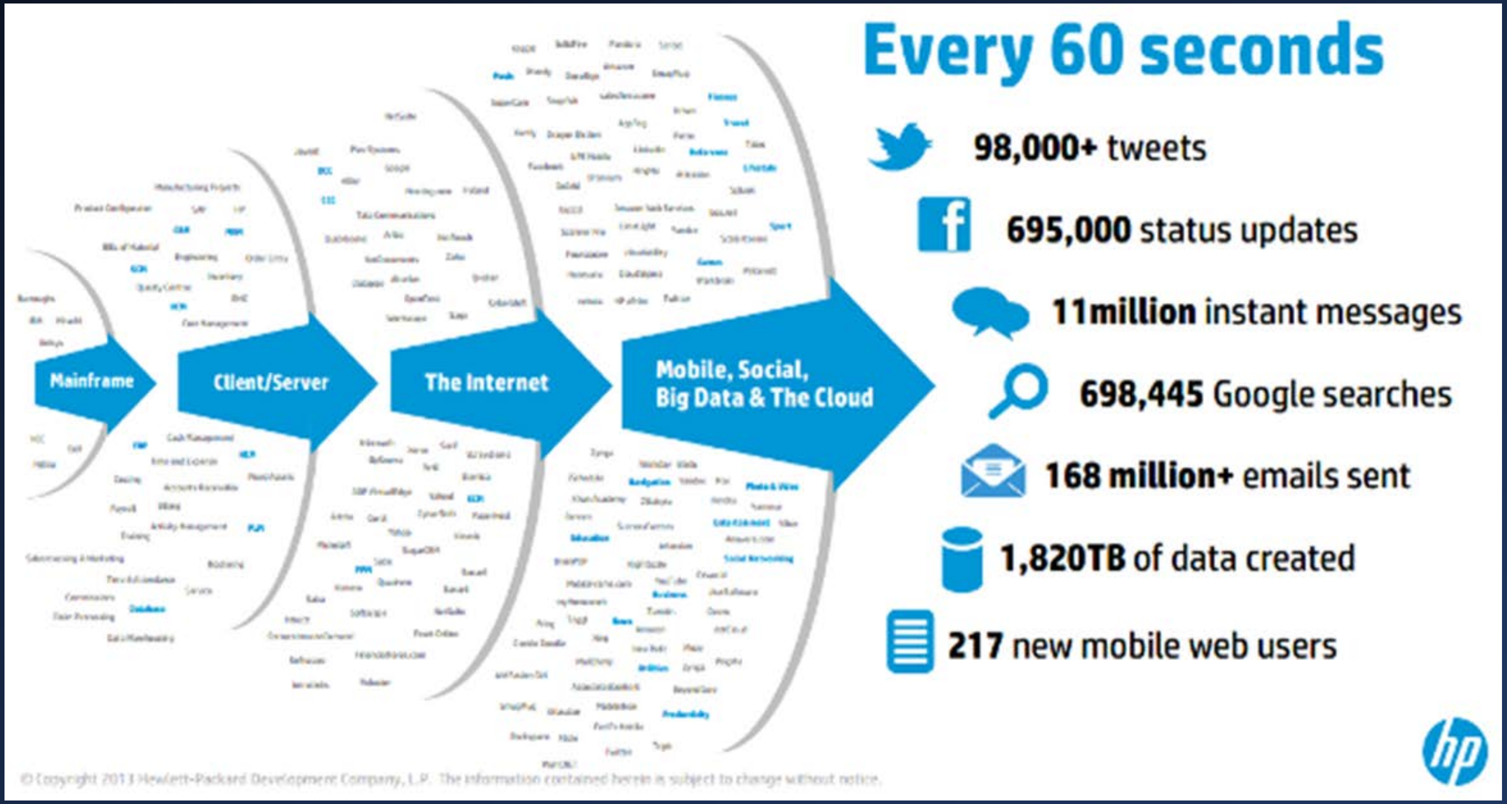
Big Data (MHI222956/MHI225101)

1.2 Big Data Introduction

The Era of Big Data



Every 60 seconds

- 98,000+ tweets
- 695,000 status updates
- 11million instant messages
- 698,445 Google searches
- 168 million+ emails sent
- 1,820TB of data created
- 217 new mobile web users

