SNS COLLEGE OF ENGINEERING

Kurumbapalayam (Po), Coimbatore – 641 107

An Autonomous Institution

Accredited by NBA – AICTE and Accredited by NAAC – UGC with ‘A’ Grade
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

DEPARTMENT OF COMPUTER SCIENCE AND TECHNOLOGY

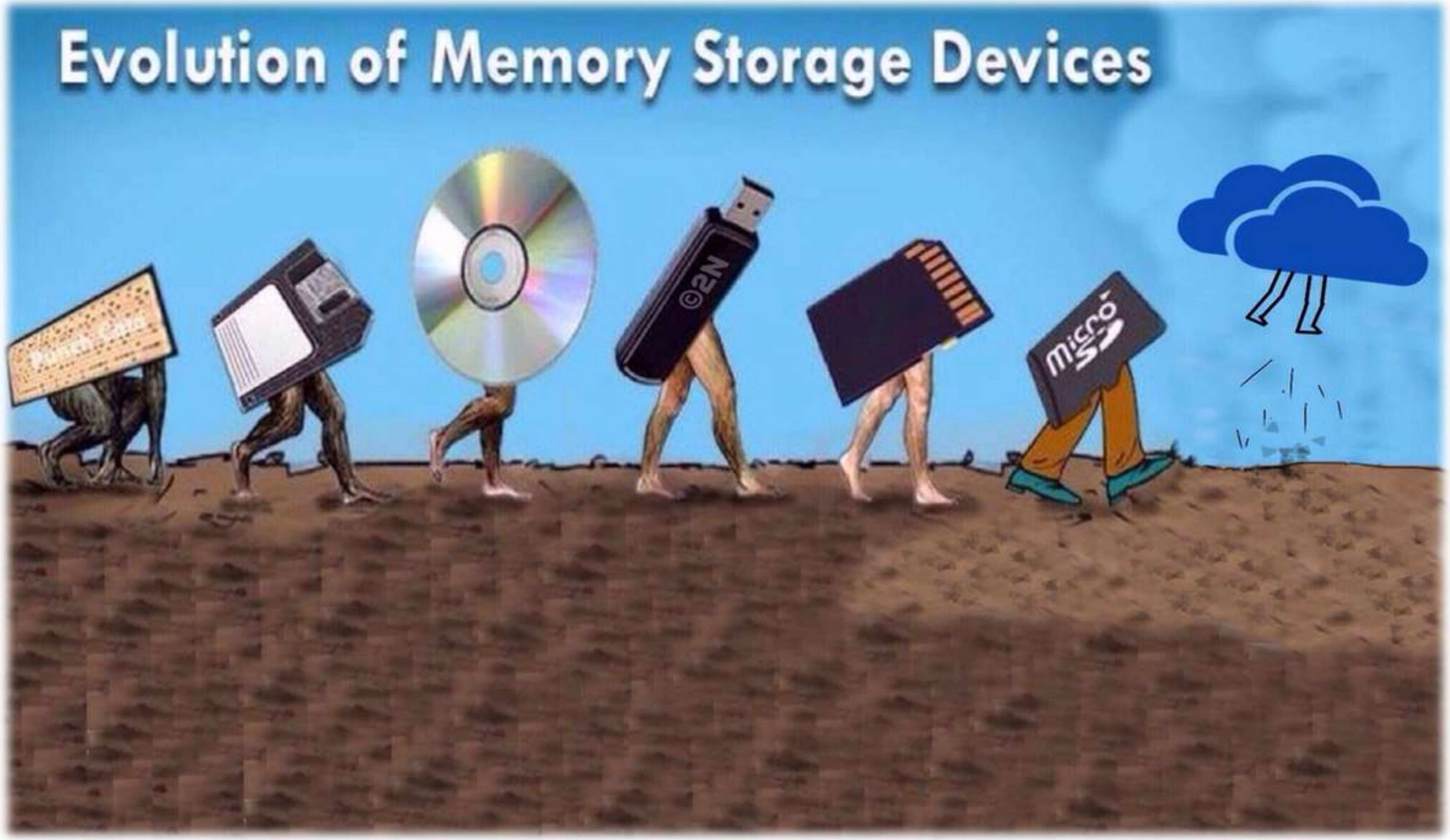
COURSE NAME :19CS407 DATA ANALYTICS WITH R

