Introduction to Big Data

Module 5

Data vs Information ??

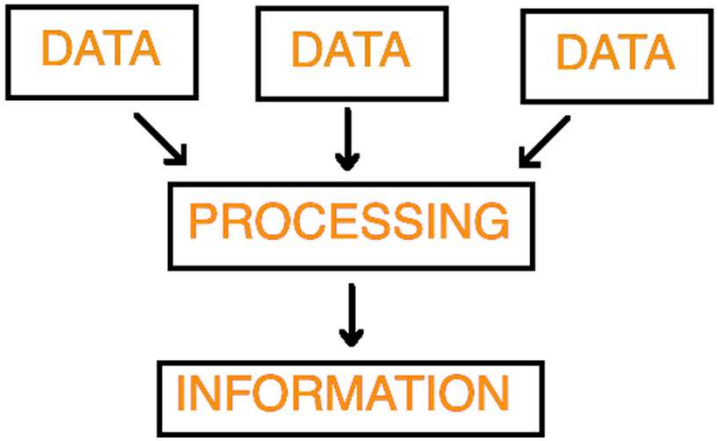
Data vs. Information

Data

- raw facts
- no context
- just numbers and text

Information

- data with context
- processed data
- value-added to data
  - summarized
  - organized
  - analyzed



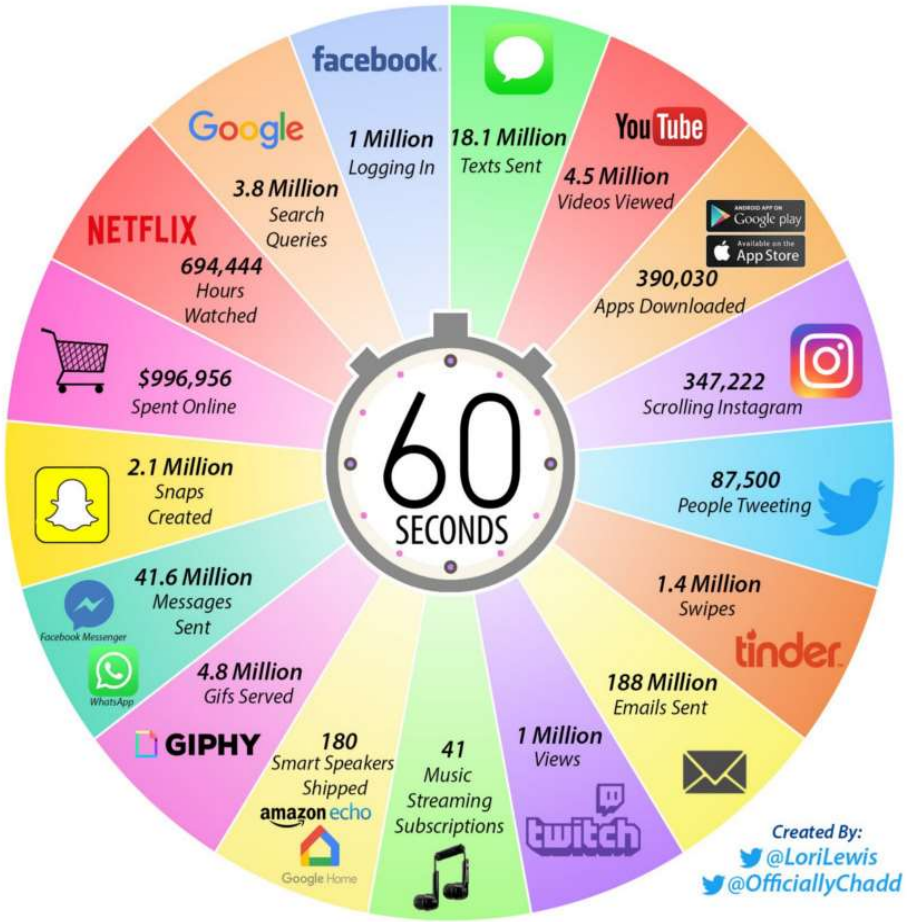
Courtesy: [1], [2]