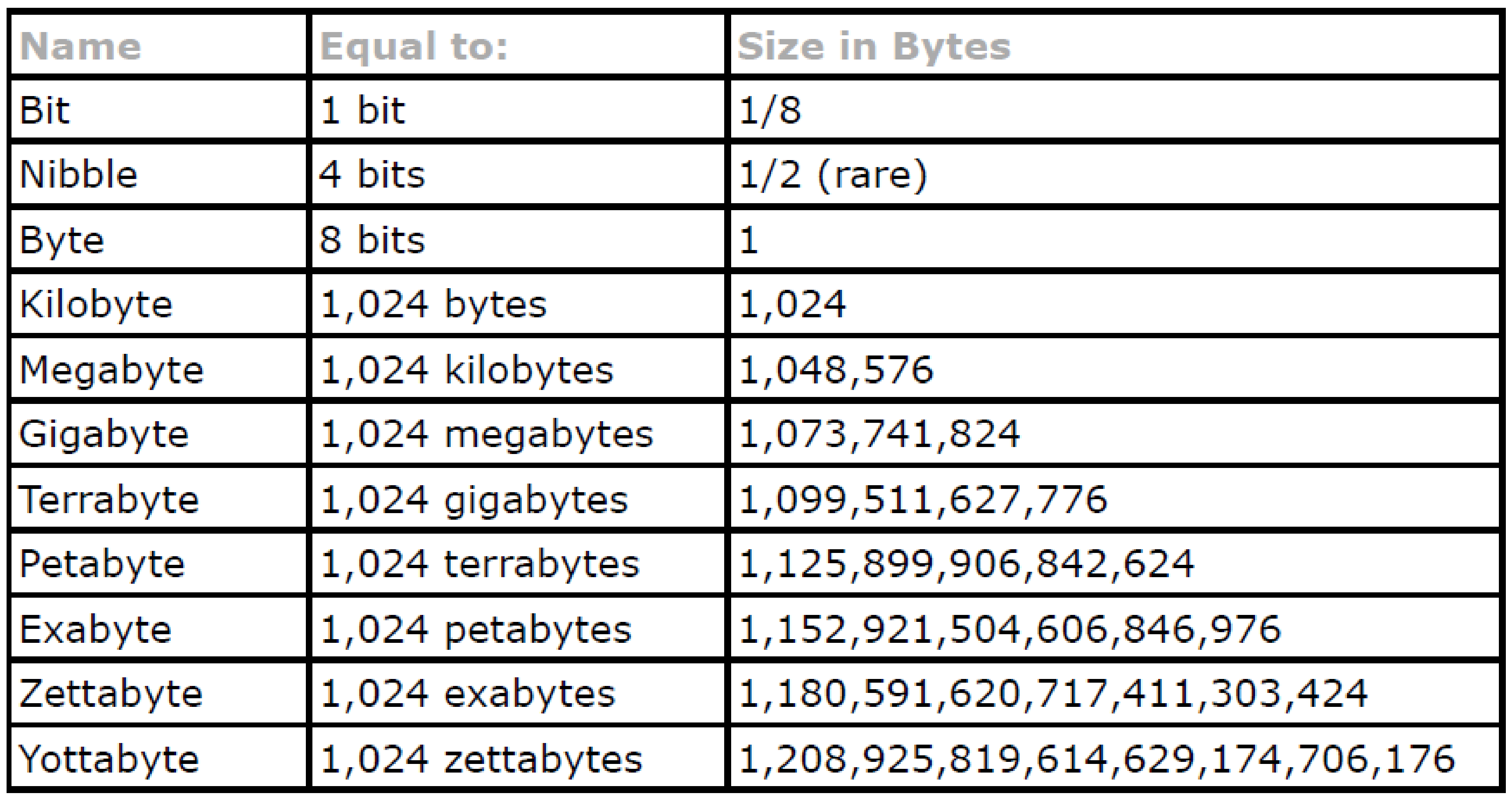
Further reading: Big data - Statistics & Facts

https://www.statista.com/topics/1464/big-data/

The data is getting generated exponentially day by day with the increasing usage of devices and digitization across the globe.

Every 2 days we create as much information as we did from the beginning of time until 2003.

A full 90 percent of all the data in the world has been generated over the last two years.



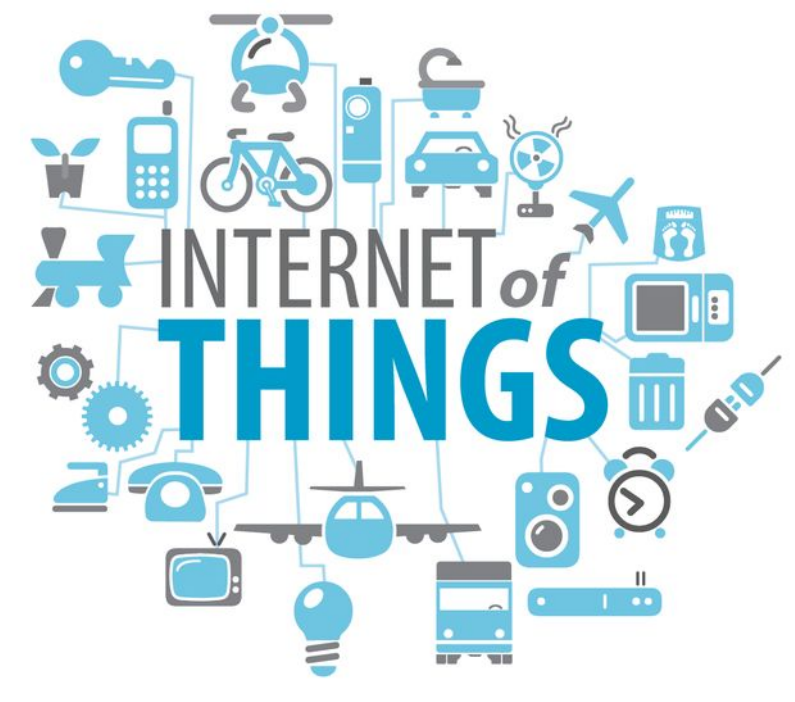
Name	Equal to:	Size in Bytes
Bit	1 bit	1/8
Nibble	4 bits	1/2 (rare)
Byte	8 bits	1
Kilobyte	1,024 bytes	1,024
Megabyte	1,024 kilobytes	1,048,576
Gigabyte	1,024 megabytes	1,073,741,824
Terrabyte	1,024 gigabytes	1,099,511,627,776
Petabyte	1,024 terrabytes	1,125,899,906,842,624
Exabyte	1,024 petabytes	1,152,921,504,606,846,976
Zettabyte	1,024 exabytes	1,180,591,620,717,411,303,424
Yottabyte	1,024 zettabytes	1,208,925,819,614,629,174,706,176

