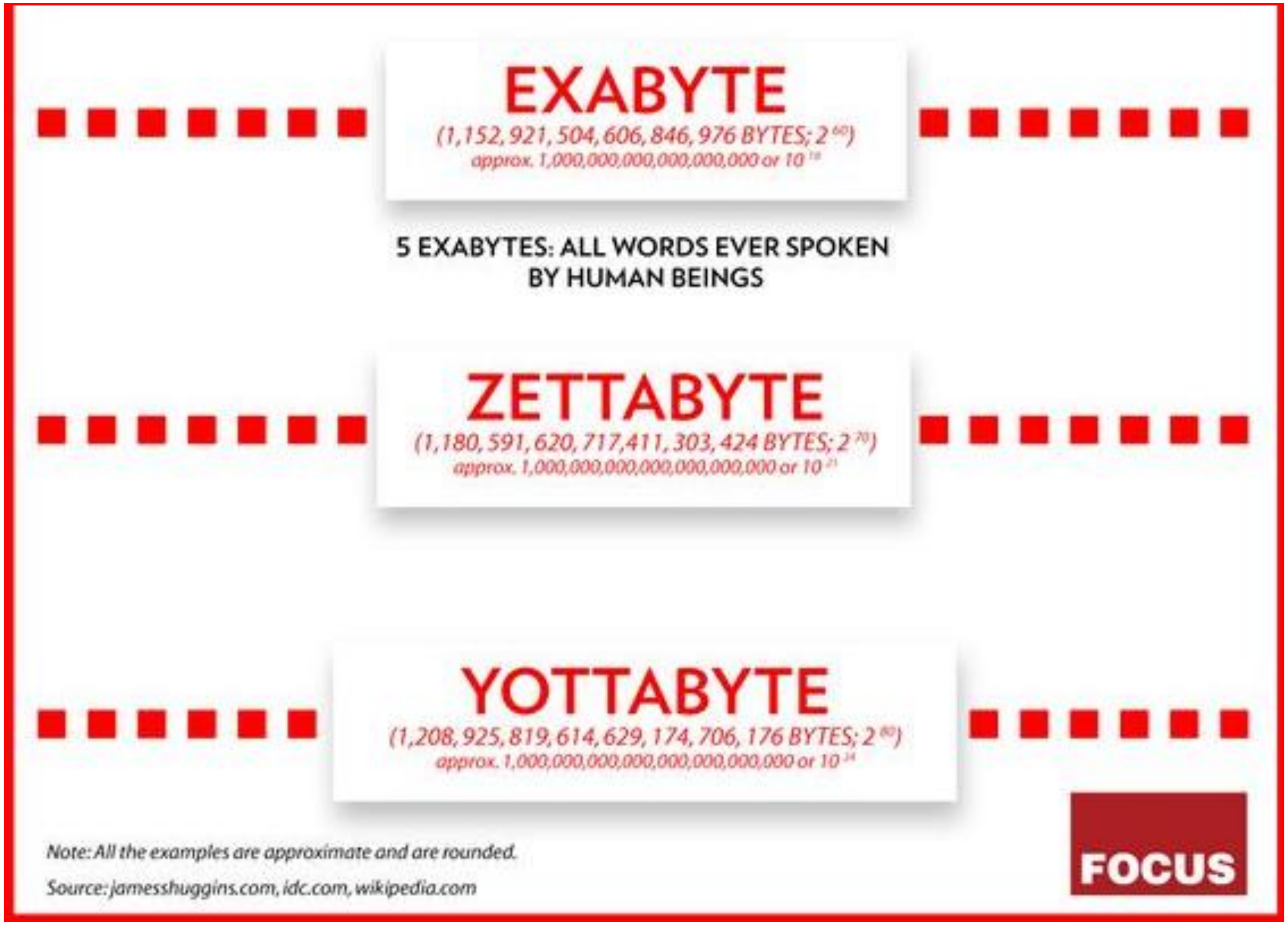


PETABYTE

(1,125,899,906,842,624 BYTES; 2^{50})

approx. 1,000,000,000,000,000 or 10^{15}





EXABYTE

(1,152, 921, 504, 606, 846, 976 BYTES; 2^{60})
approx. 1,000,000,000,000,000,000 or 10^{18}

5 EXABYTES: ALL WORDS EVER SPOKEN
BY HUMAN BEINGS

ZETTABYTE

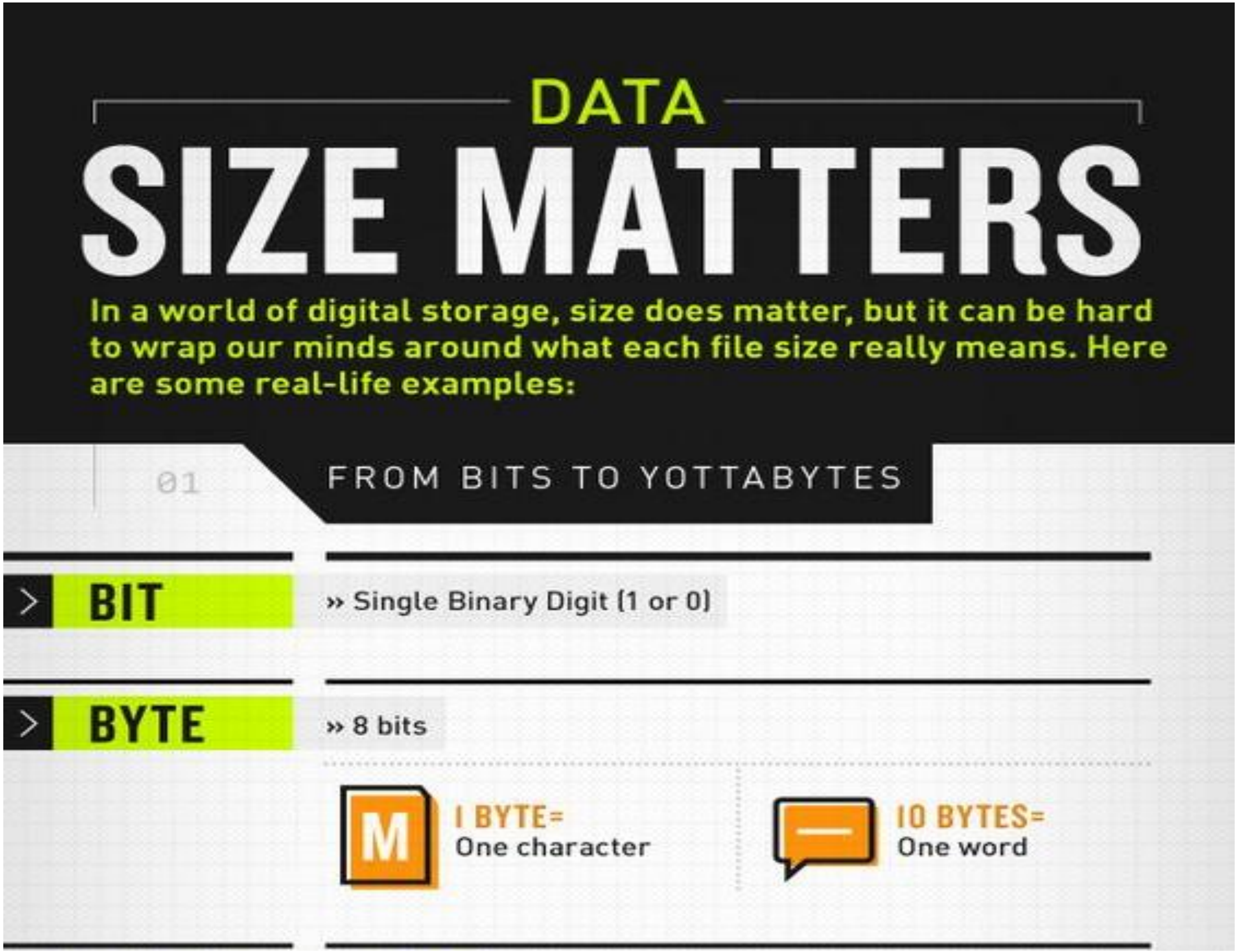
(1,180, 591, 620, 717,411, 303, 424 BYTES; 2^{70})
approx. 1,000,000,000,000,000,000,000 or 10^{21}

YOTTABYTE

(1,208, 925, 819, 614, 629, 174, 706, 176 BYTES; 2^{80})
approx. 1,000,000,000,000,000,000,000,000 or 10^{24}

Note: All the examples are approximate and are rounded.

Source: jamesshuggins.com, idc.com, wikipedia.com



DATA

SIZE MATTERS

In a world of digital storage, size does matter, but it can be hard to wrap our minds around what each file size really means. Here are some real-life examples:

01

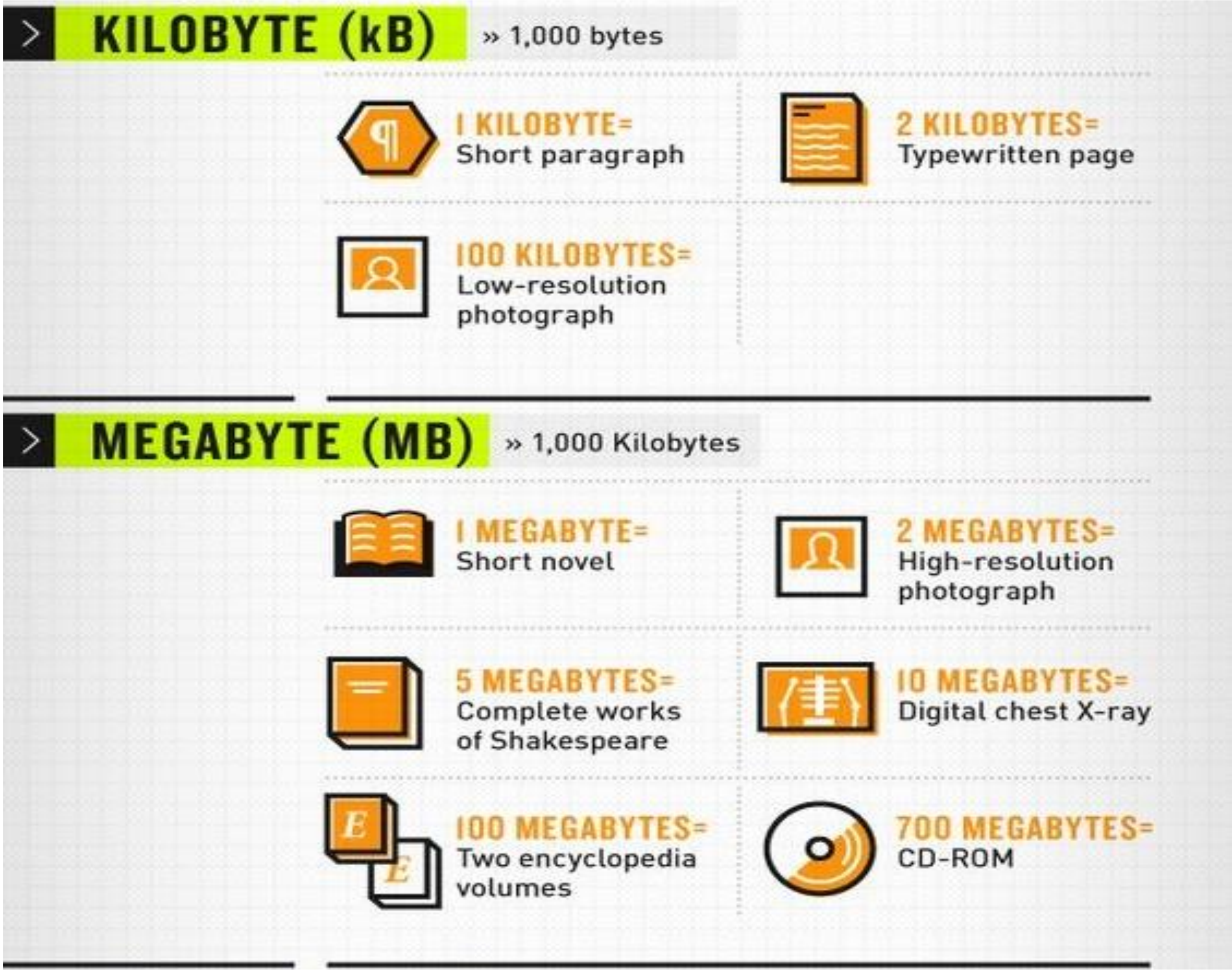
FROM BITS TO YOTTABYTES

BIT » Single Binary Digit (1 or 0)