II YEAR /IV SEMESTER

Unit 1- Introduction

Topic : Big data , Data Science

Big Data



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

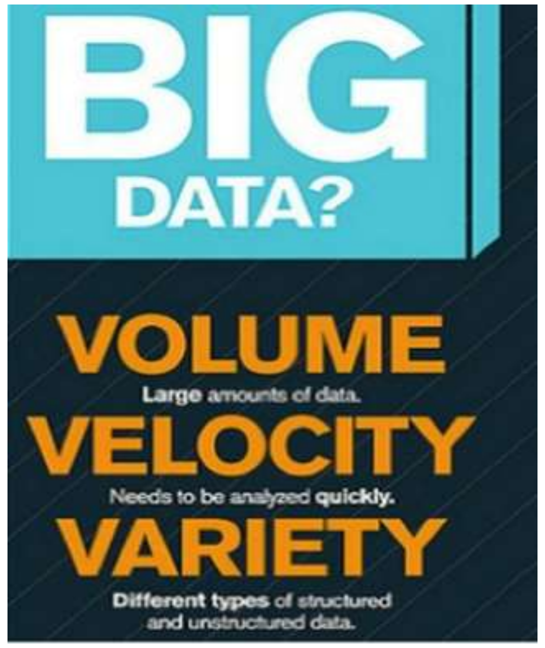
  - techniques, tools and architectures

