UNIVERSITY OF MALAYA

WQD7007 Big Data Management

Introduction to big data

Agenda

- What is big data?
- Characteristic of big data
- Big data and small data
- Four phases of big data
- When big data is needed?
- Key obstacles of the big data applications
- Application

What is Data?

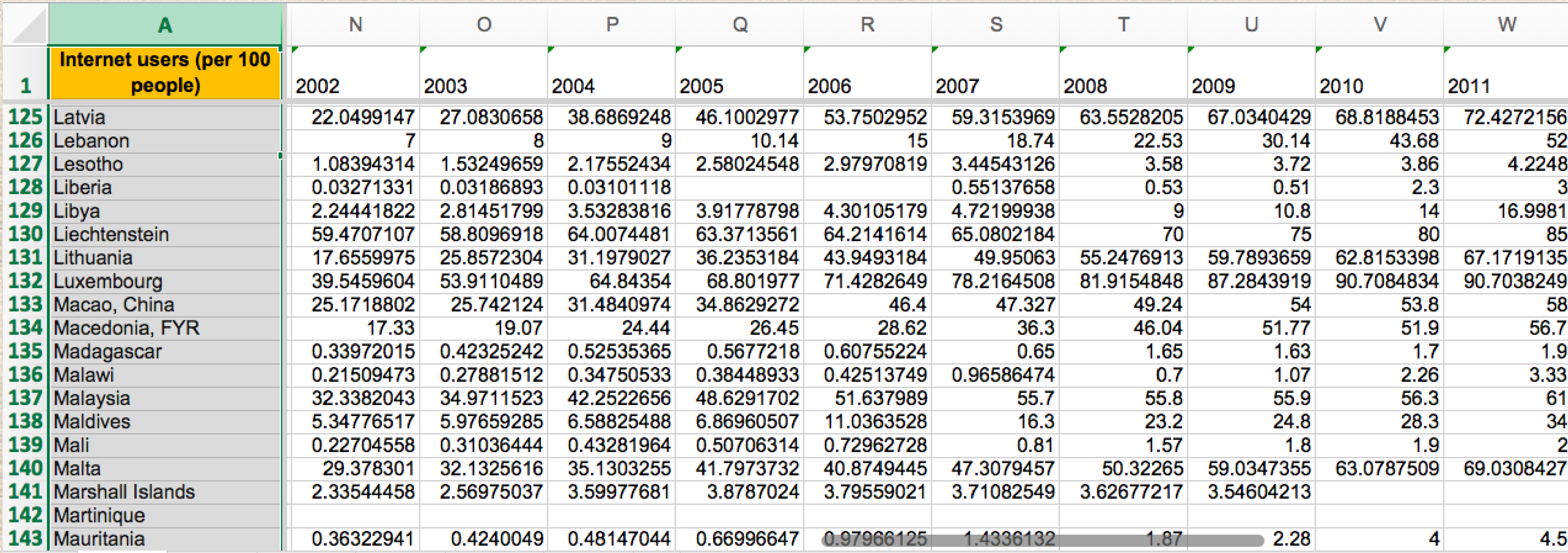
- Collection of Information
  - can be measured quantitatively or qualitatively
  - Example:
    - The room temperature now is 25 degree Celsius. $(25^{\mathrm{o}}\mathrm{C})$
    - The room is cold

What is Data?

- Collection of Information
  - can be measured quantitatively or qualitatively
  - Example:
    - The room temperature now is 25 degree Celsius. $(25^{o}\mathrm{C})$
    - The room is cold

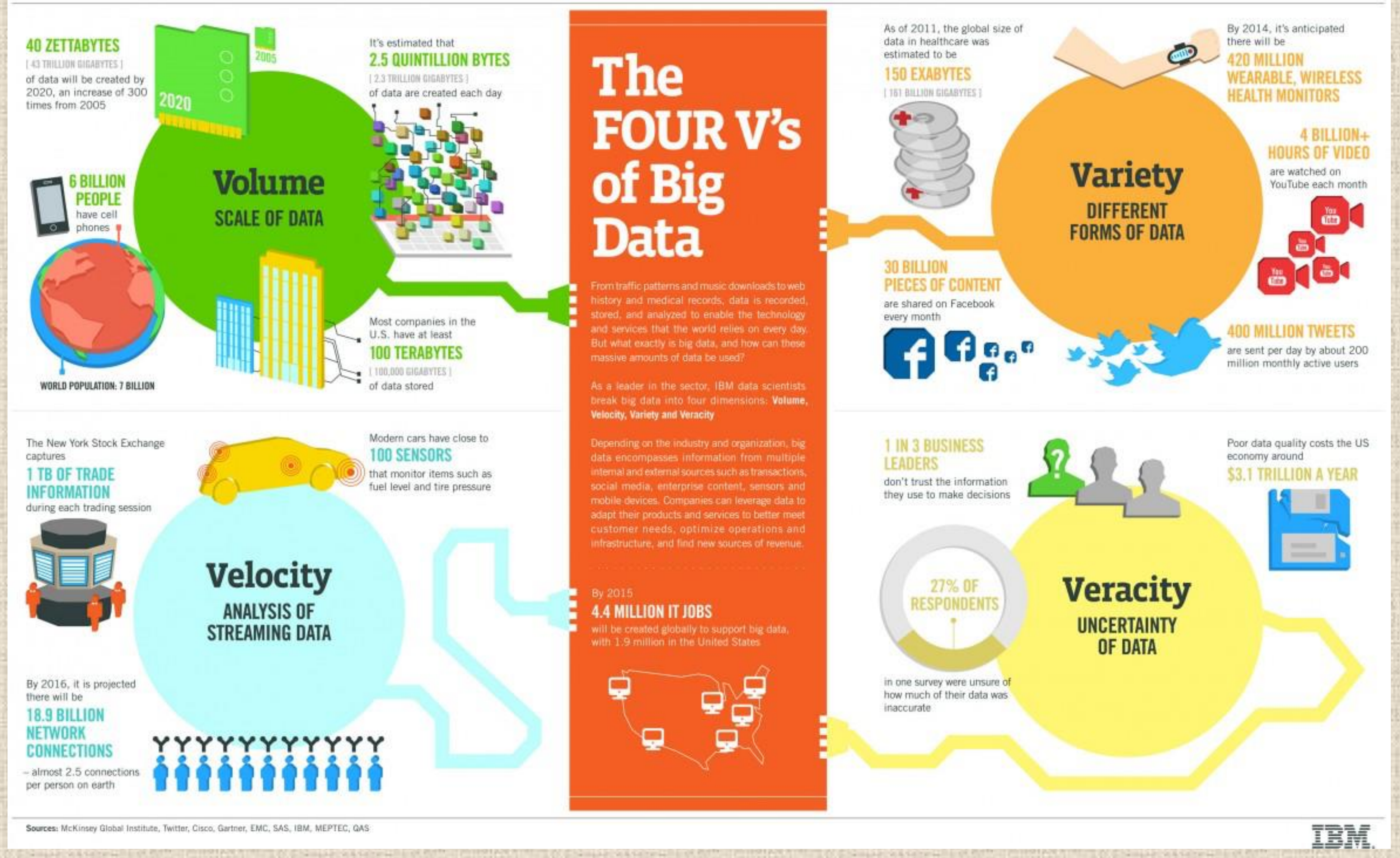
- How data is used?
- How database is setup?
- Why big data is needed?