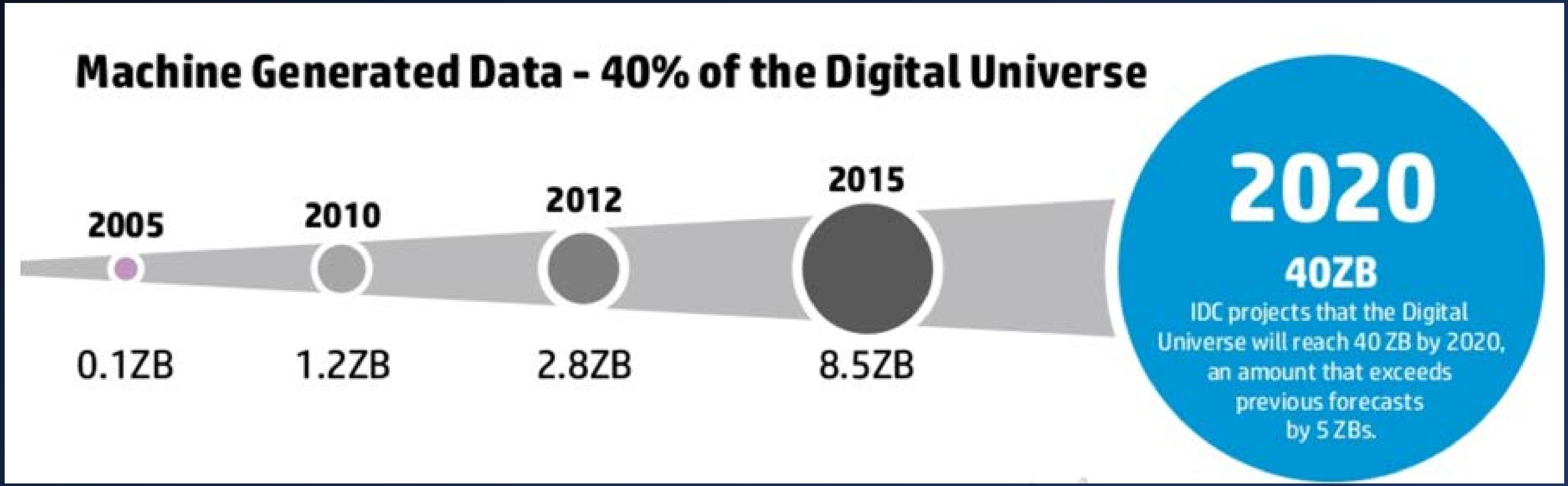
The amount of data available on the web in the year 2000 is thought to occupy 8 petabytes (theorized by Roy Williams)

Where do the data come from?

Machine-Generated Data

- It’s everywhere
- Machine to machine data
- The Internet of Things
  - Smart devices (sensing)
  - Interconnectivity