Examples of Data and Information

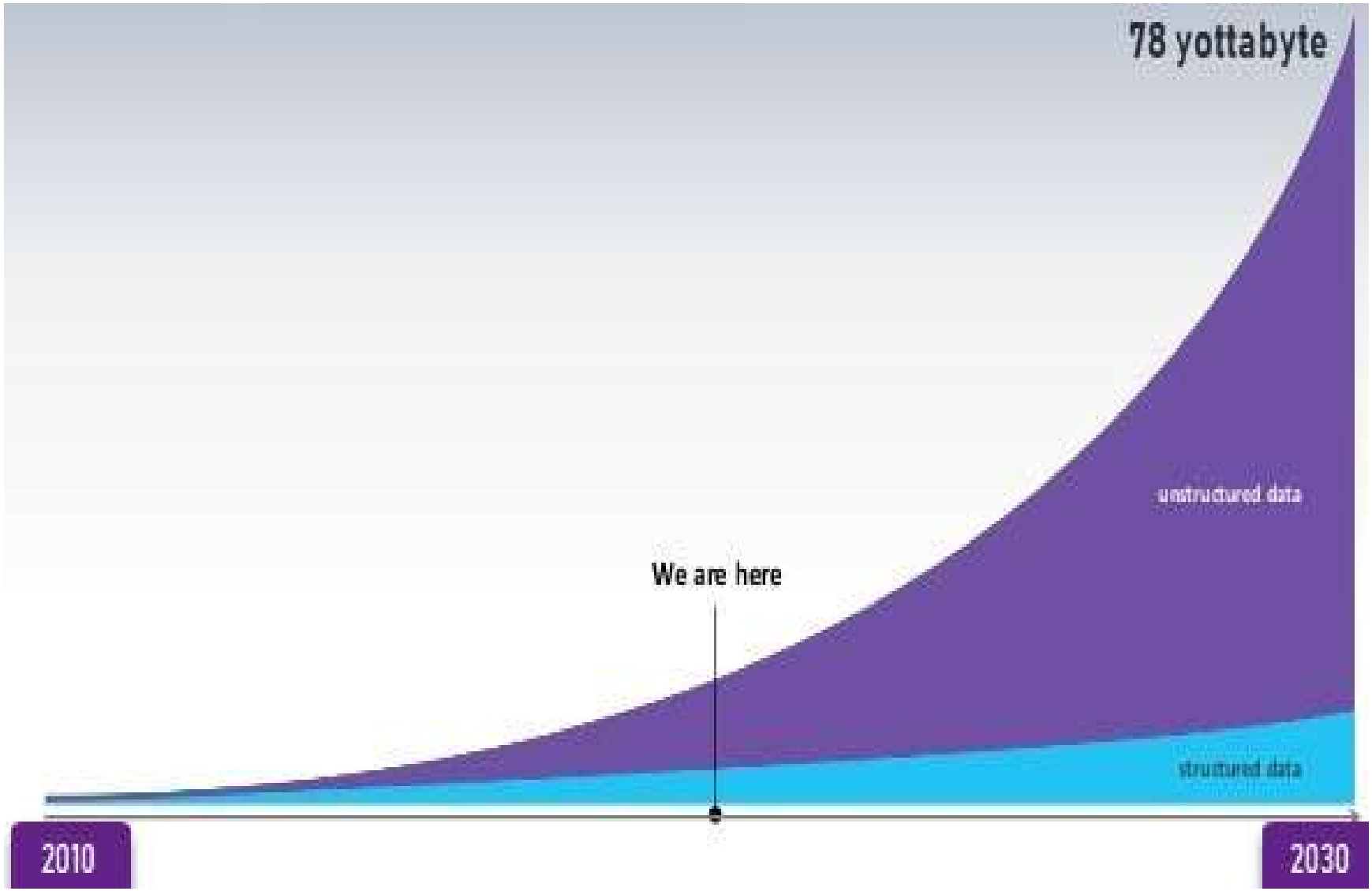
- The history of temperature readings all over the world for the past 100 years is data. If this data is organized and analyzed to find that global temperature is rising, then that is information.
- The number of visitors to a website by country is an example of data. Finding out that traffic from the U.S. is increasing while that from Australia is decreasing is meaningful information.