  …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…

Characteristics of Big data: **V3**

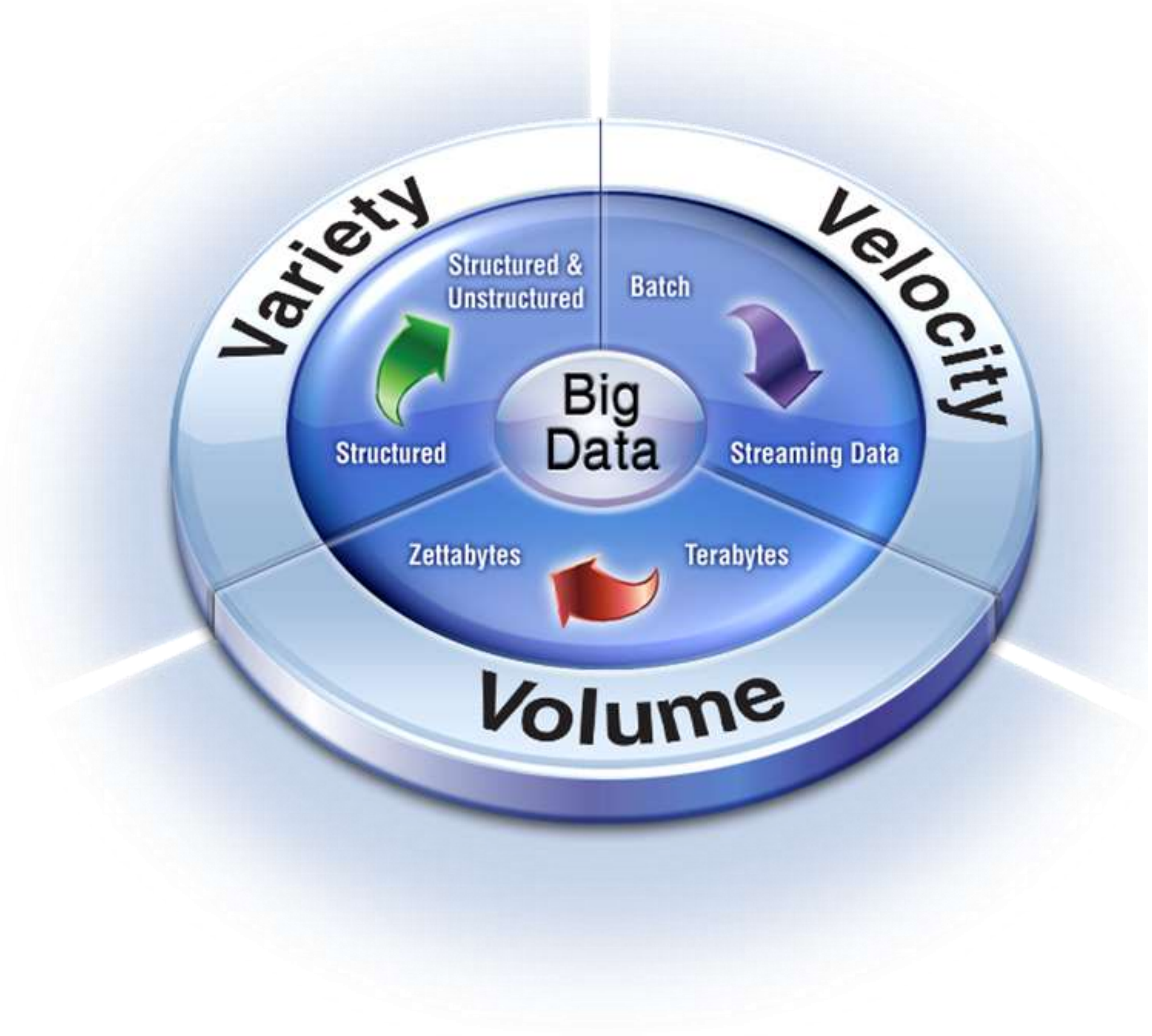


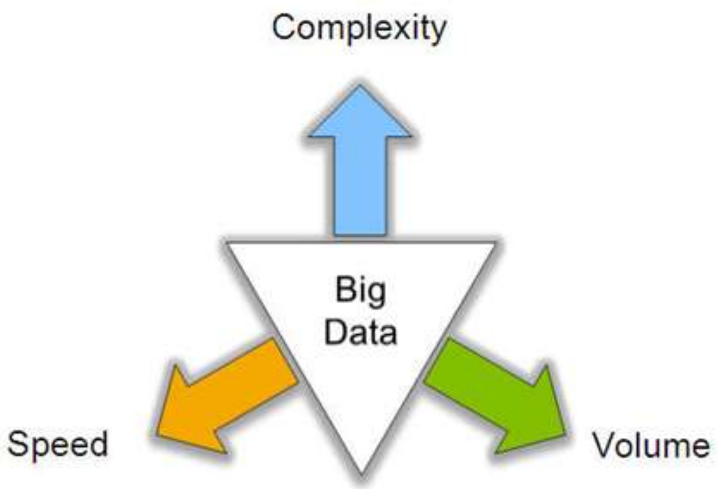
BIG DATA?

VOLUME
Large amounts of data.

VELOCITY
Needs to be analyzed quickly.