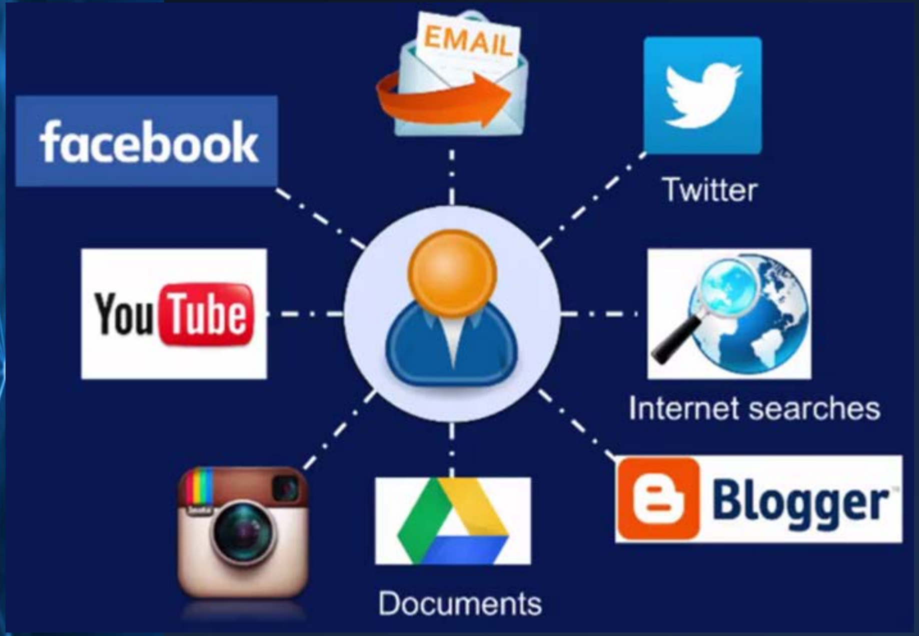




Machine Generated Data - 40% of the Digital Universe

IDC projects that the Digital Universe will reach 40 ZB by 2020, an amount that exceeds previous forecasts by 5 ZBs.

Big Data Generated by People

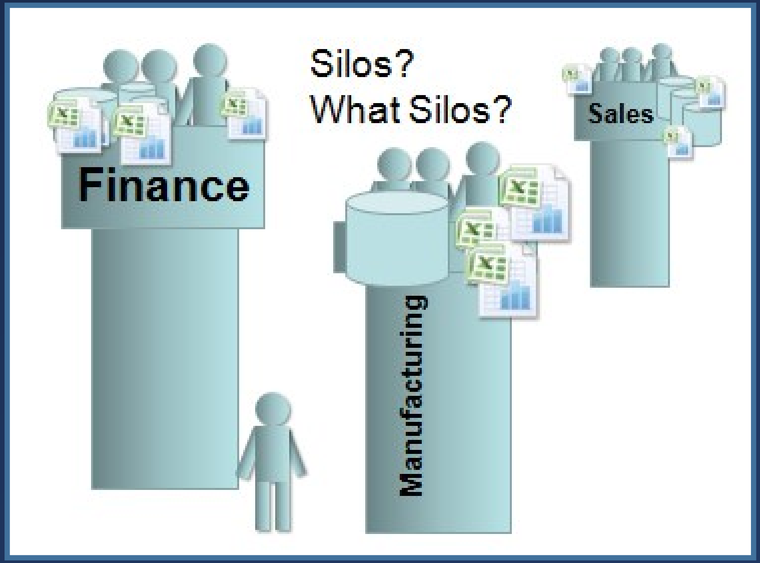


- People generated massive amounts of data every day through their actives
  - Social media network
  - Online photo sharing
  - Online video sharing
  - Blogging and Commenting
  - Email
  - etc.
- Most of this data is text-heavy and unstructured: not conforming to a predefined data model.

Organization-Generated Data

- Structured
  - Usually stored in relational database management system

- But often siloed
  - Captured at the department level
  - Without proper infrastructure and policy to share and integrate this data



Definitions of Big Data

- The first ACM digital library article which mentions Big Data was by NASA researchers Michael Cox and David Ellsworth in 1997, i.e. one year before John Mashey was credited as having coined the term. They used the term Big Data as follows:

> *"Visualization provides an interesting challenge for computer systems: data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk. We call this the problem of big data. When data sets do not fit in main memory (in core), or when they do not fit even on local disk, the most common solution is to acquire more resources."*

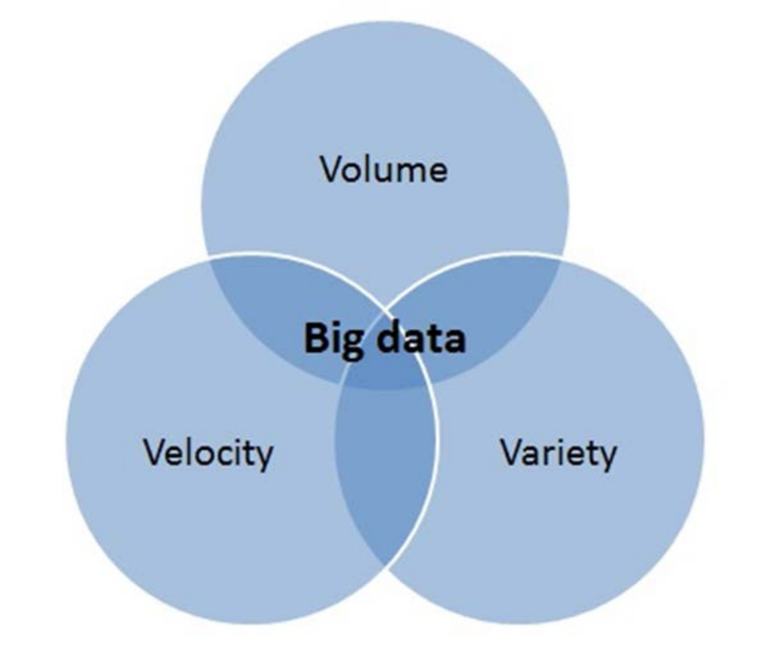
- Currently, there are a number of definitions of big data.