Example: Internet users per 100 people



	A	N	O	P	Q	R	S	T	U	V	W
1	Internet users (per 100 people)	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
125	Latvia	22.0499147	27.0830658	38.6869248	46.1002977	53.7502952	59.3153969	63.5528205	67.0340429	68.8188453	72.4272156
126	Lebanon	7	8	9	10.14	15	18.74	22.53	30.14	43.68	52
127	Lesotho	1.08394314	1.53249659	2.17552434	2.58024548	2.97970819	3.44543126	3.58	3.72	3.86	4.2248
128	Liberia	0.03271331	0.03186893	0.03101118			0.55137658	0.53	0.51	2.3	3
129	Libya	2.24441822	2.81451799	3.53283816	3.91778798	4.30105179	4.72199938	9	10.8	14	16.9981
130	Liechtenstein	59.4707107	58.8096918	64.0074481	63.3713561	64.2141614	65.0802184	70	75	80	85
131	Lithuania	17.6559975	25.8572304	31.1979027	36.2353184	43.9493184	49.95063	55.2476913	59.7893659	62.8153398	67.1719135
132	Luxembourg	39.5459604	53.9110489	64.84354	68.801977	71.4282649	78.2164508	81.9154848	87.2843919	90.7084834	90.7038249
133	Macao, China	25.1718802	25.742124	31.4840974	34.8629272	46.4	47.327	49.24	54	53.8	58
134	Macedonia, FYR	17.33	19.07	24.44	26.45	28.62	36.3	46.04	51.77	51.9	56.7
135	Madagascar	0.33972015	0.42325242	0.52535365	0.5677218	0.60755224	0.65	1.65	1.63	1.7	1.9
136	Malawi	0.21509473	0.27881512	0.34750533	0.38448933	0.42513749	0.96586474	0.7	1.07	2.26	3.33
137	Malaysia	32.3382043	34.9711523	42.2522656	48.6291702	51.637989	55.7	55.8	55.9	56.3	61
138	Maldives	5.34776517	5.97659285	6.58825488	6.86960507	11.0363528	16.3	23.2	24.8	28.3	34
139	Mali	0.22704558	0.31036444	0.43281964	0.50706314	0.72962728	0.81	1.57	1.8	1.9	2
140	Malta	29.378301	32.1325616	35.1303255	41.7973732	40.8749445	47.3079457	50.32265	59.0347355	63.0787509	69.0308427
141	Marshall Islands	2.33544458	2.56975037	3.59977681	3.8787024	3.79559021	3.71082549	3.62677217	3.54604213		
142	Martinique										
143	Mauritania	0.36322941	0.4240049	0.48147044	0.66996647	0.97966125	1.4336132	1.87	2.28	4	4.5

Characteristics of big data



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015 **4.4 MILLION IT JOBS** will be created globally to support big data, with 1.9 million in the United States

Volume
SCALE OF DATA

40 ZETTABYTES [43 TRILLION GIGABYTES] of data will be created by 2020, an increase of 300 times from 2005

It's estimated that **2.5 QUINTILLION BYTES** [2.3 TRILLION GIGABYTES] of data are created each day

6 BILLION PEOPLE have cell phones

Most companies in the U.S. have at least **100 TERABYTES** [100,000 GIGABYTES] of data stored

Velocity
ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures **1 TB OF TRADE INFORMATION** during each trading session

Modern cars have close to **100 SENSORS** that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be **18.9 BILLION NETWORK CONNECTIONS** – almost 2.5 connections per person on earth

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be **150 EXABYTES** [161 BILLION GIGABYTES]

By 2014, it's anticipated there will be **420 MILLION WEARABLE, WIRELESS HEALTH MONITORS**

4 BILLION+ HOURS OF VIDEO are watched on YouTube each month

30 BILLION PIECES OF CONTENT are shared on Facebook every month

400 MILLION TWEETS are sent per day by about 200 million monthly active users

Veracity
UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS don't trust the information they use to make decisions