BYTE » 8 bits

1 BYTE= One character

10 BYTES= One word



KILOBYTE (kB)

» 1,000 bytes

I KILOBYTE=
Short paragraph

2 KILOBYTES=
Typewritten page

100 KILOBYTES=
Low-resolution photograph

MEGABYTE (MB)

» 1,000 Kilobytes

I MEGABYTE=
Short novel

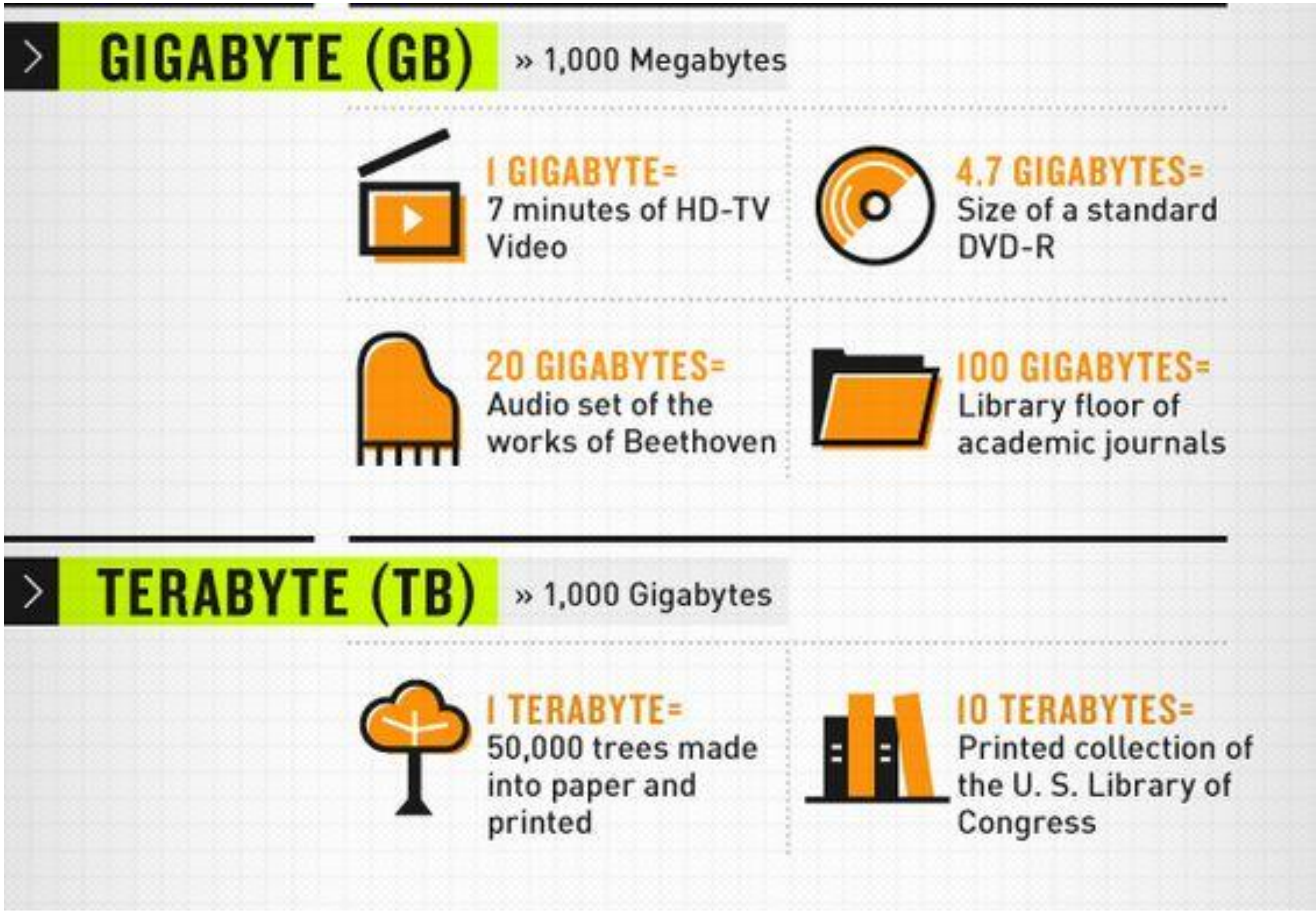
2 MEGABYTES=
High-resolution photograph

5 MEGABYTES=
Complete works of Shakespeare

10 MEGABYTES=
Digital chest X-ray

100 MEGABYTES=
Two encyclopedia volumes

700 MEGABYTES=
CD-ROM



GIGABYTE (GB)

» 1,000 Megabytes

1 GIGABYTE=
7 minutes of HD-TV Video

4.7 GIGABYTES=
Size of a standard DVD-R

20 GIGABYTES=
Audio set of the works of Beethoven

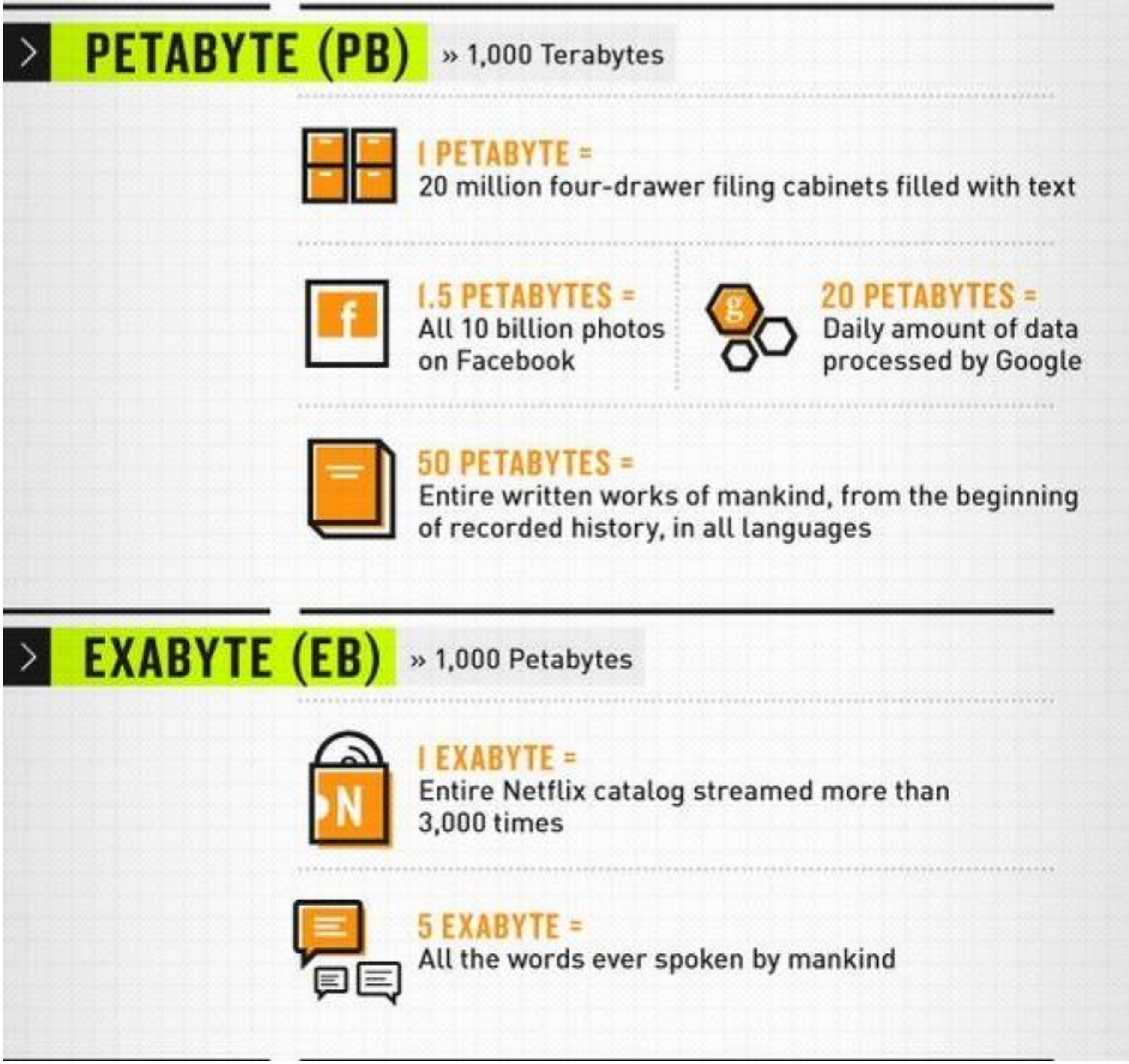
100 GIGABYTES=
Library floor of academic journals

TERABYTE (TB)

» 1,000 Gigabytes

1 TERABYTE=
50,000 trees made into paper and printed

10 TERABYTES=
Printed collection of the U. S. Library of Congress



PETABYTE (PB)

» 1,000 Terabytes

1 PETABYTE =
20 million four-drawer filing cabinets filled with text

1.5 PETABYTES =
All 10 billion photos on Facebook

20 PETABYTES =
Daily amount of data processed by Google

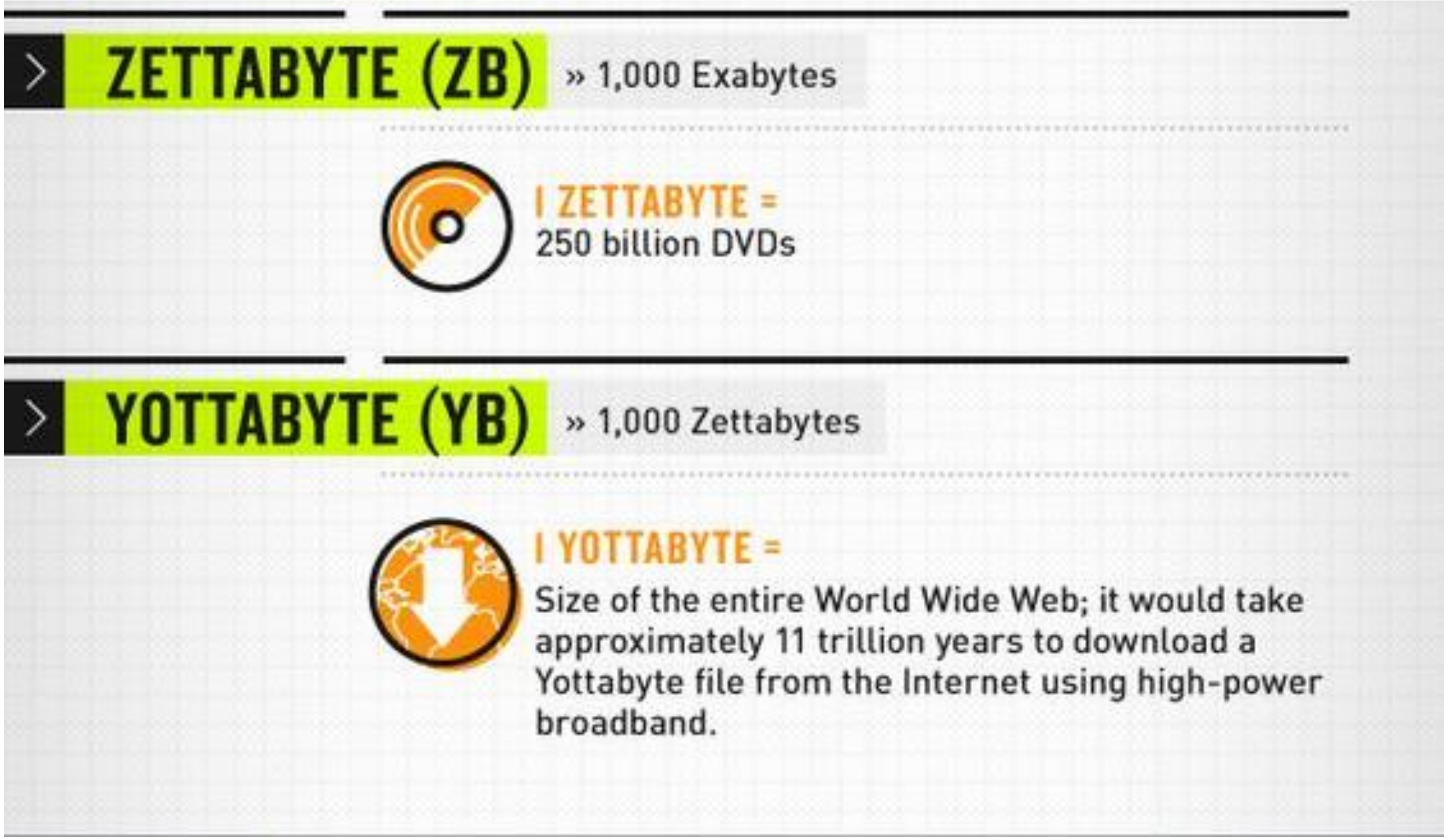
50 PETABYTES =
Entire written works of mankind, from the beginning of recorded history, in all languages