- Here is a definition from Google:

> *"Big data refers to data that would typically be too expensive to store, manage, and analyze using traditional (relational and/or monolithic) database systems. Usually, such systems are cost inefficient because of their inflexibility for storing unstructured data (such as images, text, and video), accommodating "high-velocity" (real-time) data, or scaling to support very large (petabyte-scale) data volumes."*

https://cloud.google.com/what-is-big-data

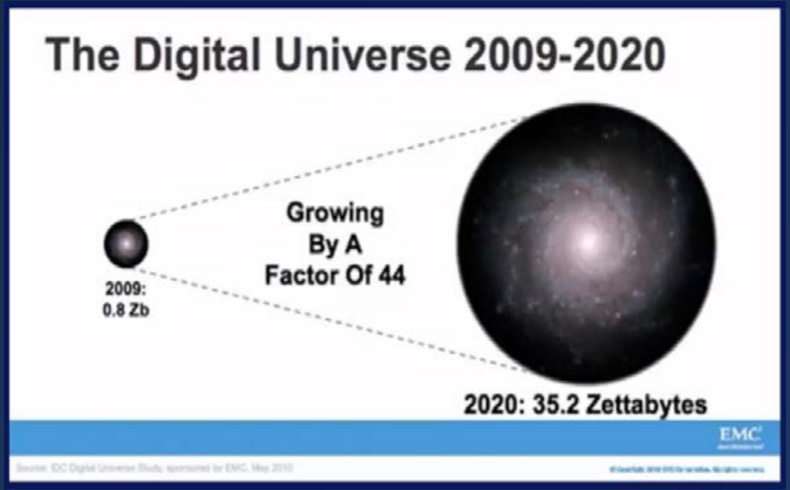
- Perhaps the most well-known version comes from IBM,
- They suggested that big data could be characterized by any or all of three “V” words:
  - volume
  - variety
  - velocity



Big Data - Volume

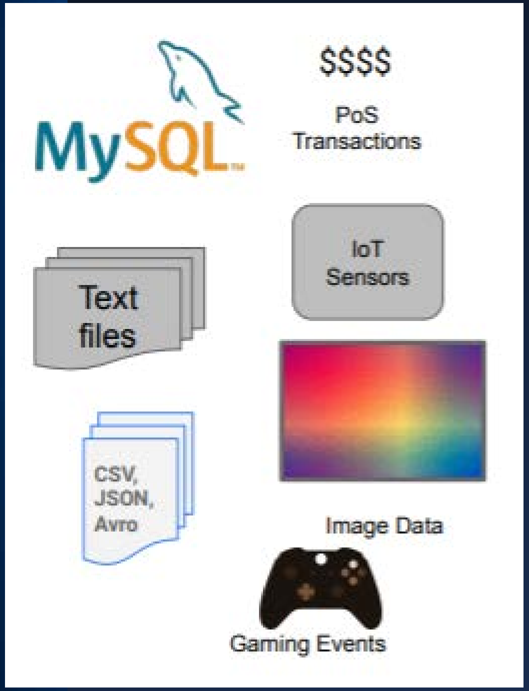
- Refers to the vast amounts of data that is generated every second, minutes, hour, and day in our digitized world

- A number of challenges related to the massive volumes of big data.
  - The amount of storage space required to store that data efficiently
  - To retrieve that large amount of data fast enough
  - To process the data and get results in a timely fashion



Source: IDC Digital Universe Study, sponsored by EMC, May 2010

Big Data - Variety



- Refers to the ever increasing different forms that data can come in such as text, images, voice, and geospatial data.

- Axes of Data Variety
  - Structural variety – formats and models
  - Media variety – medium in which data get delivered
  - Semantic variety – how to interpret and operate on data
  - Availability variations – real-time? Intermittent?

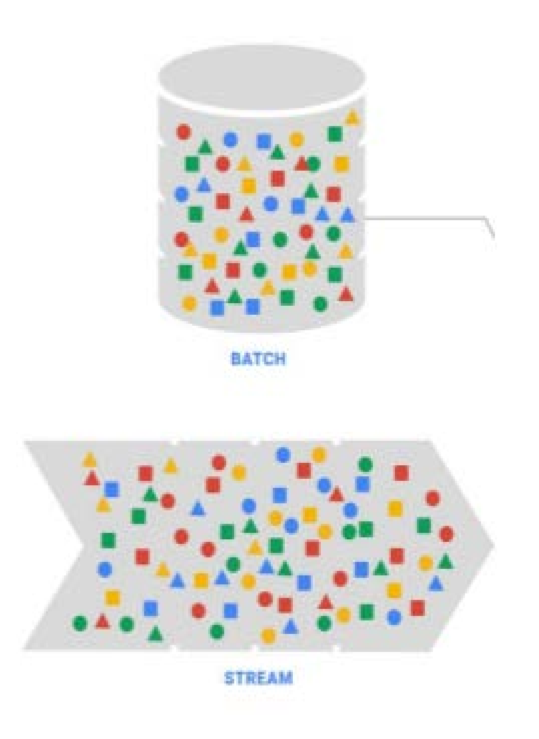
Big Data - Velocity

- Refers to the speed at which data is being generated and the increasing speed at which the data needs to be stored and analysed.