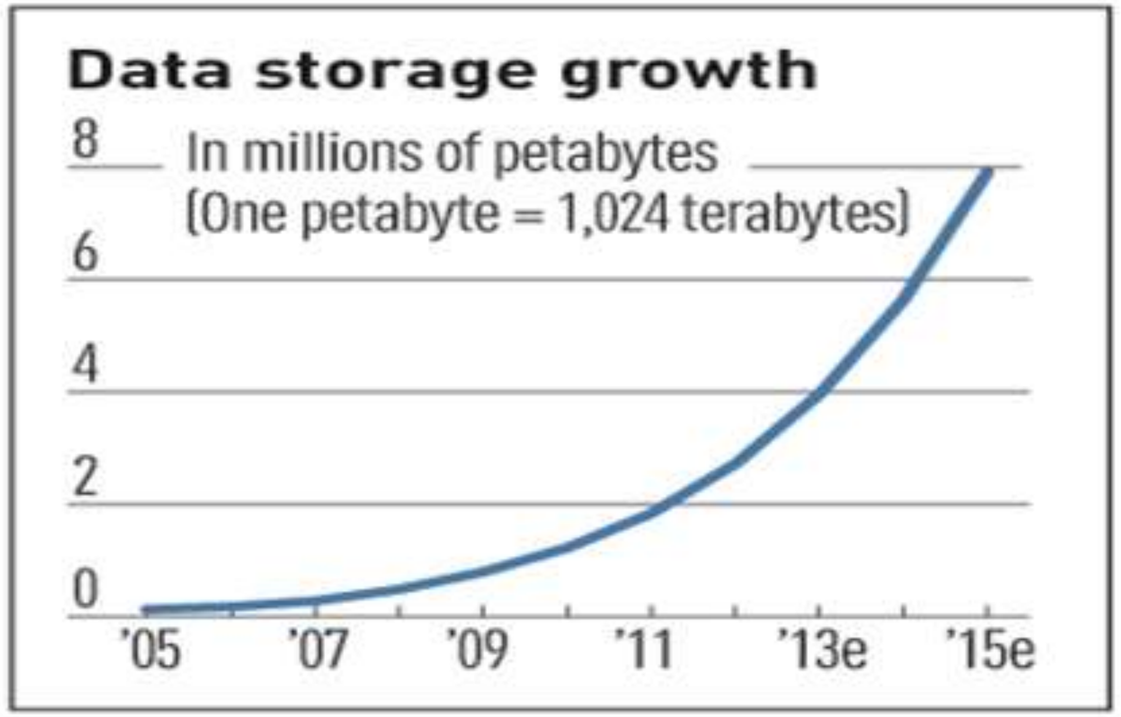
VARIETY
Different types of structured and unstructured data.