EXABYTE (EB)

» 1,000 Petabytes

1 EXABYTE =
Entire Netflix catalog streamed more than 3,000 times

5 EXABYTE =
All the words ever spoken by mankind



ZETTABYTE (ZB)

» 1,000 Exabytes

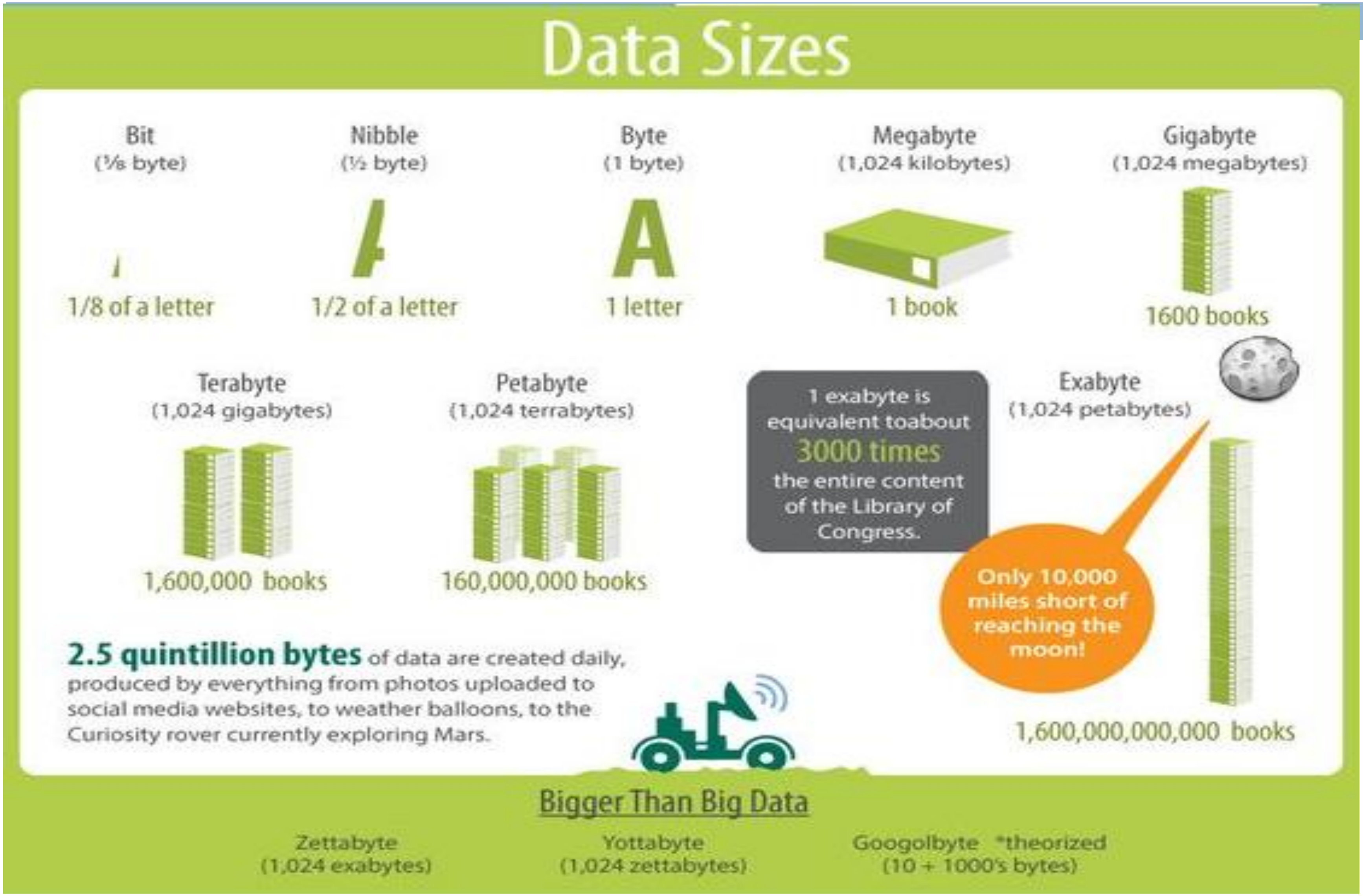
I ZETTABYTE =
250 billion DVDs

YOTTABYTE (YB)

» 1,000 Zettabytes

I YOTTABYTE =
Size of the entire World Wide Web; it would take approximately 11 trillion years to download a Yottabyte file from the Internet using high-power broadband.

Data Sizes in form of Books



Data Sizes

Bit
(⅛ byte)
1/8 of a letter

Nibble
(½ byte)
1/2 of a letter

Byte
(1 byte)
1 letter

Megabyte
(1,024 kilobytes)
1 book

Gigabyte
(1,024 megabytes)
1600 books

Terabyte
(1,024 gigabytes)
1,600,000 books

Petabyte
(1,024 terrabytes)
160,000,000 books

1 exabyte is equivalent toabout 3000 times the entire content of the Library of Congress.

Exabyte
(1,024 petabytes)
1,600,000,000,000 books

Only 10,000 miles short of reaching the moon!

2.5 quintillion bytes of data are created daily, produced by everything from photos uploaded to social media websites, to weather balloons, to the Curiosity rover currently exploring Mars.

Bigger Than Big Data

Zettabyte
(1,024 exabytes)

Yottabyte
(1,024 zettabytes)

Googolbyte *theorized
(10 + 1000's bytes)

Where Does This Data get Generated?

- There are a multitude of sources for big data. An XLS, a DOC, a PDF. etc. is unstructured data; a video on YouTube, a chat conversation on Internet Messenger, a customer feedback form on an online retail website is unstructured data; a CCTV coverage, a weather forecast report is unstructured data too.

1. Typical Internal Data Sources

- Data present within an organizations firewall. It is as follows:
  - *Data storage:* File systems, SQL (RDBMSs - Oracle, MS SQL Server, DB2, MySQL, PostgreSQL, etc.), NoSQL (MongoDB, Cassandra, etc.), and so on.
  - *Archives:* Archives of scanned documents, paper archives, customer correspondence records, patients' health records, students admission records, students' assessment records, and so on.

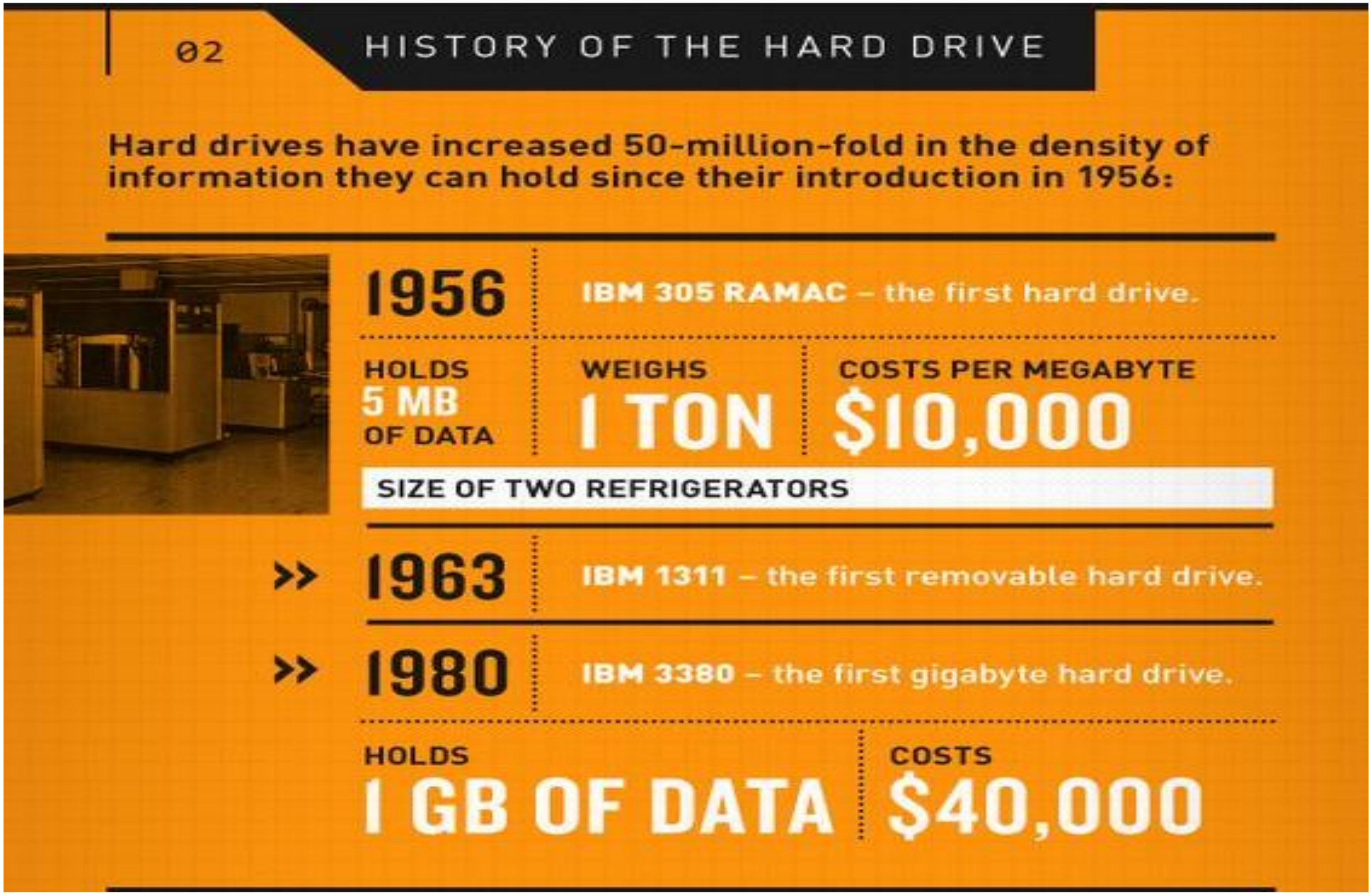
2. External Data Sources

- Data residing outside an organization’s firewall. It is as follows:
  - ***Public Web:*** Wikipedia, weather, regulatory, compliance, census, etc.

3. Both (Internal+External) Data Sources

- □ ***Sensor data:*** Car sensors, smart electric meters, office buildings, air conditioning units, refriget*- tors, and so on.
- □ ***Machine log data:*** Event logs, application logs, Business process logs, audit logs, clickstream data, etc.
- □ ***Social media:*** Twitter, blogs, Facebook, LinkedIn, YouTube, Instagram, etc.
- □ ***Business apps:*** ERP, CRM, HR, Google Docs, and so on.
- □ ***Media:*** Audio, Video, Image, Podcast, etc.
- □ ***Docs:*** Comma separated value (CSV), Word Documents, PDF, XLS, PPT, and so on.