- Processing of data in real-time to match its production rate as it gets generated is a particular goal of big data analytics.

  e.g. personalization of advertisement on the web pages

- Real-time processing and Batch processing



More Vs for Big Data:

- Veracity
- Valence
- Value
- …

Big Data - Veracity

- Refers to the quality of the data, i.e., biases, noise, and abnormality in data.

- Data is of no value if it's not accurate, the results of big data analysis are only as good as the data being analysed.

- Quality can be defined as a function of a couple of different variables:
    - Accuracy of the data
    - The trustworthiness or reliability of the data source
    - Context within analysis

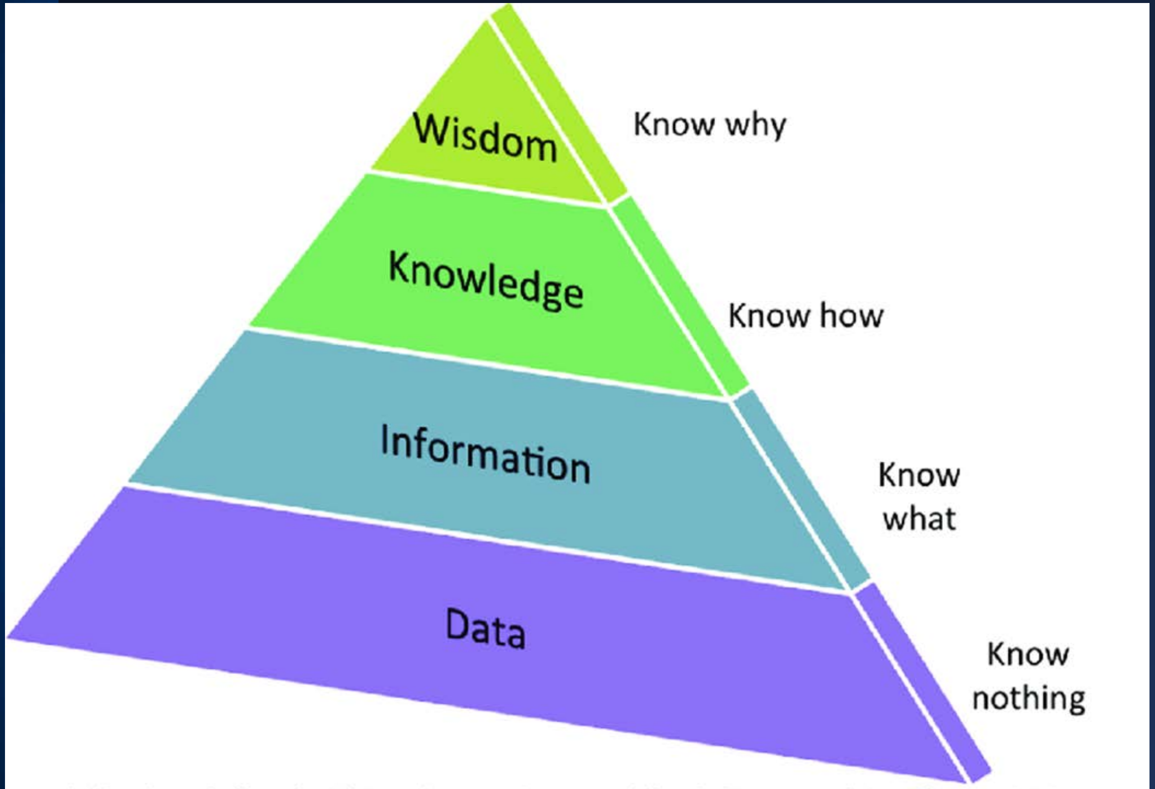
Big Data - Valence

- Refers to the connectendness of big data in forms of graphs.

- Measure of connectivity
  - Data Connectivity: Two data items are connected when they are related to each other
  - Valence: Fraction of data items that are connected out of the total number of possible connections

- Valence increase over time, makes the data connections denser

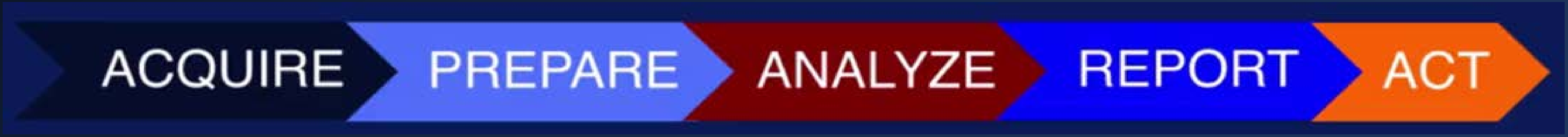
Big Data - Value



Compiled based on Ackoff, R. L. (1989). From data to wisdom. *Journal of Applied Systems Analysis*, *16*(1), 3-9. and Zeleny, M. (1987). Management support systems: towards integrated knowledge management. *Human Systems Management*, *7*(1), 59-70.