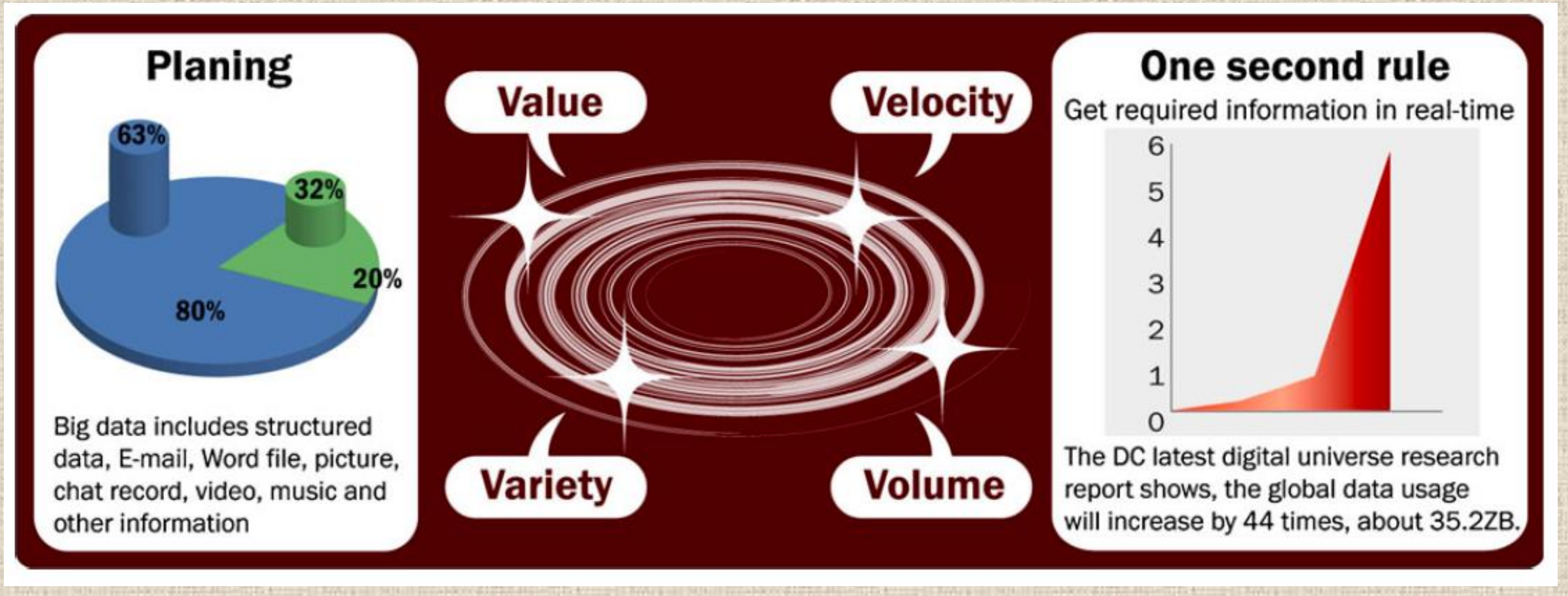
Poor data quality costs the US economy around **$3.1 TRILLION A YEAR**

27% OF RESPONDENTS in one survey were unsure of how much of their data was inaccurate

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

(Source: http://www.ibmbigdatahub.com/infographic/four-vs-big-data)