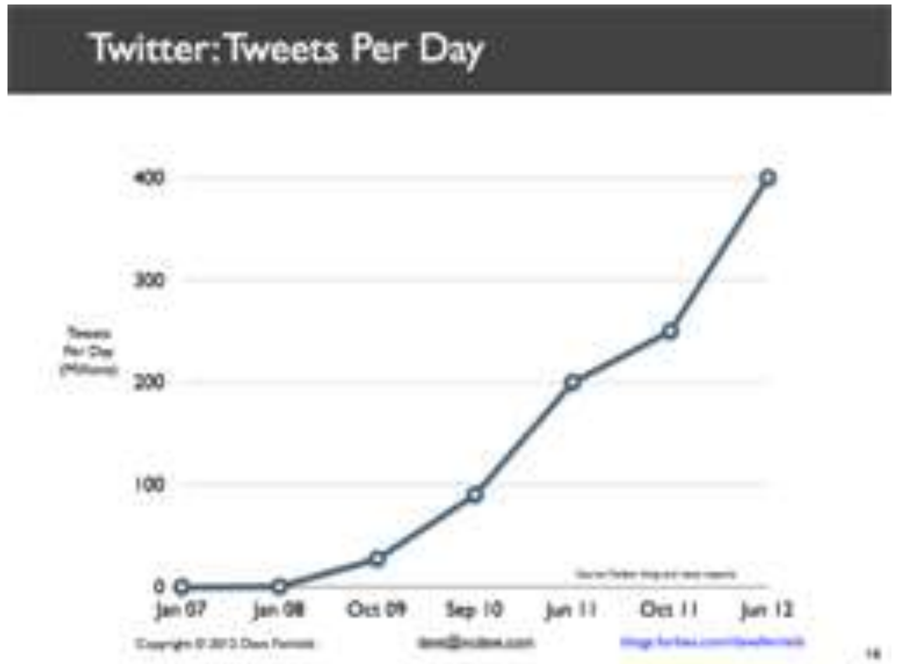




V3 : V for Volume

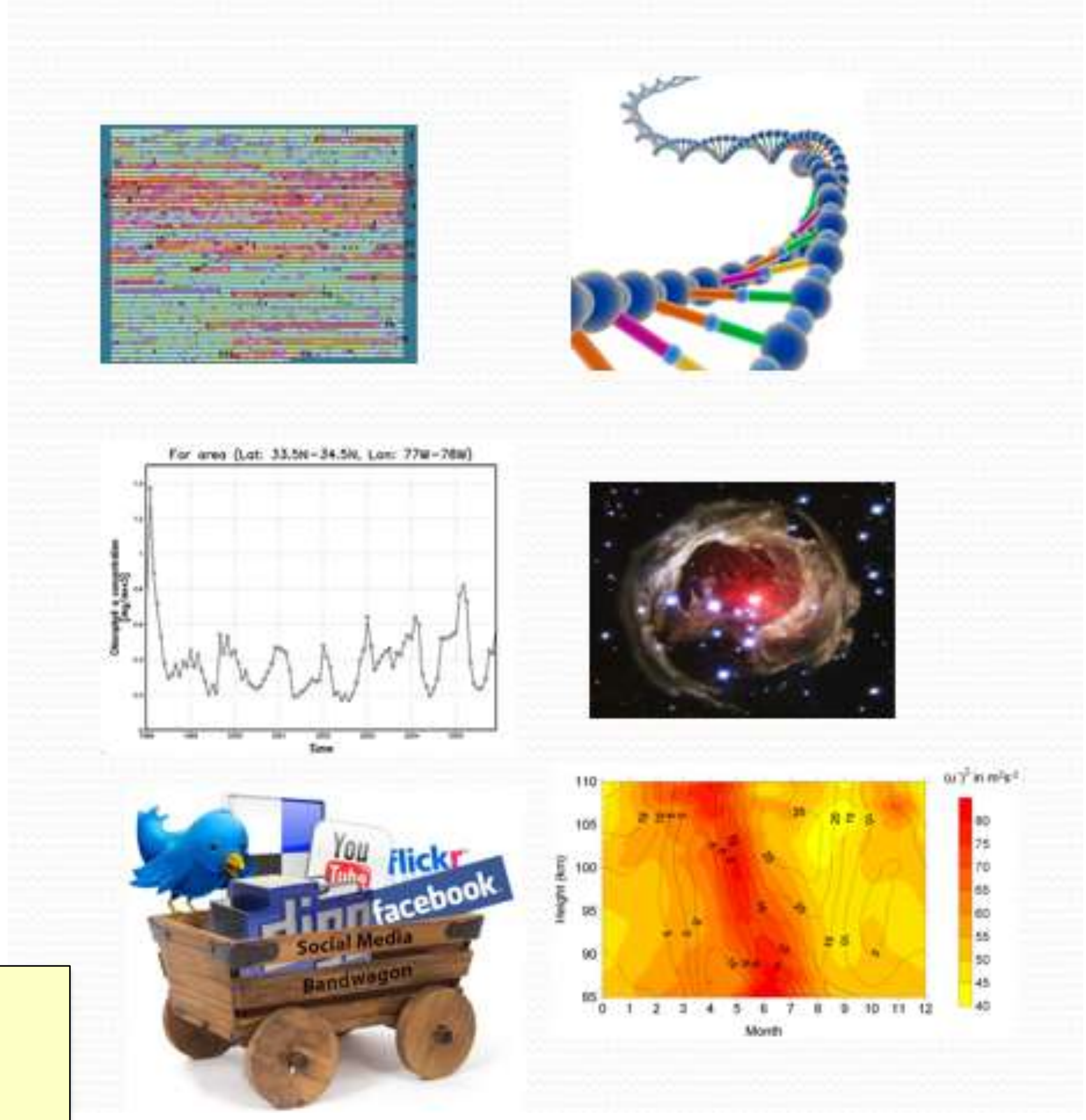
- Volume of data, which needs to be processed is increasing rapidly
  - More storage capacity
  - More computation
  - More tools and techniques