Where all this data stores?



02

HISTORY OF THE HARD DRIVE

Hard drives have increased 50-million-fold in the density of information they can hold since their introduction in 1956:

1956 IBM 305 RAMAC – the first hard drive.

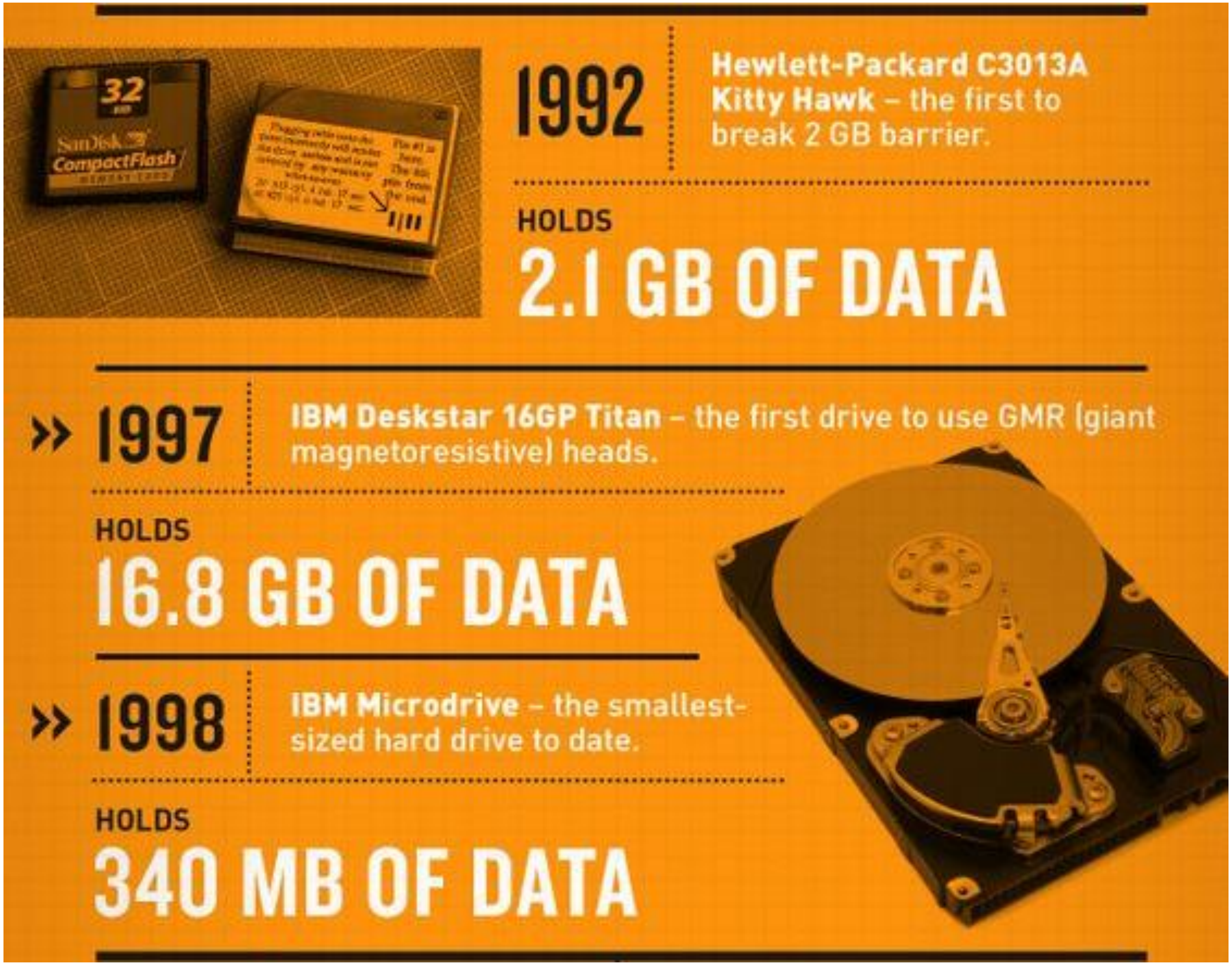
HOLDS	WEIGHS	COSTS PER MEGABYTE
5 MB OF DATA	1 TON	$10,000

SIZE OF TWO REFRIGERATORS

» **1963** IBM 1311 – the first removable hard drive.

» **1980** IBM 3380 – the first gigabyte hard drive.

HOLDS	COSTS
1 GB OF DATA	$40,000



1992

Hewlett-Packard C3013A Kitty Hawk – the first to break 2 GB barrier.

HOLDS

2.1 GB OF DATA

» 1997

IBM Deskstar 16GP Titan – the first drive to use GMR (giant magnetoresistive) heads.

HOLDS

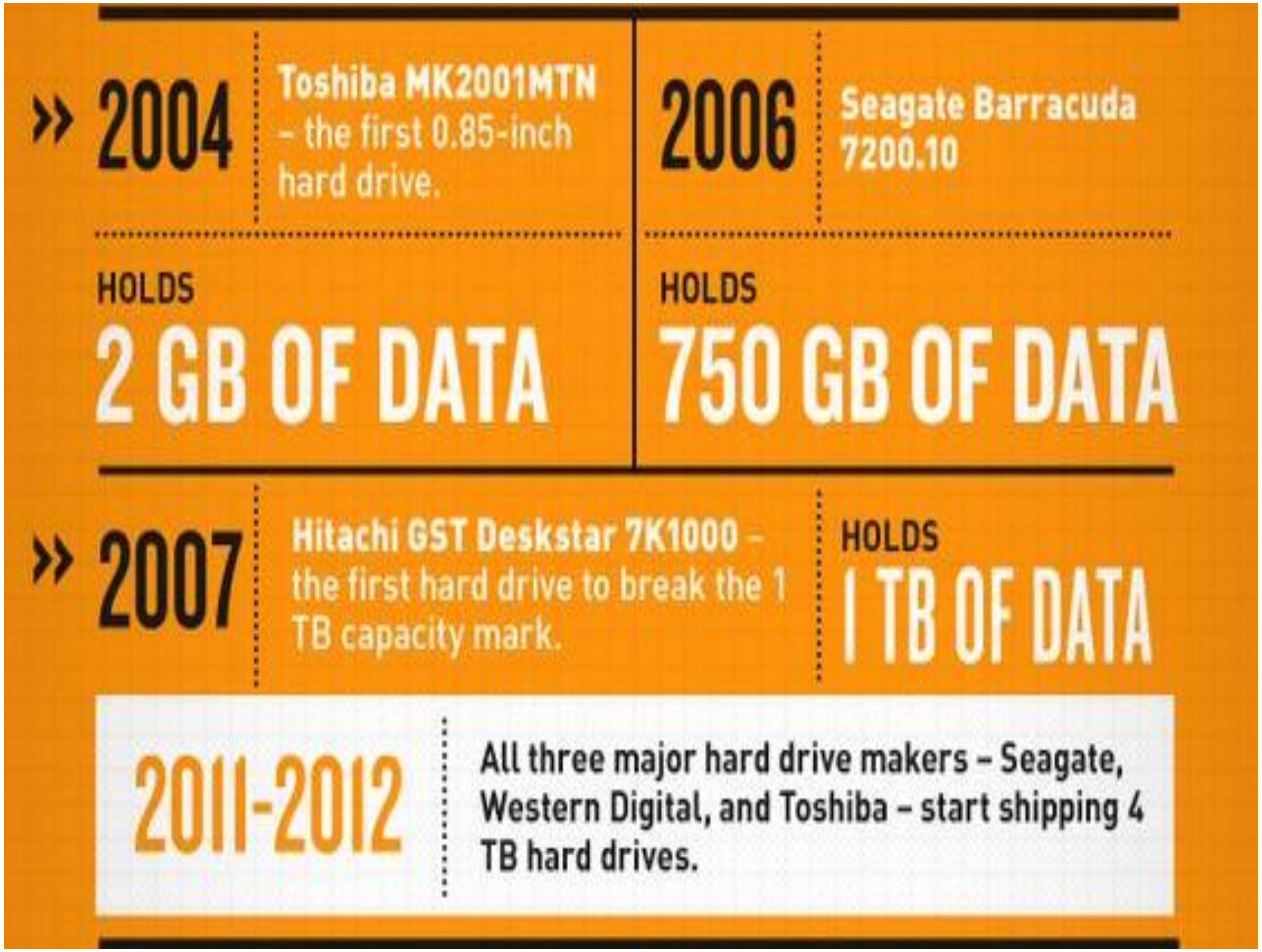
16.8 GB OF DATA

» 1998

IBM Microdrive – the smallest-sized hard drive to date.

HOLDS

340 MB OF DATA



» 2004

Toshiba MK2001MTN – the first 0.85-inch hard drive.

HOLDS

2 GB OF DATA

2006

Seagate Barracuda 7200.10

HOLDS

750 GB OF DATA

» 2007

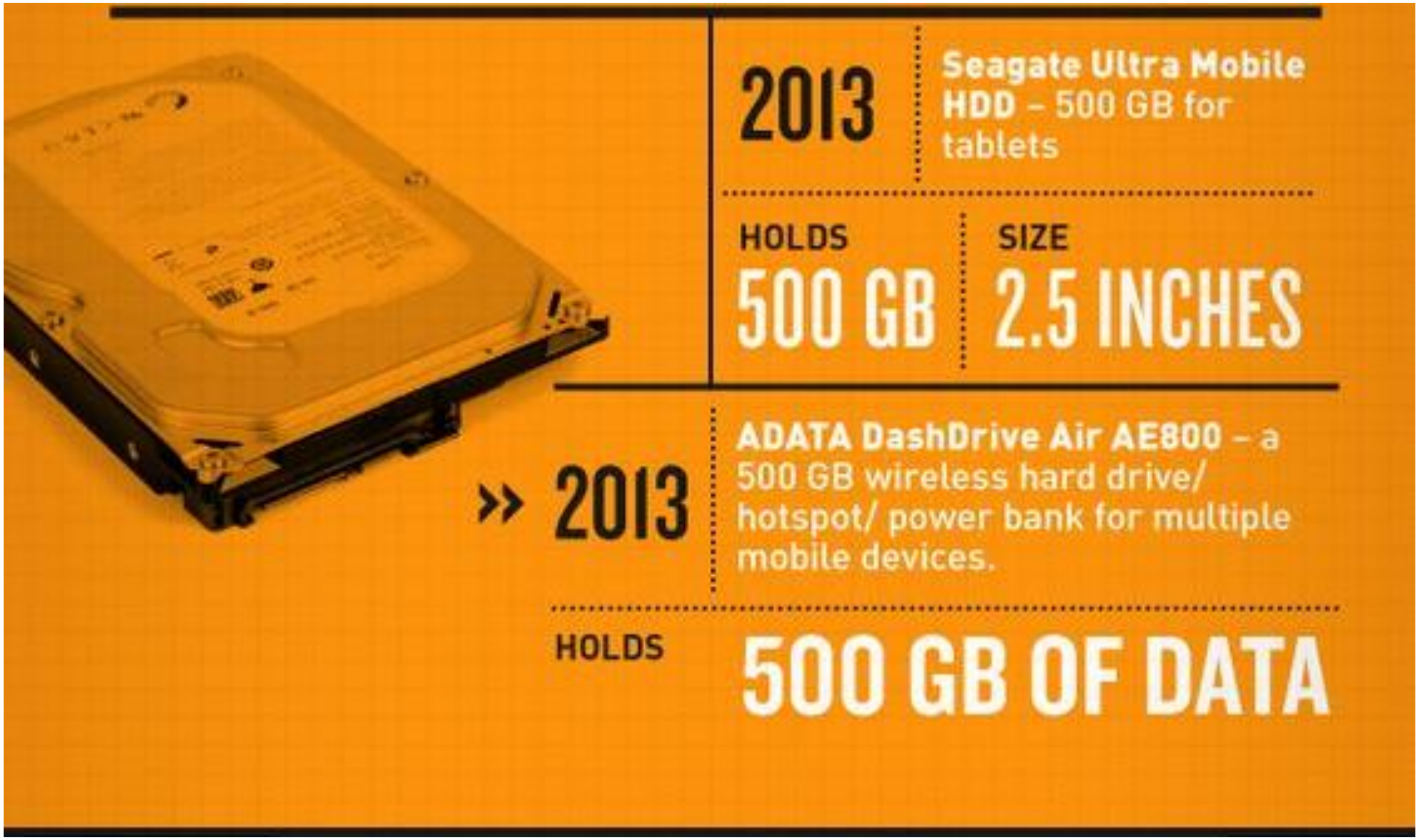
Hitachi GST Deskstar 7K1000 – the first hard drive to break the 1 TB capacity mark.