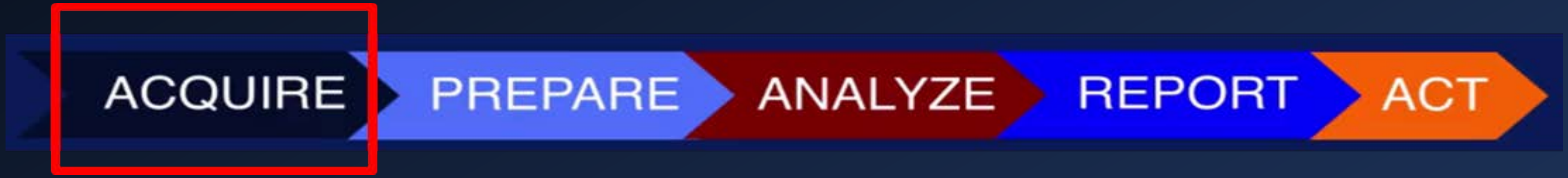
- Getting the value out of big data

A simple linear form of data science process:



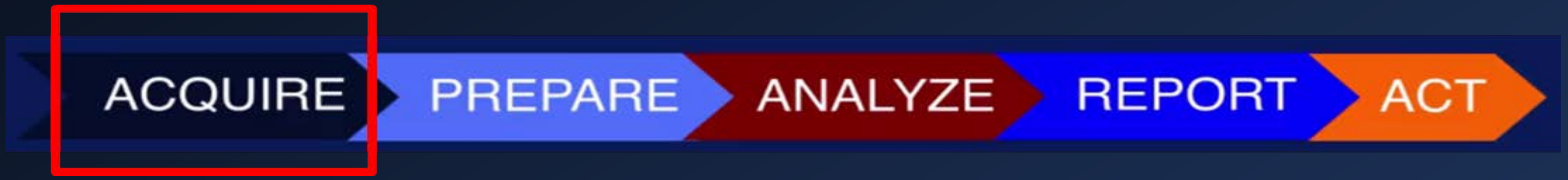
ACQUIRE → PREPARE → ANALYZE → REPORT → ACT

Acquiring Data



- The first step in acquiring data is to determine what data is available: finding the right data sources.
- Data, comes from, many places, local and remote, in many varieties, structured and un-structured. And, with different velocities.

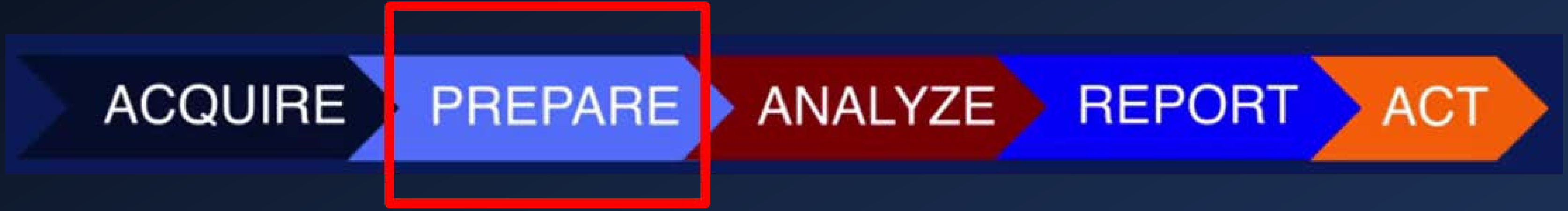
Acquiring Data



ACQUIRE PREPARE ANALYZE REPORT ACT

- There are many techniques and technologies to access these different types of data.
    - For data exists in conventional relational databases, the tool of choice to access data from databases is structured query language or SQL.
    - Data can also exist in files such as text files and Excel spreadsheets. Scripting languages (e.g., Java Script, Python, R) are generally used to get data from files.
    - An increasingly popular way to get data is from websites. Many websites host web services (e.g. REST) which produce program access to their data.
- NoSQL storage systems are increasingly used to manage a variety of data types in big data.