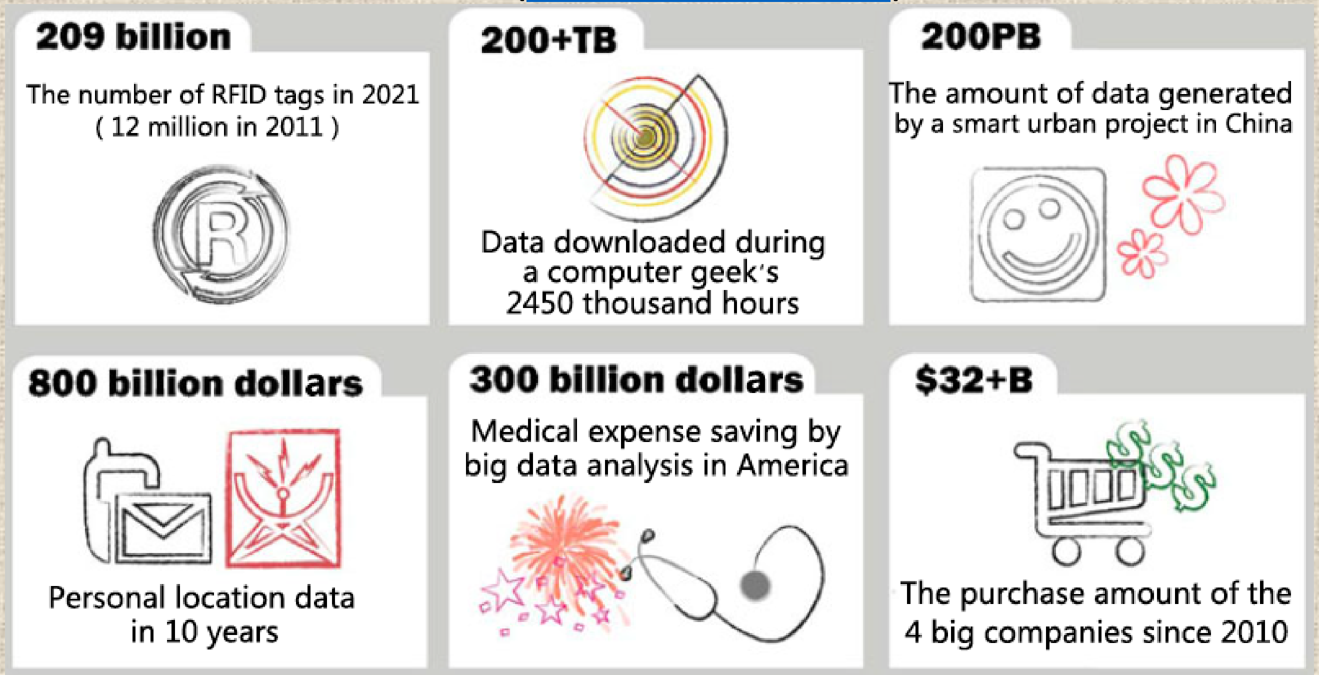
Characteristics of Big Data


Planing
63%
32%
20%
80%
Big data includes structured data, E-mail, Word file, picture, chat record, video, music and other information
Value
Velocity
Variety
Volume
One second rule
Get required information in real-time
6
5
4
3
2
1
0
The DC latest digital universe research report shows, the global data usage will increase by 44 times, about 35.2ZB.

[1] M. Chen, S. Mao and Y. Liu, "Big Data: A Survey", *Mobile Networks and Applications*, vol. 19, no. 2, pp. 171-209, 2014.

Characteristics of Big Data

- Volume
  - Large **amount** of data, big scale of data
  - Example:
    - 40 Zettabytes of data will be created by 2020, an increase of 300 times from 2015 (size of data units)



209 billion
The number of RFID tags in 2021
(12 million in 2011)

200+TB
Data downloaded during a computer geek's 2450 thousand hours

200PB
The amount of data generated by a smart urban project in China

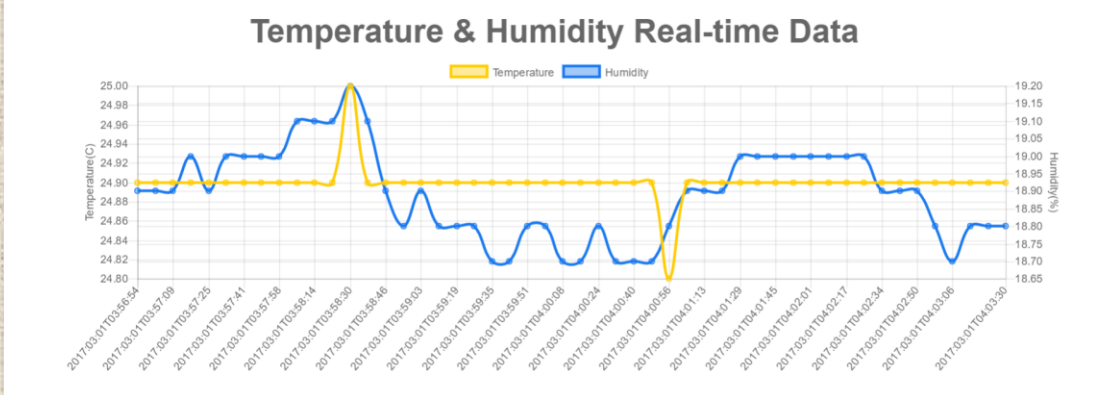
800 billion dollars
Personal location data in 10 years

300 billion dollars
Medical expense saving by big data analysis in America

$32+B
The purchase amount of the 4 big companies since 2010

Characteristics of Big Data

- Velocity
  - Analysis of **streaming** data
    - Trade data, Internet-of-Things (IoT) sensing, network data
  - the content of the data is constantly changing, through
    - the absorption of complementary data collections
    - previously archived data or legacy collections
    - from streamed data arriving from multiple sources



Source: https://docs.microsoft.com/en-us/azure/iot-hub/iot-hub-live-data-visualization-in-web-apps

Characteristics of Big Data

- Variety
  - Different **forms** of data
  - Example: traditional databases, text, images, documents, and complex records

Source: https://makeawebsitehub.com/social-media-sites/

Characteristics of Big Data

- Veracity
  - How to make sure the data is **accurate**?
  - Uncertainty of data exist, due to the **poor data quality**

  - Example: marketing automation system with false names and inaccurate contact information
  - One of three business leaders don't trust the information they use to make decisions

Characteristics of Big Data

- Variability
  - data whose **meaning is constantly changing**.
  - Example: language processing.
- Visualization
  - presenting data using graphs and charts
- Value
  - Get business insights
  - Know your customer

Source: http://dataconomy.com/2014/05/seven-vs-big-data/, https://www.impactradius.com/blog/7-vs-big-data/

Difference between big data and massive data

- Big Data
  - Most of the V’s should apply
- Massive data
  - enormous collections of simple-format records

- Big data resources are **not equivalent** to a large spreadsheet, and a big data resource is not analyzed in its totality.
- Big Data analysis is a **multistep** process whereby data is extracted, filtered, and transformed, with analysis often proceeding in a piecemeal, sometimes recursive, fashion.

Big data and small data

- Big data is **NOT**
  - small data that has become bloated to the point that it can no longer fit on a spreadsheet, or
  - a database that happens to be very large
- **False impression** from professionals who customarily work with relatively small data:
  - they can apply their spreadsheet and database skills directly to big data resources without mastering new skills and without adjusting to new analytic paradigms
  - Only the computer need to be improved

Big data and Small data - Difference

1. Goal:

- small data—Usually designed to answer a **specific** question or serve a particular goal.
- Big Data—Usually designed with a **goal in mind**, but the goal is **flexible** and the questions posed are protean.

Big data and Small data - Difference

1. Goal:

- Example: “to combine high-quality data from fisheries, Coast Guard, commercial shipping, and coastal management agencies for a growing data collection that can be used to support a variety of governmental and commercial management studies in the lower peninsula.”
  - a vague goal
  - there really is no way to completely specify
    - what the big data resource will contain
    - how the various types of data held in the resource will be organized, connected to other data resources, usefully analyzed.
  - Nobody can specify, with any degree of confidence, the ultimate destiny of any big data project. It usually comes as a surprise.