HOLDS

1 TB OF DATA

2011-2012

All three major hard drive makers – Seagate, Western Digital, and Toshiba – start shipping 4 TB hard drives.



2013

Seagate Ultra Mobile HDD – 500 GB for tablets

HOLDS

500 GB

SIZE

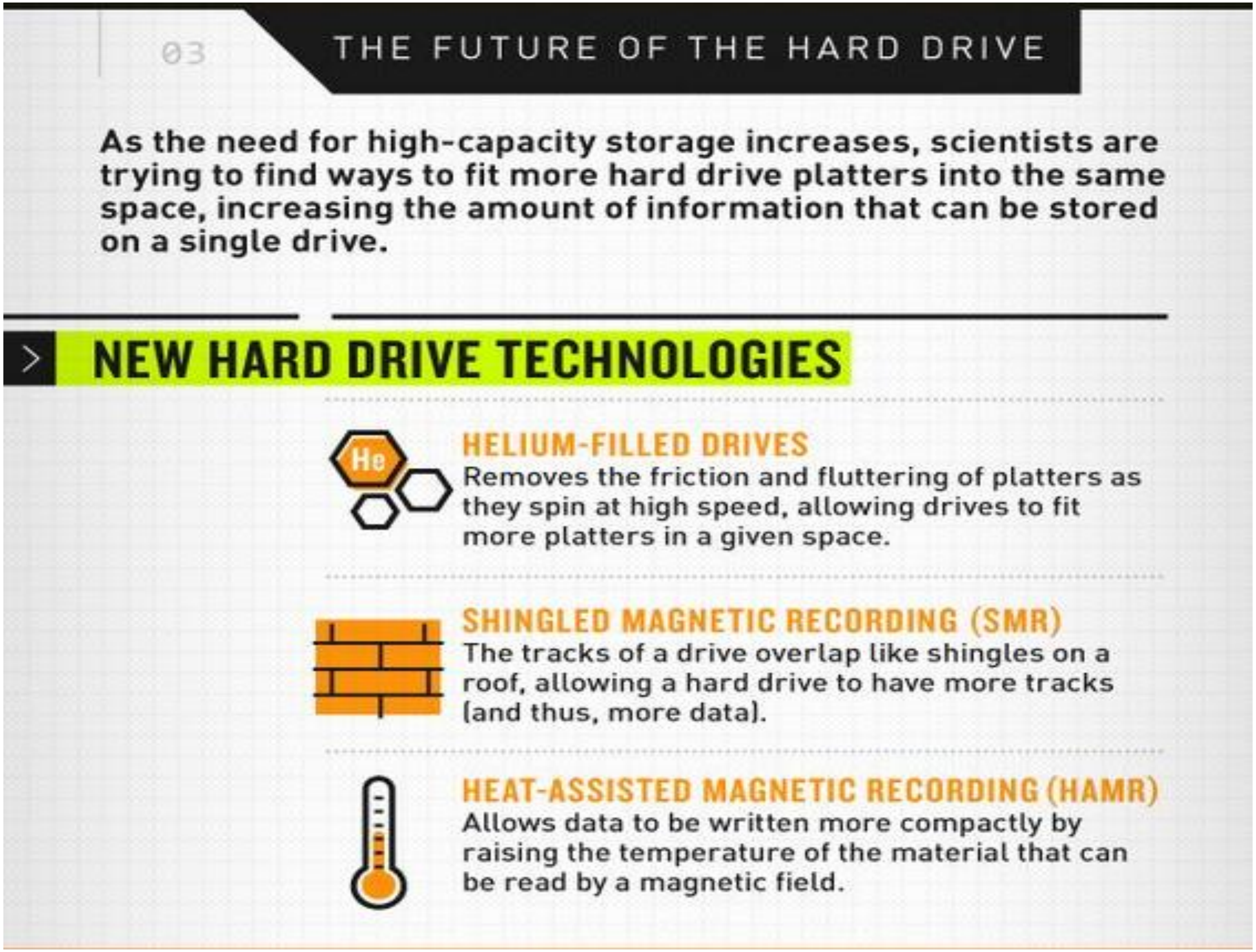
2.5 INCHES

›› 2013

ADATA DashDrive Air AE800 – a 500 GB wireless hard drive/ hotspot/ power bank for multiple mobile devices.

HOLDS

500 GB OF DATA



03

THE FUTURE OF THE HARD DRIVE

As the need for high-capacity storage increases, scientists are trying to find ways to fit more hard drive platters into the same space, increasing the amount of information that can be stored on a single drive.

NEW HARD DRIVE TECHNOLOGIES

HELIUM-FILLED DRIVES

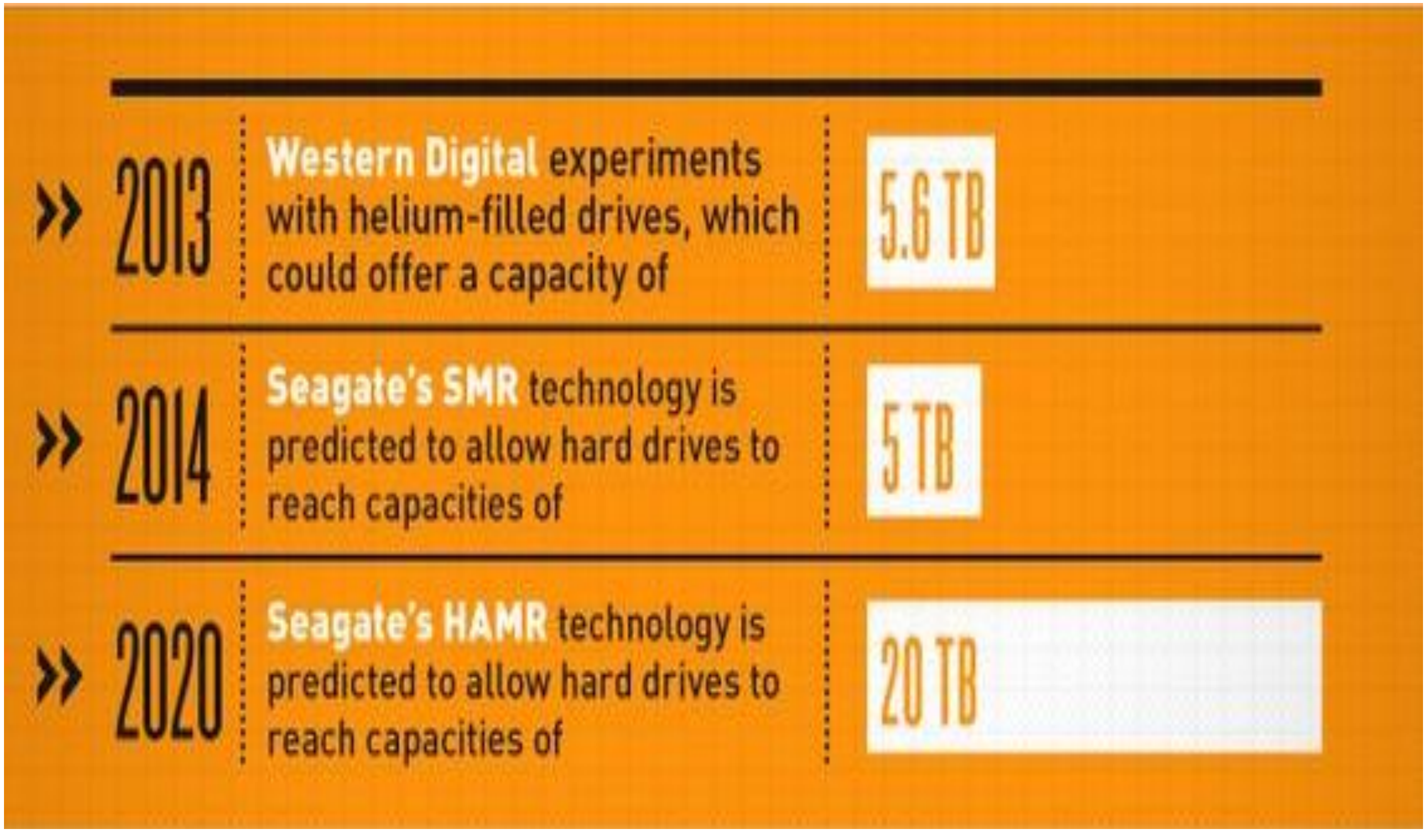
Removes the friction and fluttering of platters as they spin at high speed, allowing drives to fit more platters in a given space.

SHINGLED MAGNETIC RECORDING (SMR)

The tracks of a drive overlap like shingles on a roof, allowing a hard drive to have more tracks (and thus, more data).

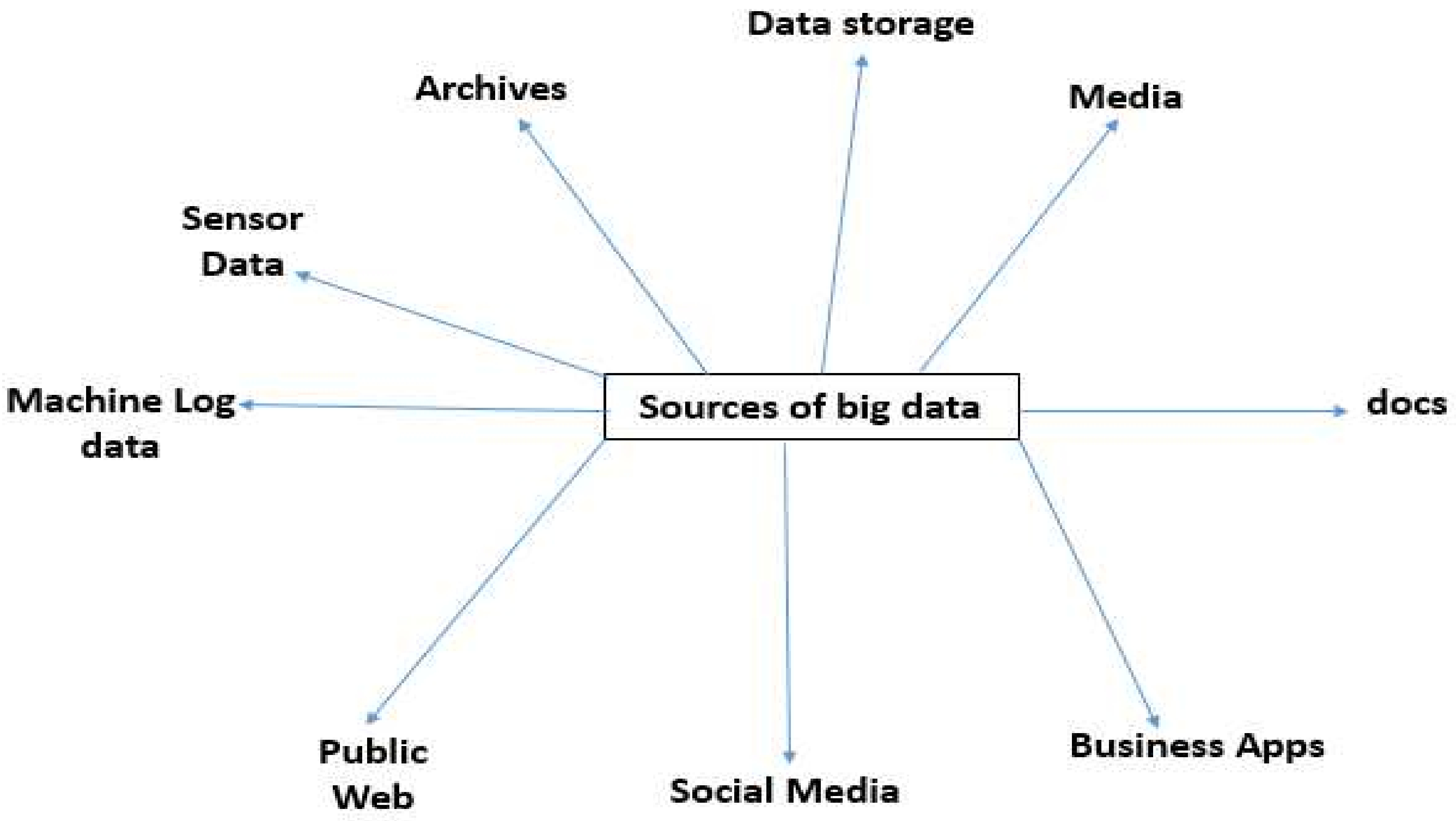
HEAT-ASSISTED MAGNETIC RECORDING (HAMR)

Allows data to be written more compactly by raising the temperature of the material that can be read by a magnetic field.