How much data do we generate?

- **Structured Data**
  - Relational Databases
  - Well defined schema
- **Unstructured Data**
  - Videos, audio, images etc.
- **Semi-structured Data**
  - structured in form but not well defined (no schema)
  - XML files



The Rise of Unstructured Data



How large your data is?

- What is the maximum file size you have dealt so far?
  - Movies/files/streaming video that you have used?
- What is the maximum download speed you get?
  - To retrieve data stored in distant locations?
- How fast your computation is?
  - Time to transfer, process and get result?

Memory unit	Size	Binary size
kilobyte (kB/KB)	10^3	2^{10}
megabyte (MB)	10^6	2^{20}
gigabyte (GB)	10^9	2^{30}
terabyte (TB)	10^{12}	2^{40}
petabyte (PB)	10^{15}	2^{50}
exabyte (EB)	10^{18}	2^{60}
zettabyte (ZB)	10^{21}	2^{70}
yottabyte (YB)	10^{24}	2^{80}

Courtesy: [3]

Sources of Data

- "Every day, we create 2.5 quintillion (10^{18}) bytes of data
  - 90% of the data in the world today has been created in the last two years alone.
  - The data come from several sources
    - *sensors used to gather climate information*
    - *posts to social media sites,*
    - *digital pictures and videos*
    - *purchase transaction records*
    - *cell phone GPS signals*

      etc.

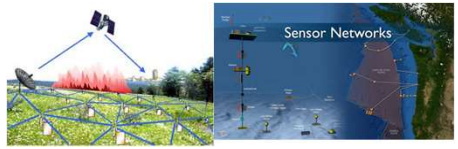
Courtesy: [3]

Examples

Social media and networks
(All of us are generating data)

Scientific instruments
(Collecting all sorts of data)

Mobile devices
(Tracking all objects all the time)



Sensor technology and networks
(Measuring all kinds of data)

Courtesy: [3]

Now Data is Big data!

- No single standard definition!
- 'Big-data' is similar to 'Small-data', but bigger

  …but having data bigger consequently requires different approaches

    - techniques, tools and architectures

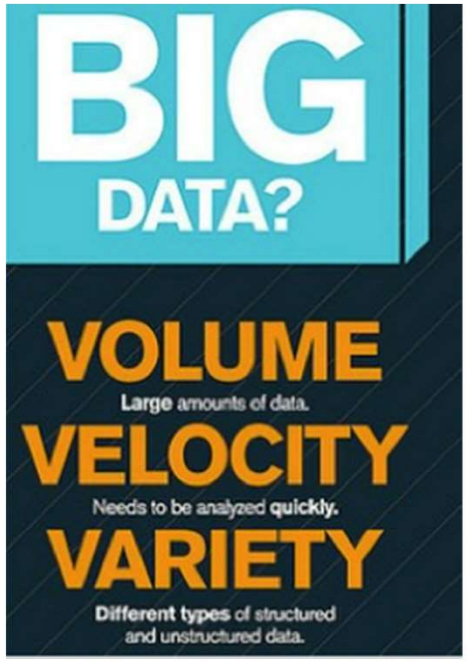
  …to solve: new problems

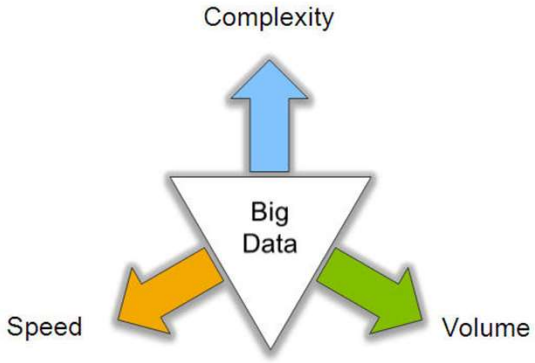
    …and, of course, in a better way

Big data is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and **analytics** to manage it and extract value and hidden knowledge from it.