Preparing Data - data exploring & pre-processing



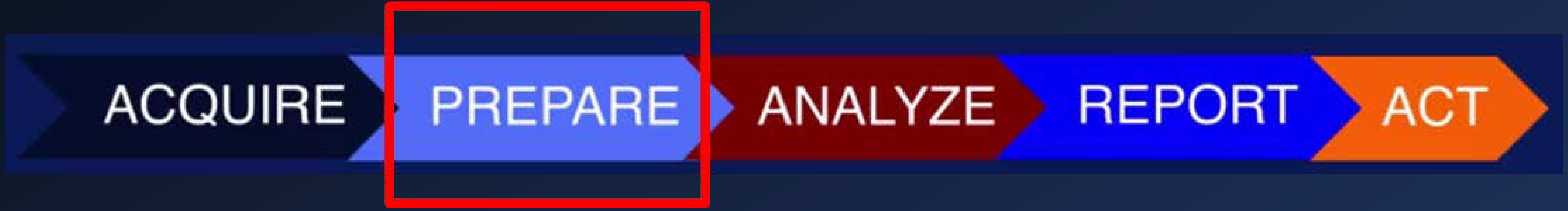
- Exploring data is the first part of the data preparation process.
- Techniques for data exploring:
  - Preliminary investigation

    Correlations, General trends, Outliers
  - Summary statistics

    Mean, Median, Range, standard deviation
  - Visualization techniques

    Histogram, Scatter plots

Preparing Data - data exploring & pre-processing



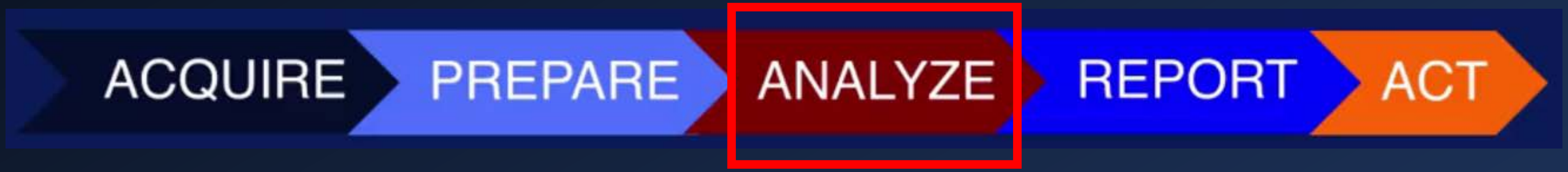
ACQUIRE PREPARE ANALYZE REPORT ACT

- Two major purposes of data pre-processing
  - To clean the data to address data quality issues

    e.g., Missing values, Invalid data, outliers

  - To transform the raw data to make it suitable for analysis
    - Scaling
    - Transformation
    - Feature selection
    - Dimensionality reduction
    - Data manipulation

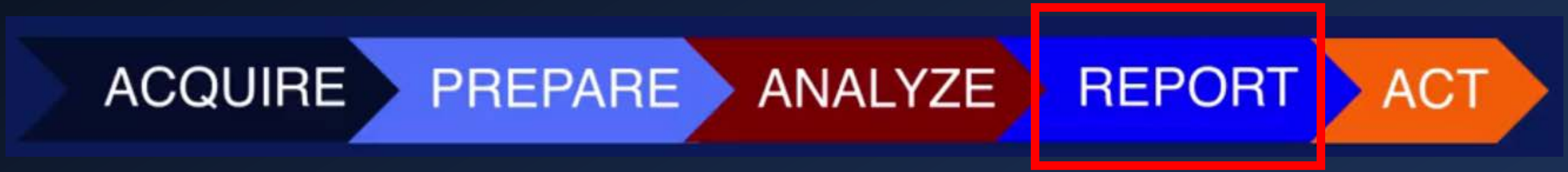
Analysing Data



ACQUIRE → PREPARE → ANALYZE → REPORT → ACT

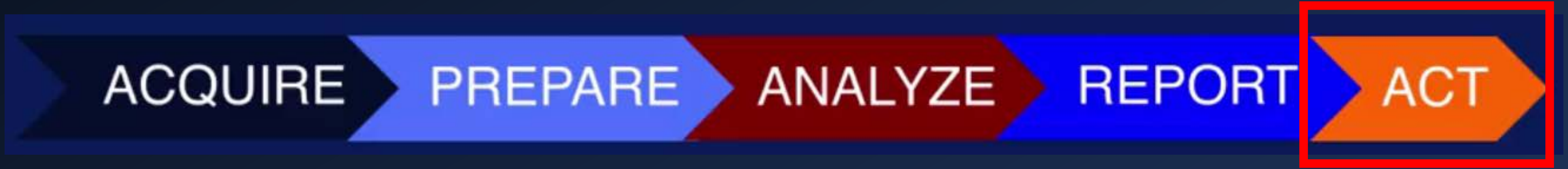
- Building a model from your data!
- Different types of analysis techniques
  - Classification
  - Regression
  - Clustering
  - Association analysis
  - Graph analysis

Communication results



- To report the insights gained from your analysis
- All findings must be presented so that informed decisions can be made.
- Visualization is an important tool in presenting your results.
  - Python
  - R
  - D3
  - Leaflet
  - Tableau

Turn insights into action



- To determine what action or actions should be taken, based on the insights gained
- Is there additional analysis that need to be performed in order to yield even better results?
- What data should be revisited?

Summary

- The Era of Big Data
- Where do big data come from?
- Definitions of Big Data
- Characteristics of Big Data
- Steps in Data Science Process