»	2013	**Western Digital** experiments with helium-filled drives, which could offer a capacity of	5.6 TB
»	2014	**Seagate's SMR** technology is predicted to allow hard drives to reach capacities of	5 TB
»	2020	**Seagate's HAMR** technology is predicted to allow hard drives to reach capacities of	20 TB

Sources of Big Data



Sources of big data

Data storage

Archives

Media

Sensor Data

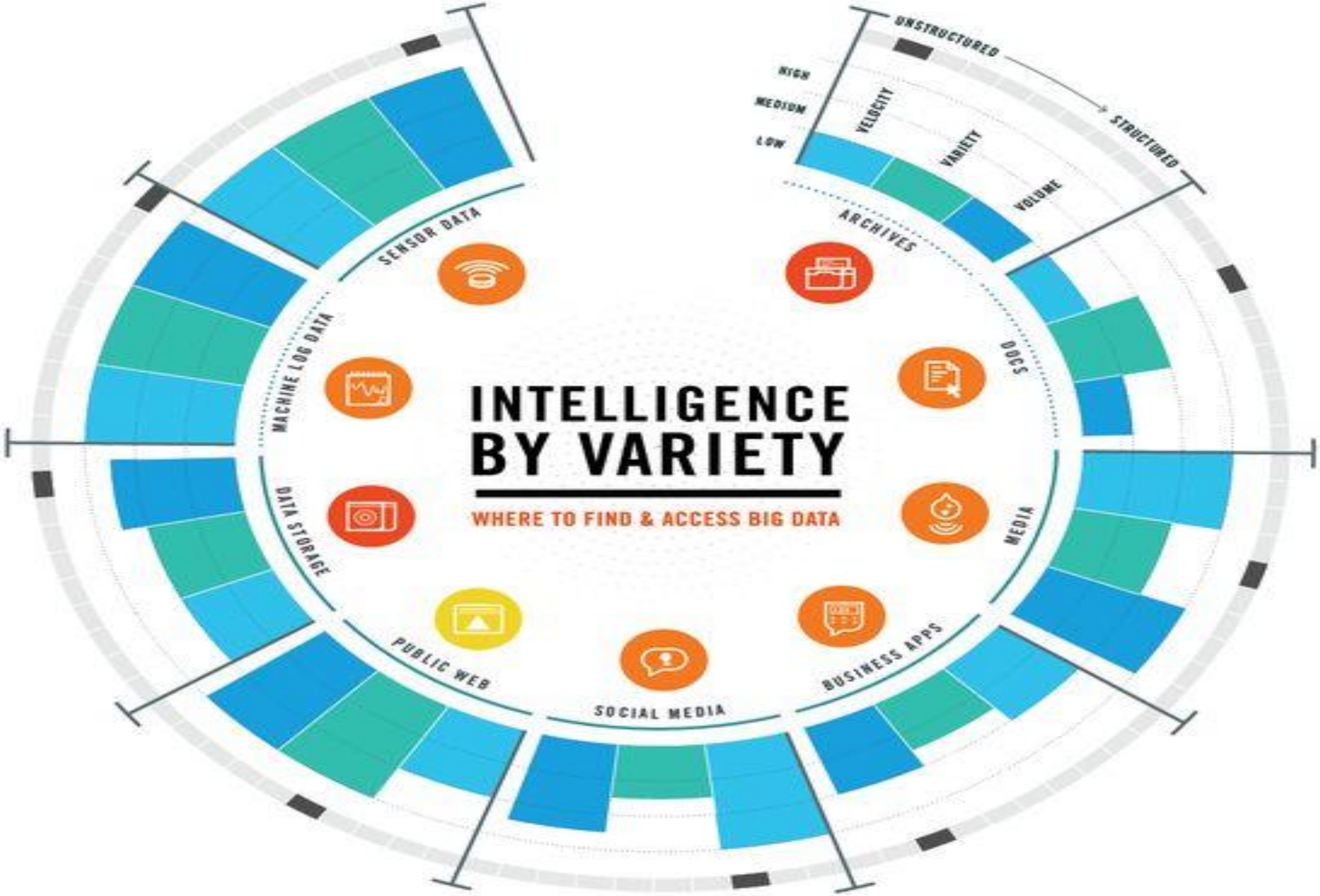
Machine Log data

docs

Public Web

Social Media

Business Apps



INTELLIGENCE BY VARIETY

WHERE TO FIND & ACCESS BIG DATA

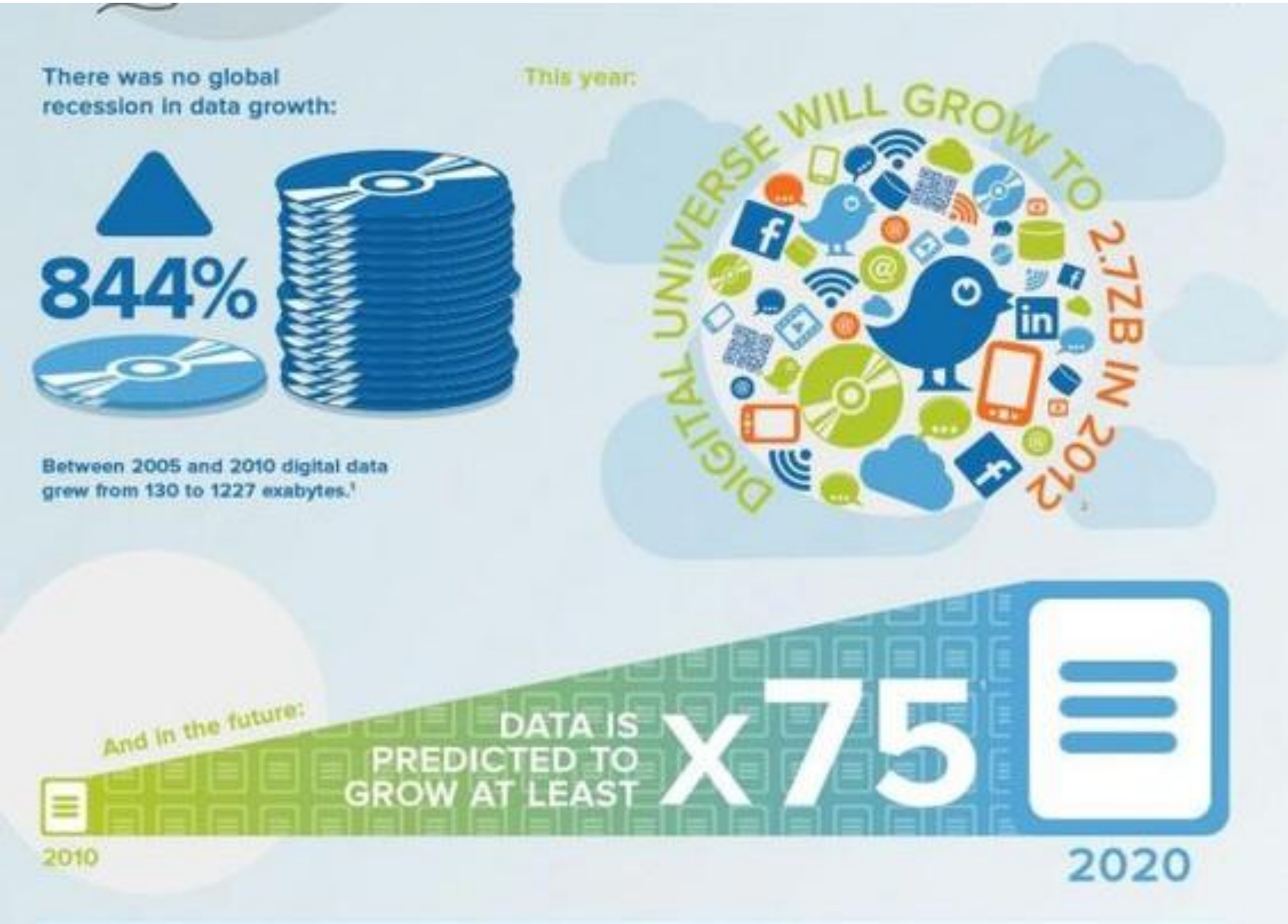
USING BIG DATA, ORGANIZATIONS CAN GENERATE ACTIONABLE INSIGHTS THAT ENABLE THEM TO DRIVE THEIR BUSINESS FORWARD. RAPID INTEGRATION OF THE EVER-EXPANDING POOL OF DATA SOURCES AND TYPES IS OPENING A WHOLE NEW WORLD OF POSSIBILITIES.

KEY

Velocity

- We have moved from the days of batch processing (remember payroll applications) to real-time processing (when you buy a product the website shows related product)

Batch → Periodic → Near real time → Real-time processing



There was no global recession in data growth:

844%

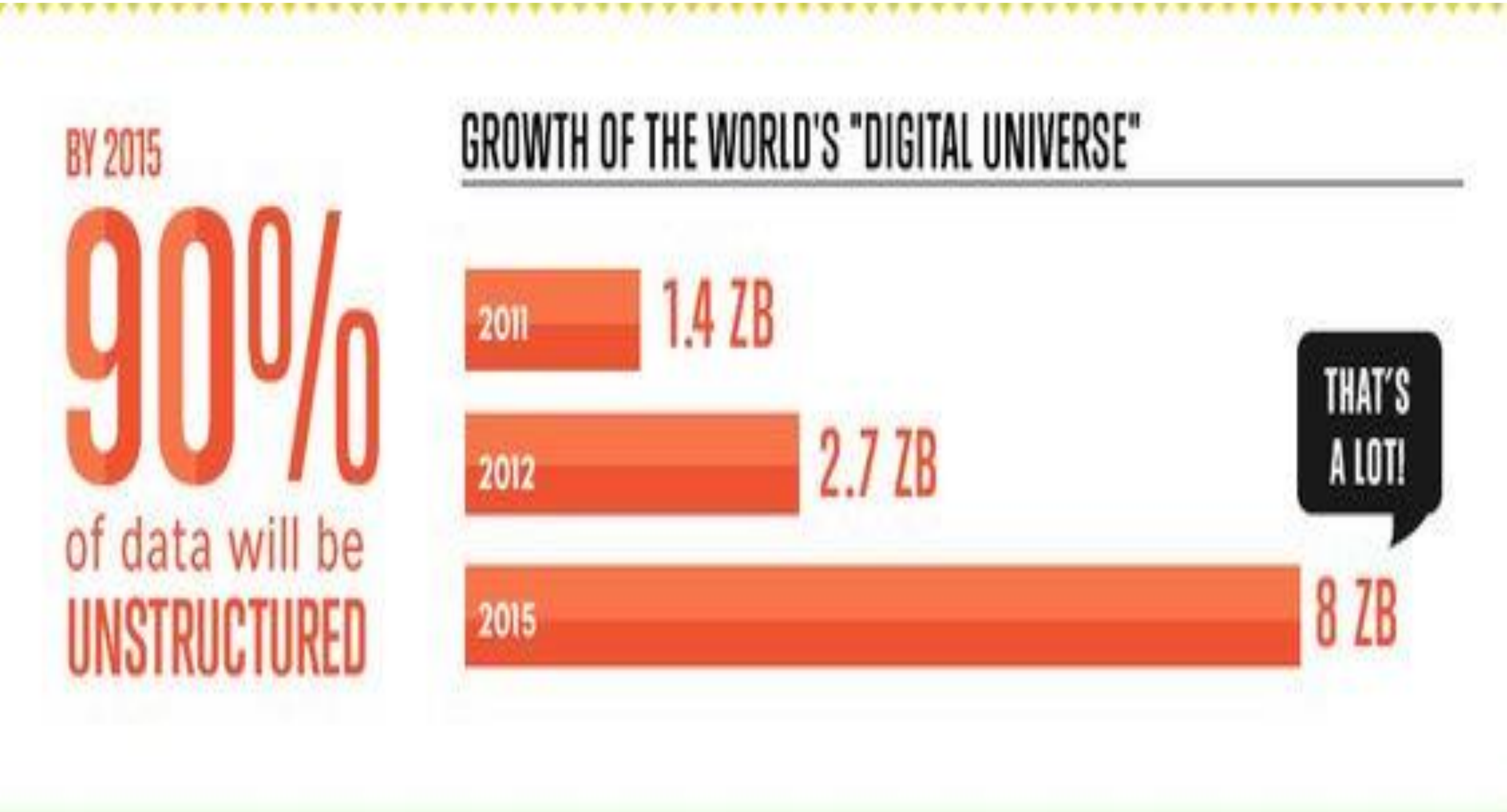
Between 2005 and 2010 digital data grew from 130 to 1227 exabytes.[1]