Courtesy: [3]

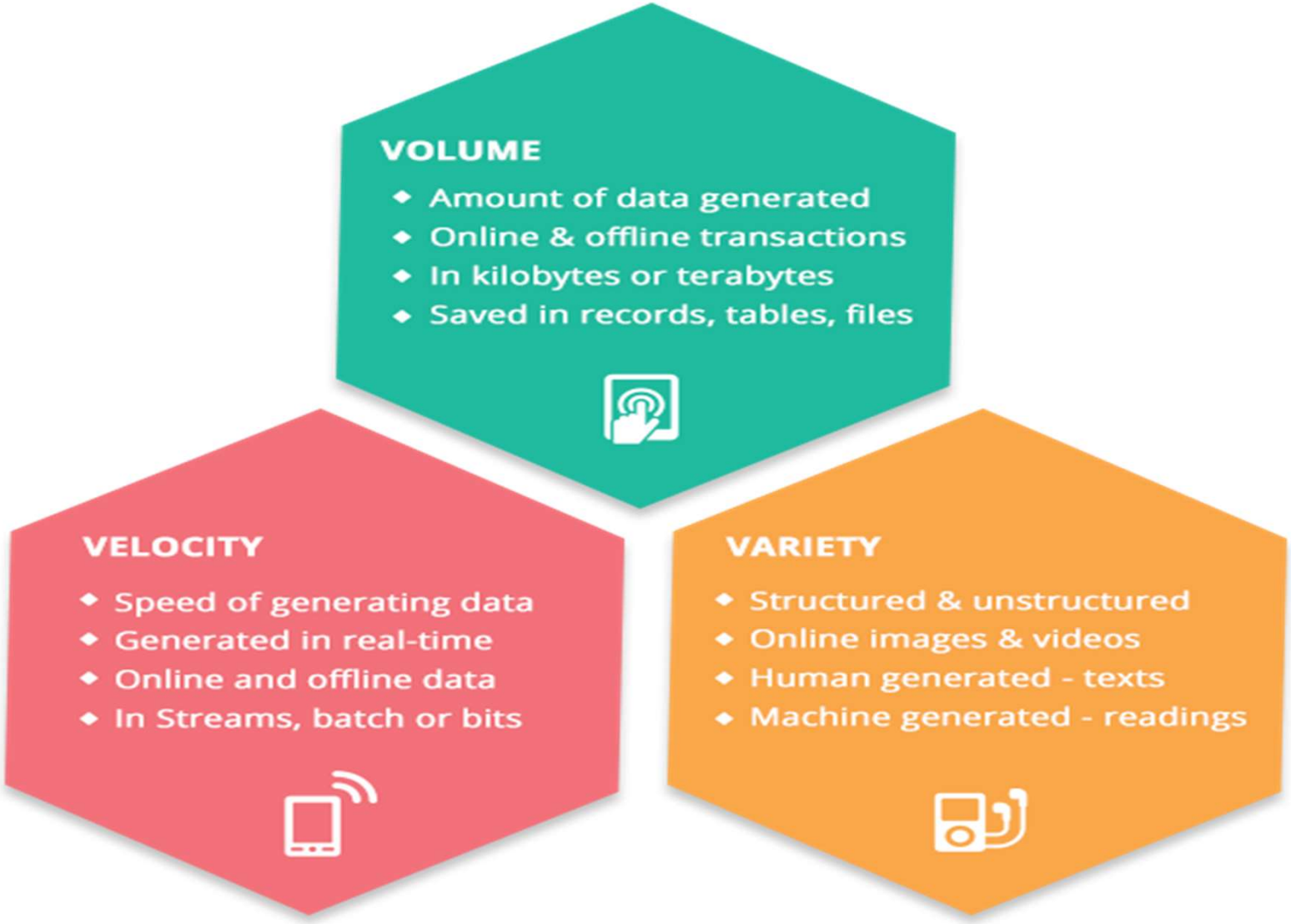
Characteristics of Big data: **V3**





Courtesy: [3]