Big data and Small data - Difference

2. Location:

- small data—Typically, small data is contained **within one institution**, often on one computer, sometimes in one file.
- Big Data—Typically spread throughout electronic space, typically parceled onto **multiple Internet servers**, located anywhere on earth.

Big data and Small data - Difference

3. Data structure and content:

- small data—Ordinarily contains **highly structured data**.
  - The data domain is restricted to a single discipline or subdiscipline. The data often comes in the form of uniform records in an ordered spreadsheet.
- Big Data —**Must be capable of absorbing unstructured data** (e.g., such as free-text documents, images, motion pictures, sound recordings, physical objects).
  - The subject matter of the resource may cross multiple disciplines, and the individual data objects in the resource may link to data contained in other big data resources.

Big data and Small data - Difference

4. Data preparation

- small data —In many cases, the data user prepares her **own data**, for her **own purposes**.
- Big Data —The data comes from **many diverse sources,** and it is prepared by many people. People who use the data are seldom the people who have prepared the data.

Big data and Small data - Difference

5. Longevity

- small data —When the data project ends, the data is kept for a limited time (seldom longer than 7 years, the traditional academic life span for research data) and then discarded.
- Big Data —Big Data projects typically contain data that **must be stored in perpetuity**.
    - Ideally, data stored in a Big Data resource will be absorbed into another resource when the original resource terminates. Many Big Data projects extend into the future and the past (e.g., legacy data), accruing data prospectively and retrospectively.

Big data and Small data - Difference

6. Measurements

- small data —Typically, the data is measured using one experimental protocol, and the data can be represented using **one set of standard units**
- Big Data —Many different types of data are delivered in **many different electronic formats**.
  - Measurements, when present, may be obtained by **many different protocols**. Verifying the quality of big data is one of the most difficult tasks for data managers.

Big data and Small data - Difference

7. Reproducibility

- small data —Projects are typically **repeatable**.
  - If there is some question about the quality of the data, reproducibility of the data, or validity of the conclusions drawn from the data, the entire project can be repeated, yielding a new data set.
- Big Data —Replication of a Big Data project is **seldom feasible**.
  - In most instances, all that anyone can hope for is that bad data in a big data resource will be found and flagged as such.

Big data and Small data - Difference

8. Stakes/cost

- small data —Project costs are **limited**.
  - Laboratories and institutions can usually recover from the occasional small data failure.
- Big Data —Big Data projects can be obscenely **expensive**.

Big data and Small data - Difference

9. Introspection/observation

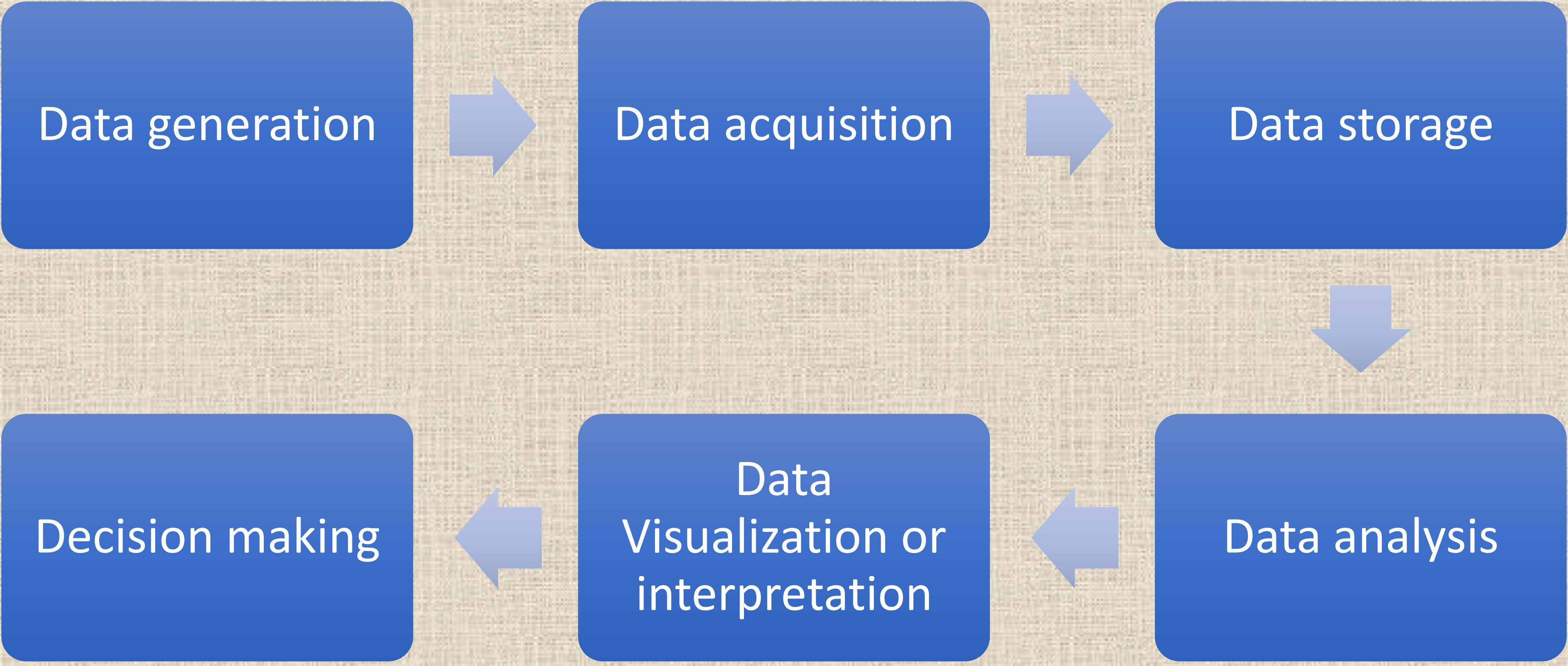
- small data —Individual data points are identified by their row and column location within a spreadsheet or database table.
  - If you know the row and column headers, you can find and specify all of the data points contained within.
- Big Data —Unless the Big Data resource is exceptionally well designed, the contents and organization of the resource can be inscrutable, even to the data managers