This year:

DIGITAL UNIVERSE WILL GROW TO 2.7ZB IN 2012

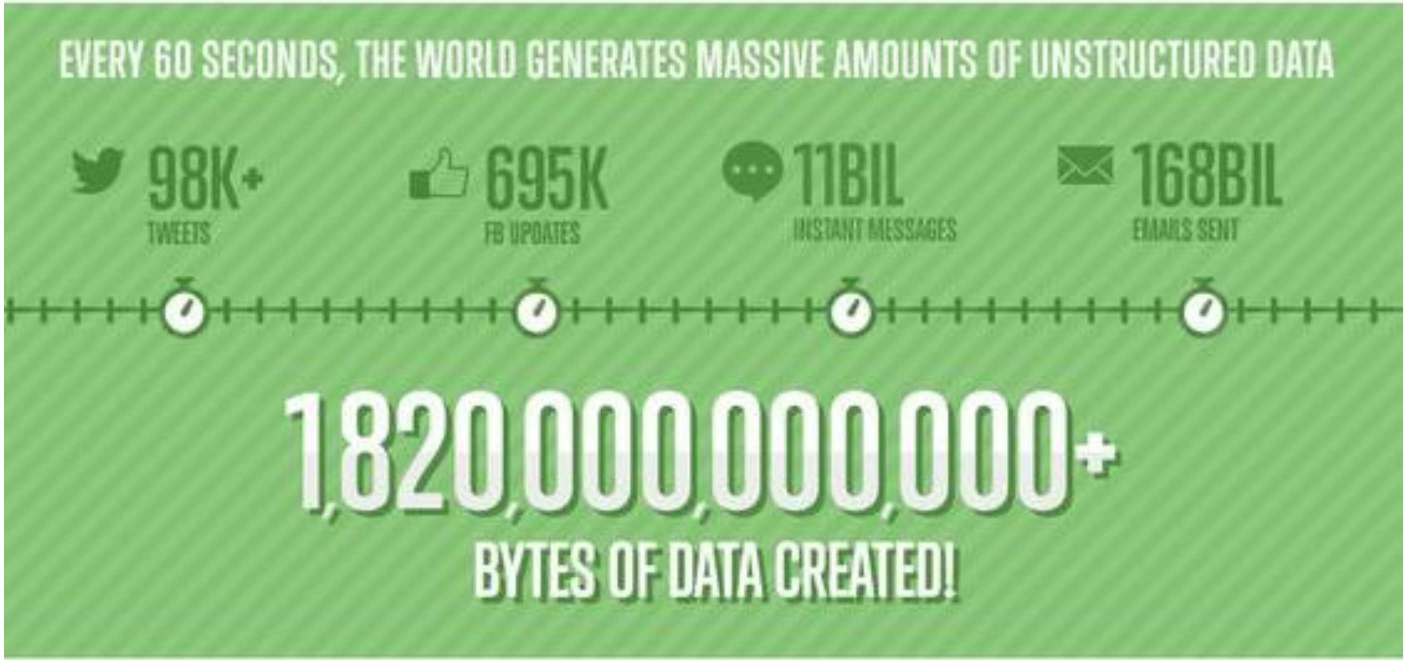
And in the future:

DATA IS PREDICTED TO GROW AT LEAST X75



BY 2015 90% of data will be UNSTRUCTURED

GROWTH OF THE WORLD'S "DIGITAL UNIVERSE"



EVERY 60 SECONDS, THE WORLD GENERATES MASSIVE AMOUNTS OF UNSTRUCTURED DATA

98K+ TWEETS

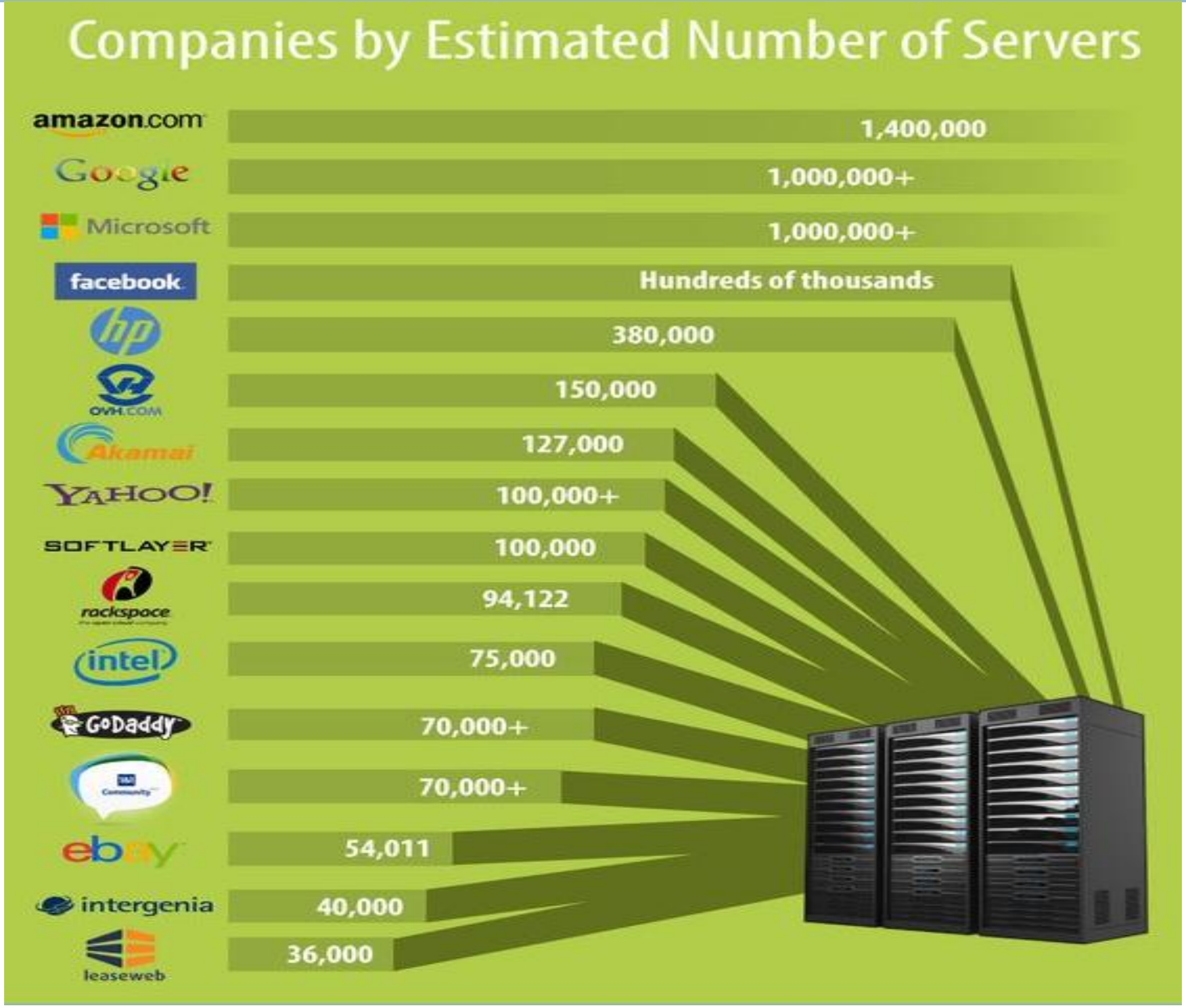
695K FB UPDATES

11BIL INSTANT MESSAGES

168BIL EMAILS SENT

1,820,000,000,000+ BYTES OF DATA CREATED!

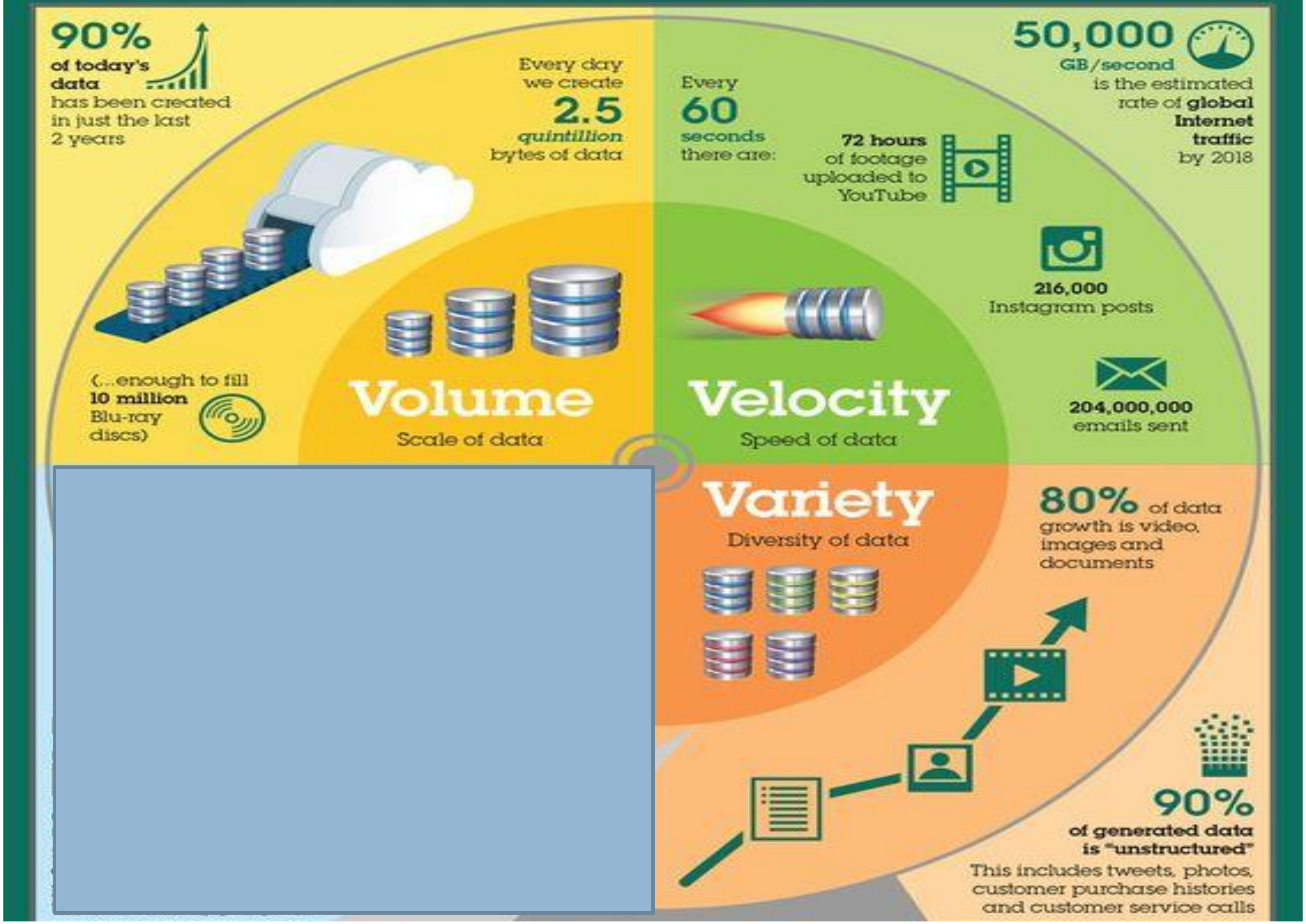
What volume does?