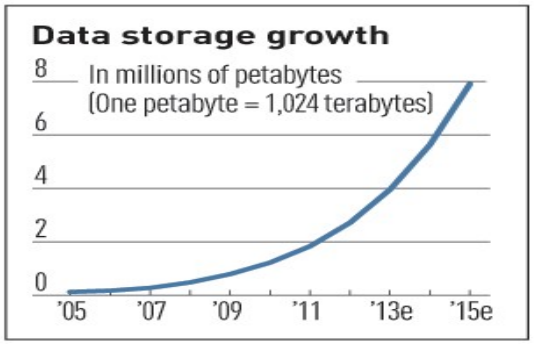
THE 3Vs OF BIG DATA

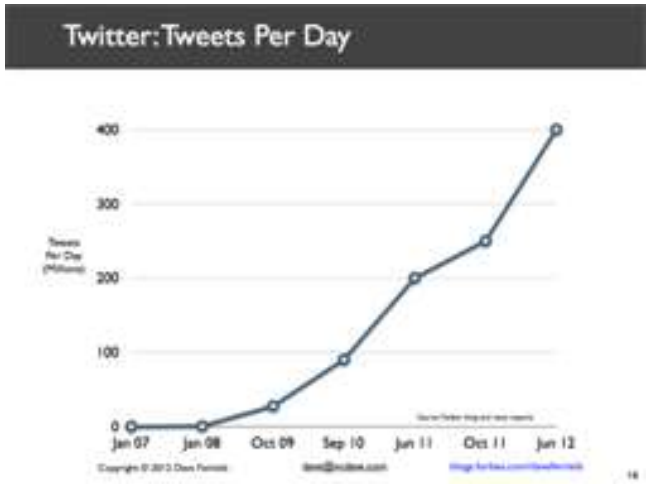


www.whishworks.com

V3 : V for Volume

- Large volume of data, More computation
  - More tools and techniques required
- needs to be processed rapidly
  - More storage capacity



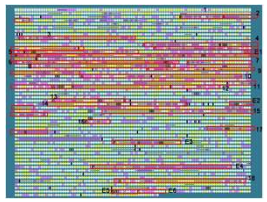


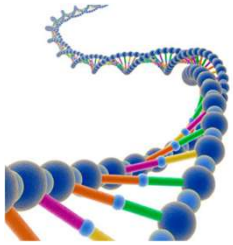
Exponential increase in collected/generated data

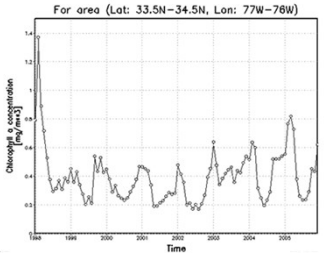
Courtesy: [3]

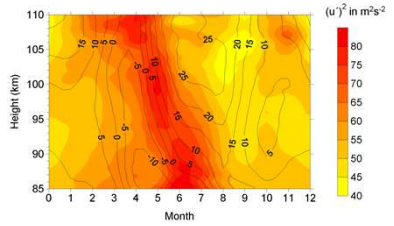
V3: V for Variety

- Various formats, types, and structures
  - Text, numerical, images, audio, video, sequences, time series, social media data, multi-dimensional arrays, etc…
- Static data vs. streaming data
- A single application can be generating/collecting many types of data