Big data and Small data - Difference

10. Analysis

- small data —In most instances, all of the data contained in the data project can be **analyzed together, and all at once**.
- Big Data —With few exceptions, such as those conducted on supercomputers or in parallel on multiple computers, Big Data is ordinarily analyzed **in incremental steps**.
  - The data are extracted, reviewed, reduced, normalized, transformed, visualized, interpreted, and reanalyzed with different methods.

Six phases of Big Data



Data Generation

- Enterprise Data
  - online trading data and online analysis data, most of which are historically static data
  - Internal data e.g. production data, inventory data, sales data, and financial data which aims to capture informationized and data-driven activities in enterprises, so as to record all activities of enterprises
- IoT (Internet-of-Things) data
  - smart cities, industry, agriculture, traffic, transportation, medical care, public departments, and families
- Bio-medical data
  - Medical images and electronic medical records

Data Acquisition

- Data collection
  - Log files, sensing, network data
- Data transmission/transportation
  - Inter-DCN transmissions: Inter-DCN transmissions are from data source to data center, which is generally achieved with the existing physical network infrastructure.
  - Intra-DCN Transmissions: Intra-DCN transmissions are the data communication flows within data centers.
- Data pre-processing
  - Integration, cleaning and redundancy elimination

(DCN – Dynamic circuit network)

Data Storage

- Storage system for massive data
  - Direct Attached Storage (DAS)
  - Network Attached Storage (NAS)
  - Storage Area Network (SAN)
  - Distributed storage system
- Storage mechanism:
  - Traditional relational databases
    - cannot meet the challenges on categories and scales brought about by big data.
  - NoSQL databases (i.e., non traditional relational databases) feature flexible modes, support for simple and easy copy, simple API, eventual consistency, and support of large volume data.
    - Example: Key-value databases, column-oriented databases, and document-oriented databases, each based on certain data models.

Data Analysis

- Traditional data analysis
  - Cluster Analysis
  - Factor Analysis
  - Correlation Analysis
  - Regression Analysis
  - A/B testing
  - Statistical analysis
  - Data mining algorithms

- Big data analytics method:
  - Bloom Filter
  - Hashing
  - Index
  - Triel
  - Parallel Computing

Data Visualization and Interpretation

- To represent knowledge more intuitively and effectively by using different graphs
  - e.g. Tableau

Decision Making

- Decision is made based on the analytical results
  - Is Big Data really helpful in improving decision making process?

When Big Data is needed?

Case 1:

- An entity has collected a lot of data, in the course of its **normal activities**, and seeks to organize the data so that materials can be retrieved, as needed.
- The Big Data effort is intended to streamline the regular activities of the entity. In this case, **the data is just waiting to be used.** The entity is **not looking to discover anything or to do anything new**.
  - Example: **Medical center** as an “accidental” Big Data resource. The administrative staff anticipated that the collected data can be used to achieve mandated goals: e.g. improving quality of service, increasing staff efficiency, and reducing operational costs.

When Big Data is needed?

Case 2:

- An entity has collected a lot of data in the course of its normal activities and **decides that there are many new activities that could be supported** by their data.
  - Example: Consider modern corporations—these entities **do not restrict themselves to one manufacturing process or one target audience**. They are constantly looking for **new opportunities**.
  - Their collected data may enable them to develop **new products** based on the preferences of their loyal customers, to **reach new markets**, or to market and distribute items via the Web. These entities will become hybrid Big Data/manufacturing enterprises.

When Big Data is needed?

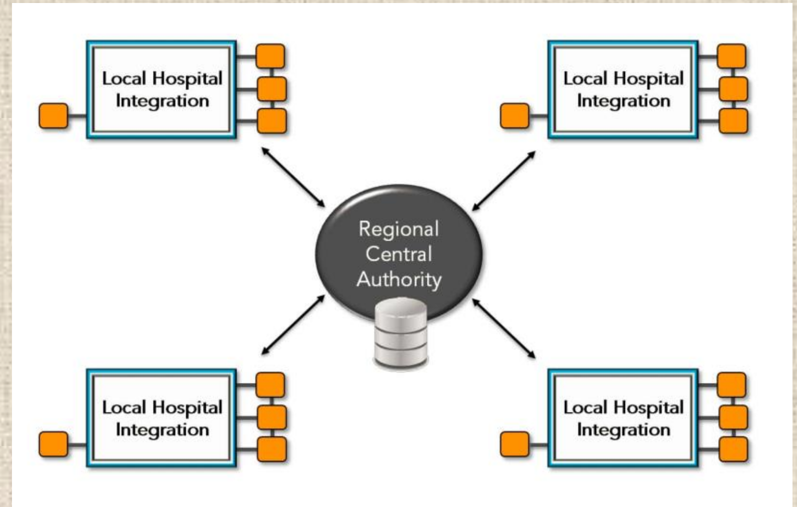
Case 3:

- An entity **plans a business model** based on a Big Data resource. Unlike the previous entities, this entity starts with Big Data and adds a physical component secondarily.
  - Amazon and FedEx may fall into this category, as they began with a plan for **providing a data-intense service** (e.g., the Amazon Web catalog and the FedEx package-tracking system). The traditional tasks of warehousing, inventory, pickup, and delivery had been available all along, but lacked the novelty and efficiency afforded by Big Data.