Variety

- **Structured data:** From traditional transaction processing systems and RDBMS, etc.
- **Semi-structured data:** For example: Hyper Text Markup Language (HTML), eXtensible Markup Language (XML).
- **Unstructured data:** For example: unstructured text documents, audio, video, email, photos, PDFs, social media, etc.

Extracting business value from the 3 V's of big data



Other Characteristics of Data

- Characteristics of data which are not Definitional Traits of Big Data
  - Veracity and Validity
  - Volatility
  - Variability

Veracity and Validity

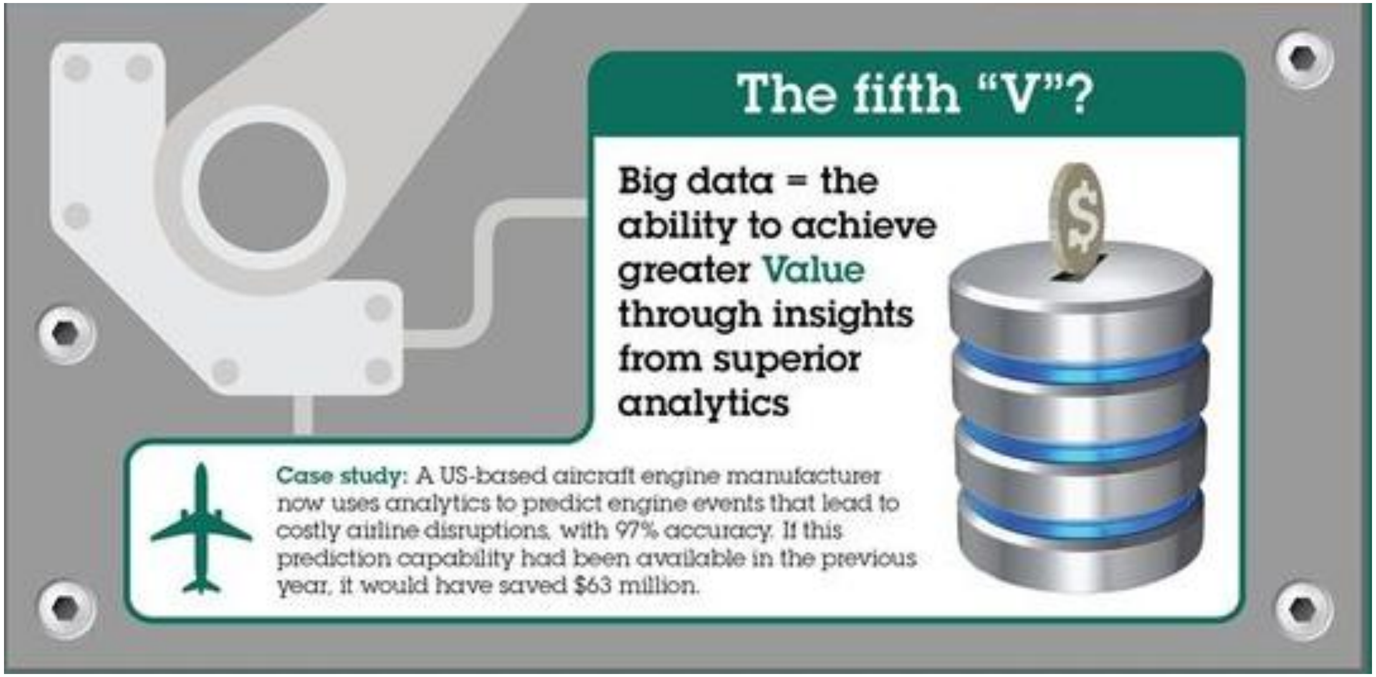
- *Veracity* refers to biases, noise, and abnormality in data. The key question here is: “Is all the data that is being stored, mined, and analyzed meaningful and pertinent to the problem under consideration?” *Validity* refers to the accuracy and correctness of the data. Any data that is picked up for analysis needs to be accurate. It is not just true about big data alone.

Volatility

- Volatility of data deals with, how long is the data valid? And how long should it be stored? There is some data that is required for long-term decisions and remains valid for longer periods of time. However, there are also pieces of data that quickly become obsolete minutes after their generation.

Variability

- Data flows can be highly inconsistent with periodic peaks. For example: An online retailer announces the "big sale day" for a particular week. The retailer is likely to experience an upsurge in customer traffic to the website during this week. In the same way, he/she might experience a slump in his/her business immediately after the festival season. This reemphasizes the point that one might witness spikes in data at some point in time and at other times, the data flow can go flat.



The fifth "V"?

Big data = the ability to achieve greater Value through insights from superior analytics

Case study: A US-based aircraft engine manufacturer now uses analytics to predict engine events that lead to costly airline disruptions, with 97% accuracy. If this prediction capability had been available in the previous year, it would have saved $63 million.

Challenges with Big Data

Following are a few challenges with big data:

1. **Usefulness:** Data today is growing at an exponential rate. Most of the data that we have today has been generated in the last 2—3 years. This high tide of data will continue to rise incessantly (persistently). The key questions here are: “Will all this data be useful for analysis?”, “Do we work with all this data or a subset of it?”, “How will we separate the knowledge from the noise?”, etc.

Challenges with Big Data cont…

2. **Cloud computing and virtualization:** Cloud computing is the answer to managing infrastructure for big data as far as cost-efficiency, elasticity, and easy upgrading/downgrading is concerned. This further complicates the decision to host big data solutions outside the enterprise.

Challenges with Big Data cont…

3. **Retention:** The other challenge is to decide on the period of retention of big data. Just how long should one retain this data? A tricky question indeed as some data is useful for making long-term decisions, whereas in few cases, the data may quickly become irrelevant and obsolete just a few hours after having being generated.

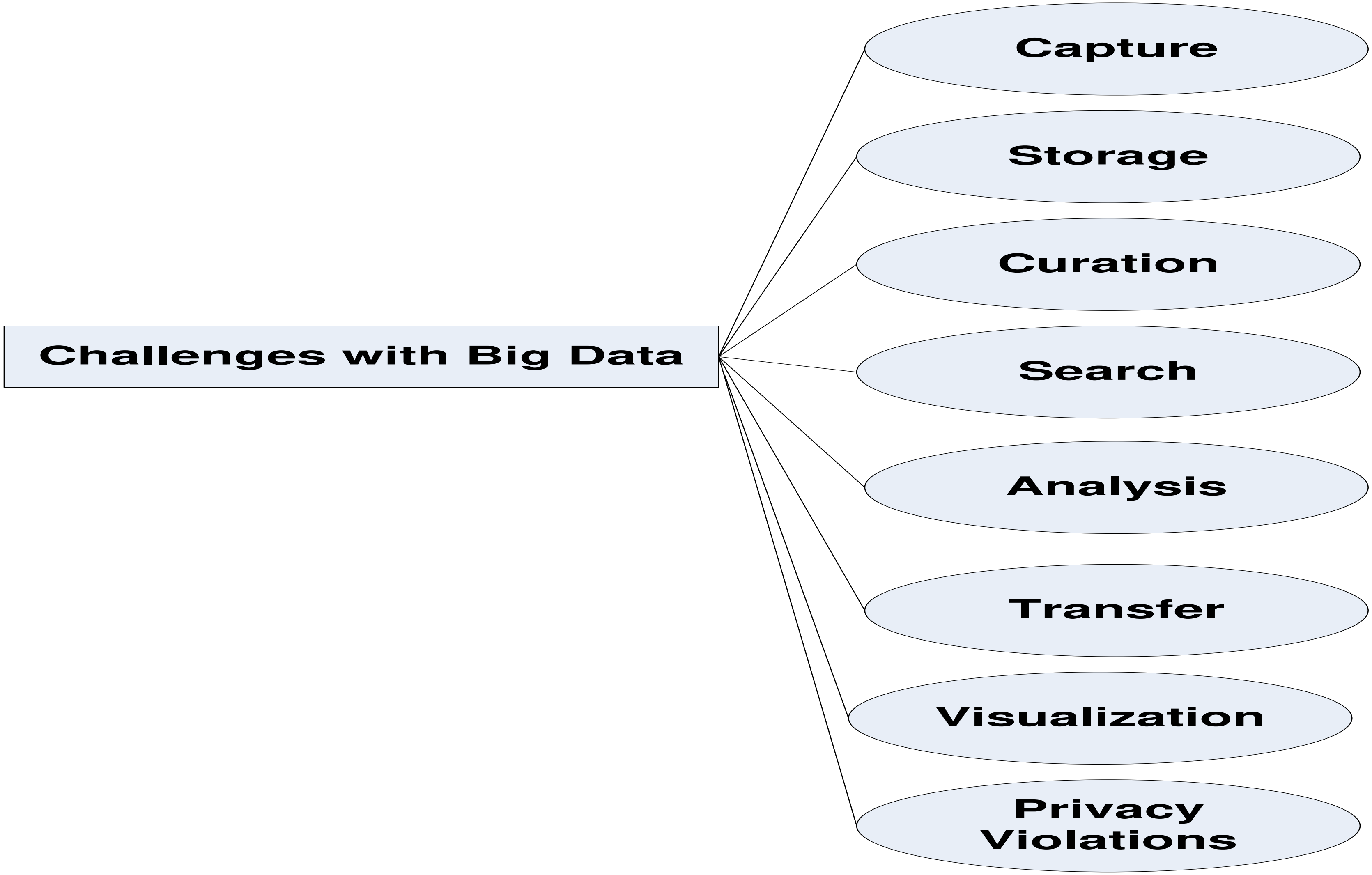
Challenges with Big Data cont…

4. **Scarcity of Data Scientist:** There is a dearth (shortage) of skilled professionals who possess a high level of proficiency in data sciences that is vital in implementing big data solutions.
5. **Data Visualization:** Then, of course, there are other challenges with respect to capture, storage, preparation, search, anal-ysis, transfer, security, and visualization of big data. Data visualization is becoming popular as a separate discipline.

Challenges with Big Data cont…

6. **Storage Capacity:** Big data refers to datasets whose size is typically beyond the storage capacity of traditional database software tools. There is no explicit definition of how big the dataset should be for it to be considered “big data.” Here we are to deal with data that is just too big, moves way to fast, and does not fit the structures of typical database systems.

Challenges with Big Data cont…



Fill in the blanks

- Big data is high-volume, __________, and high-variety information assets that demand __________, innovative forms of information processing for enhanced insight and decision making.
- ________, a Gartner analyst coined the term, ‘Big Data’
- ________, is the characterstic of data dealing with its retention
- ________, is the large data repository that stores data in its native format until it is needed
- ________ characteristic of data explains the spikes in data.
- Near real time or real time processing deals with __________ of data.

Question's Answer ??

- How is traditional BI environment different from the Big Data environment?

- Share your experience as a customer on an e-commerce site. Comment on the big data that gets created on a typical e-commerce site.