To extract knowledge➔ all these types of data need to be linked together

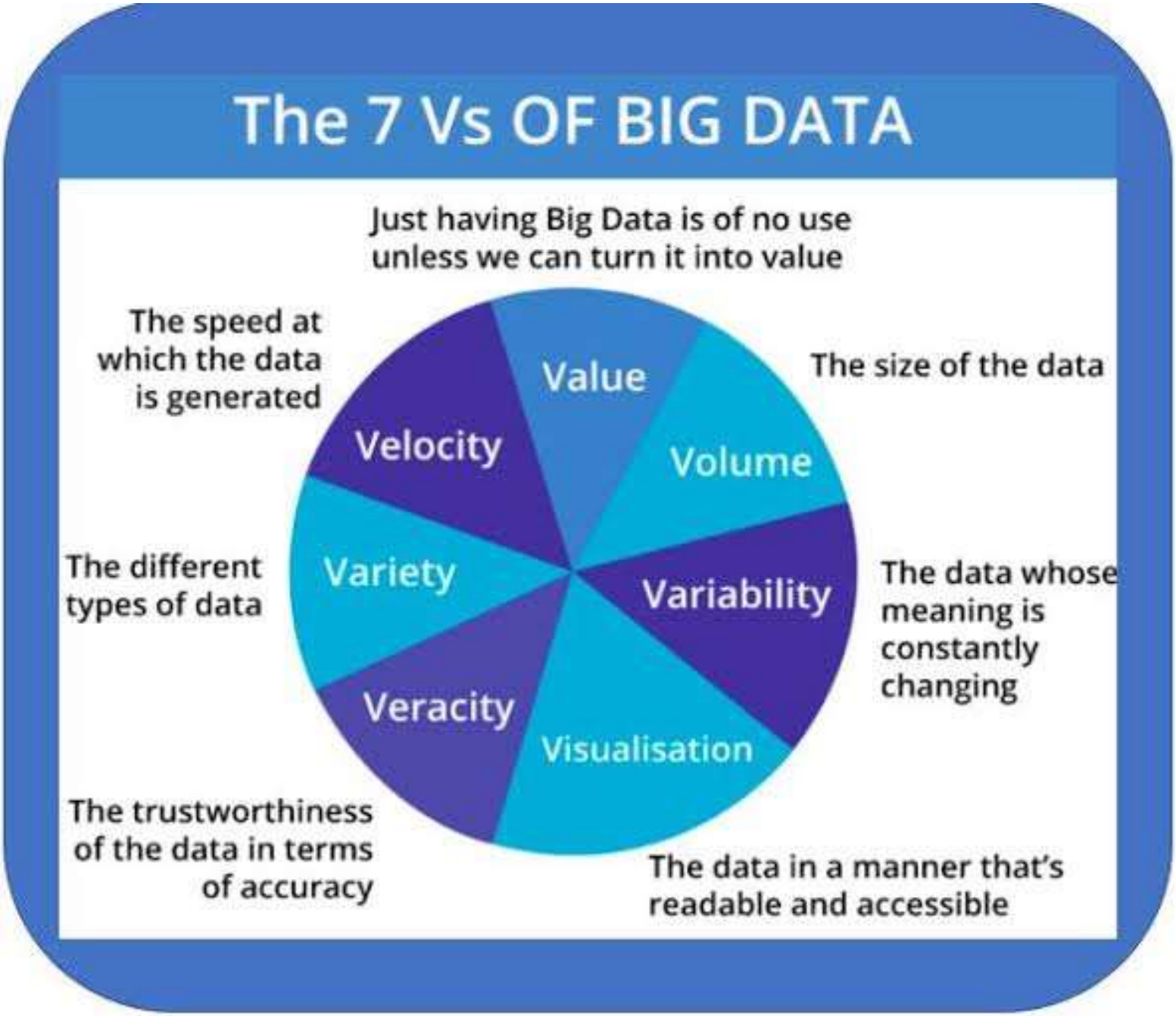
Courtesy: [3]

V3: V for Velocity

- Data is being generated fast and need to be processed fast
  - For time-sensitive processes such as catching fraud, big data must be used as it streams into your enterprise in order to maximize its value
  - Scrutinize 5 million trade events created each day to identify potential fraud
  - Analyze 500 million daily call detail records in real-time to predict customer churn faster
- Sometimes, 2 minutes is too late!
  - The latest we have heard is 10 ns (nano seconds) delay is too much

Courtesy: [3]

The 7 Vs OF BIG DATA



References

[1] https://www.slideshare.net/edjuma/data-vs-information-4

[2] https://www.exetercityfutures.com/qsteps-blog/

[3] Dr. Debasis Samanta, Data Analytics, Lecture #1.

[4] https://www.diffen.com/difference/Data_vs_Information

[5] https://bigdatapath.wordpress.com/2019/11/13/understanding-the-7-vs-of-big-data/