When Big Data is needed?

Case 4:

- An entity is part of a group of entities that have large data resources, all of whom understand that it would be to their mutual advantage to **federate** (centralize) their data resources.
  - Example: hospital databases that share electronic medical health records

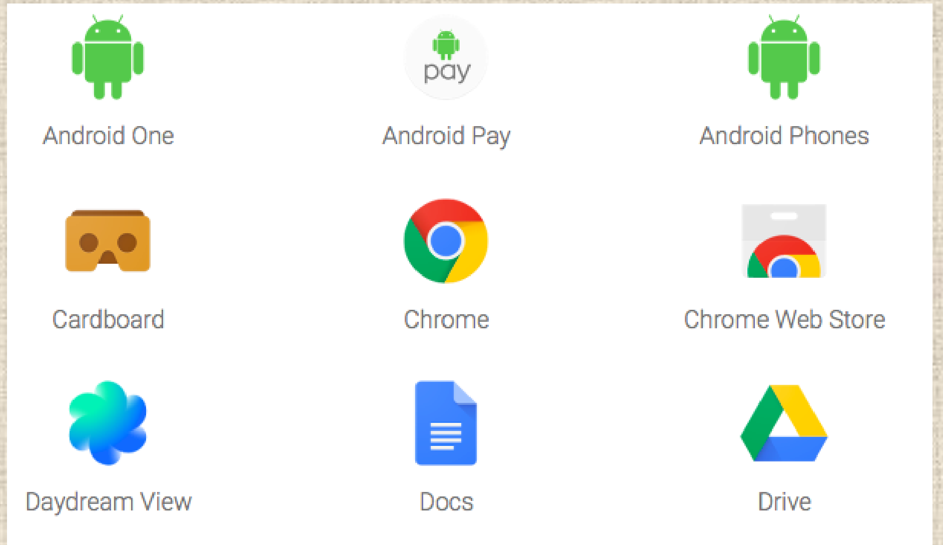

Local Hospital Integration
Local Hospital Integration
Regional Central Authority
Local Hospital Integration
Local Hospital Integration

Source: https://corepointhealth.com/health-information-exchange-architecture-types

When Big Data is needed?

Case 5:

- An entity with skills and vision **develops a project wherein large amounts of data are collected and organized** to the **benefit of themselves and their user-clients**.
  - Example: Google, and its many services



Source: https://www.google.com/about/products/

When Big Data is needed?

Case 6:

- An entity has **no data and has no particular expertise** in Big Data technologies, but it **has money and vision**.
- The entity seeks to **fund and coordinate** a group of data creators and data holders who will build a Big Data resource that can be used by others.
    - Example: Government agencies have been the major benefactors. These Big Data projects are justified if they lead to important discoveries that could not be attained at a lesser cost, with smaller data resources.

When Big Data is needed?

The most common purpose of Big Data is to produce SMALL DATA

- Imagine using a restaurant locater on your smartphone:
  1. With a few taps, it lists the Italian restaurants located within a 10 block radius of your current location.
  2. The database being queried is big and complex (a map database, a collection of all the restaurants in the world, their longitudes and latitudes, their street addresses, and a set of ratings provided by patrons, updated continuously)
  3. the data that it yields is small (e.g., five restaurants, marked on a street map, with pop-ups indicating their exact address, telephone number, and ratings).
  4. Your task comes down to selecting one restaurant from among the five and dining thereat.

When Big Data is needed?

The most common purpose of Big Data is to produce SMALL DATA

- your data selection was drawn from a large data set, but **your ultimate analysis was confined to a small dataset** (i.e., five restaurants meeting your search criteria).
  - The purpose of the Big Data resource was to proffer the small dataset. No analytic work was performed on the Big Data resource—just search and retrieval. **The real labor of the Big Data resource involved collecting and organizing complex data so that the resource would be ready for your query.** Along the way, the data creators had many decisions to make
  - e.g., Should bars be counted as restaurants? What about take-away only shops? What data should be collected? How should missing data be handled? How will data be kept current?

Industrial Application

- Big data VS cloud services
  - Oracle and IBM
- Big data VS Internet-of-Things
  - Smarthome, VYROX
- Big data VS data center
  - Digital realty trust, Equinix
- Big data VS Hadoop
  - Cloudera, Hortonworks

Source: https://www.smarthome.com , http://vyrox.com , http://www.datacenterknowledge.com/archives/2017/01/20/here-are-the-10-largest-data-center-providers-in-the-world

Key Obstacles of the Big Data applications

1. Data representation
   - many datasets have certain levels of heterogeneity in type, structure, semantics, organization, granularity, and accessibility.
   - aims to make data more meaningful for computer analysis and user interpretation.
   - an **improper data representation will reduce the value of the original data** and may even obstruct effective data analysis.
   - Efficient data representation shall reflect data structure, class, and type, as well as integrated technologies, so as to **enable efficient operations on different datasets**.

Key Obstacles of the Big Data applications

2. Redundancy reduction and data compression
    - Generally, the redundancy level in datasets are high.
    - Aim to **reduce the indirect cost of the entire system** on the premise that the potential values of the data are not affected.
        - Example: most data generated by sensor networks are highly redundant, which may be filtered and compressed at orders of magnitude.

Key Obstacles of the Big Data applications

3. Data life cycle management
   - compared with the relatively slow advances of storage systems, pervasive sensing and computing are generating data at unprecedented rates and scales.
   - current storage system could not support such massive data.
   - **In general**, values hidden in big data depend on **data freshness**. Therefore, a data importance principle related to the analytical value should be developed to decide **which data shall be stored and which data shall be discarded**

Key Obstacles of the Big Data applications

4. Analytical mechanism

   - traditional RDBMSs are strictly designed with a lack of scalability and expandability
   - non-relational databases have shown their unique advantages in the processing of unstructured data. But still exist problems in their performance in particular applications.
   - **a compromising solution between RDBMSs and non-relational databases should be identified.**
     - Example: some enterprises have utilized a mixed database architecture that integrates the advantages of both types of database (e.g., Facebook and Taobao).

Key Obstacles of the Big Data applications

5. Data confidentiality
   - most big data service providers or owners at present could not effectively maintain and analyze such huge datasets because of their limited capacity.
   - They must **rely on professionals or tools** to analyze such data, which **increase the potential safety risks**.
     - Example: the transactional dataset generally includes a set of complete operating data to drive key business processes.
     - Such data contains details of the lowest granularity and **some sensitive information such as credit card numbers**.
     - Therefore, analysis of big data may be delivered to a **third party** for processing only when proper preventive measures are taken to protect such sensitive data, to ensure its safety.

Key Obstacles of the Big Data applications

6. Energy management
    - the energy consumption of mainframe computing systems has drawn much attention from both **economy and environment** perspectives.
    - With the increase of data volume and analytical demands, the processing, storage, and transmission of big data will inevitably consume more and more electric energy.
        - Therefore, system-level power consumption control and management mechanism shall be established for big data while the expandability and accessibility are ensured.

Key Obstacles of the Big Data applications

7. Expendability and scalability
   - the analytical system of big data must support **present and future datasets**.
     - The analytical algorithm must be able to process **increasingly expanding and more complex datasets**.
8. Cooperation
   - analysis of big data is an interdisciplinary research, which requires **experts in different fields cooperate to harvest the potential of big data**.
     - A comprehensive big data network architecture must be established to help scientists and engineers in various fields access different kinds of data and fully utilize their expertise, so as to cooperate to complete the analytical objectives.

Coming Next

- Big data pipeline using Hadoop

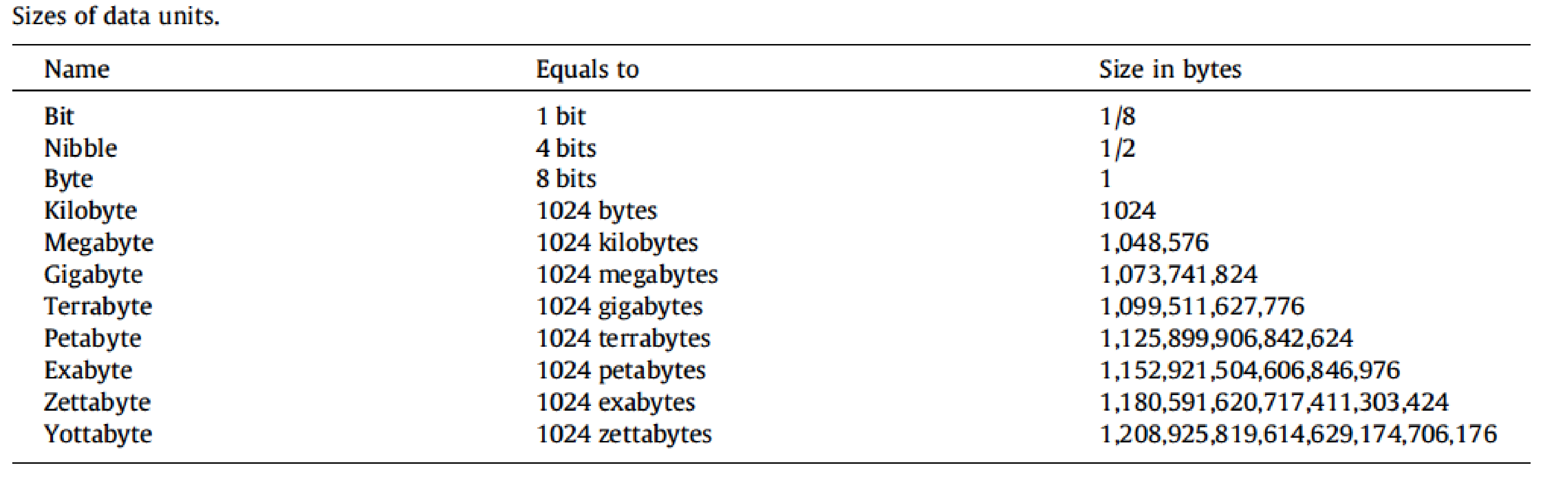
References

[1] M. Chen, S. Mao and Y. Liu, "Big Data: A Survey", *Mobile Networks and Applications*, vol. 19, no. 2, pp. 171-209, 2014.

[2] C. Philip Chen and C. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data", 2014.

[3] J. Berman, *Principles of big data*. Amsterdam: Elsevier, 2013.

Size of data units



Sizes of data units.

Name	Equals to	Size in bytes
Bit	1 bit	1/8
Nibble	4 bits	1/2
Byte	8 bits	1
Kilobyte	1024 bytes	1024
Megabyte	1024 kilobytes	1,048,576
Gigabyte	1024 megabytes	1,073,741,824
Terrabyte	1024 gigabytes	1,099,511,627,776
Petabyte	1024 terrabytes	1,125,899,906,842,624
Exabyte	1024 petabytes	1,152,921,504,606,846,976
Zettabyte	1024 exabytes	1,180,591,620,717,411,303,424
Yottabyte	1024 zettabytes	1,208,925,819,614,629,174,706,176

Back