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

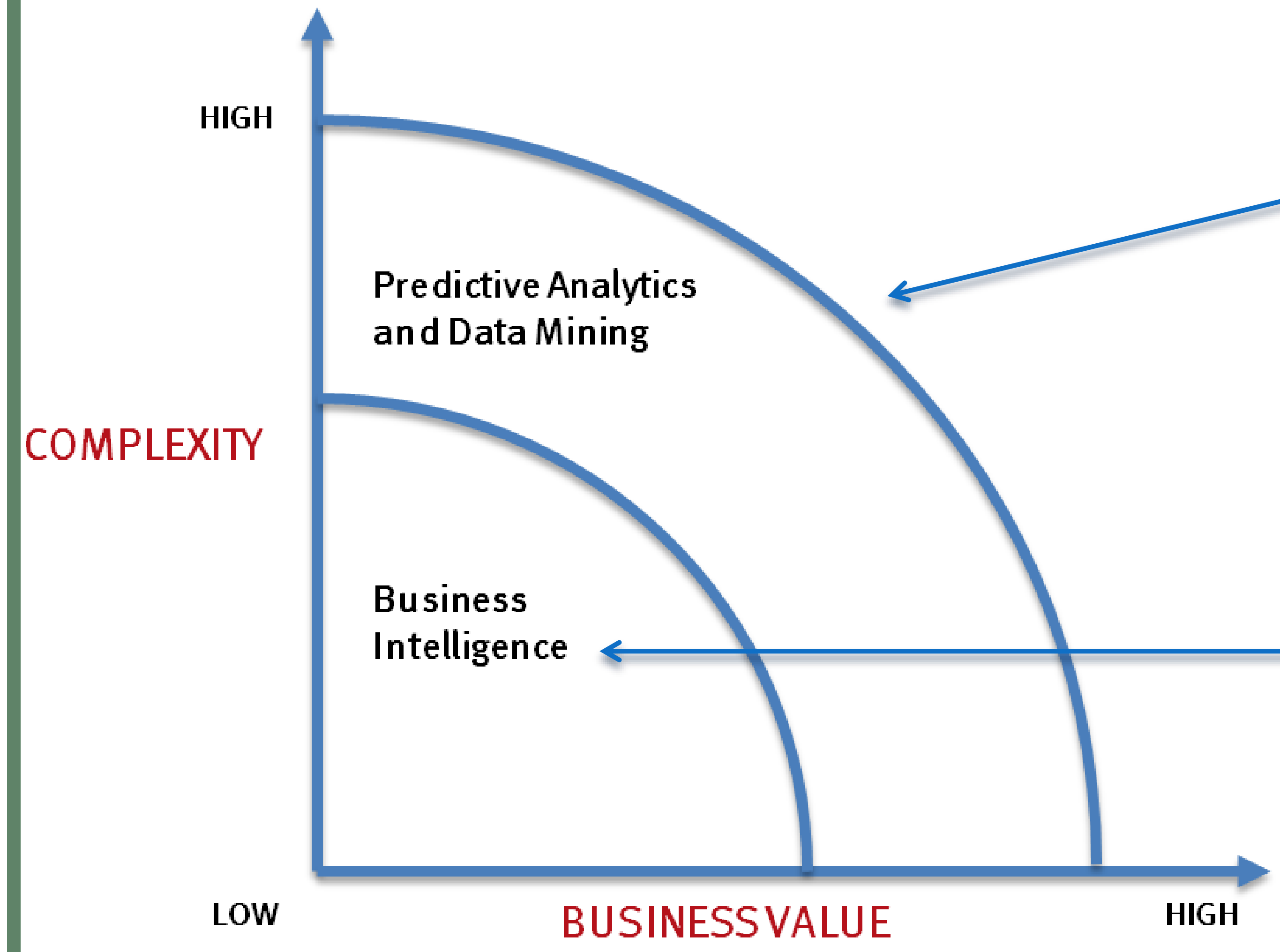


To extract knowledge➔ all these types of data need to be linked together

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

Big data vs. small data

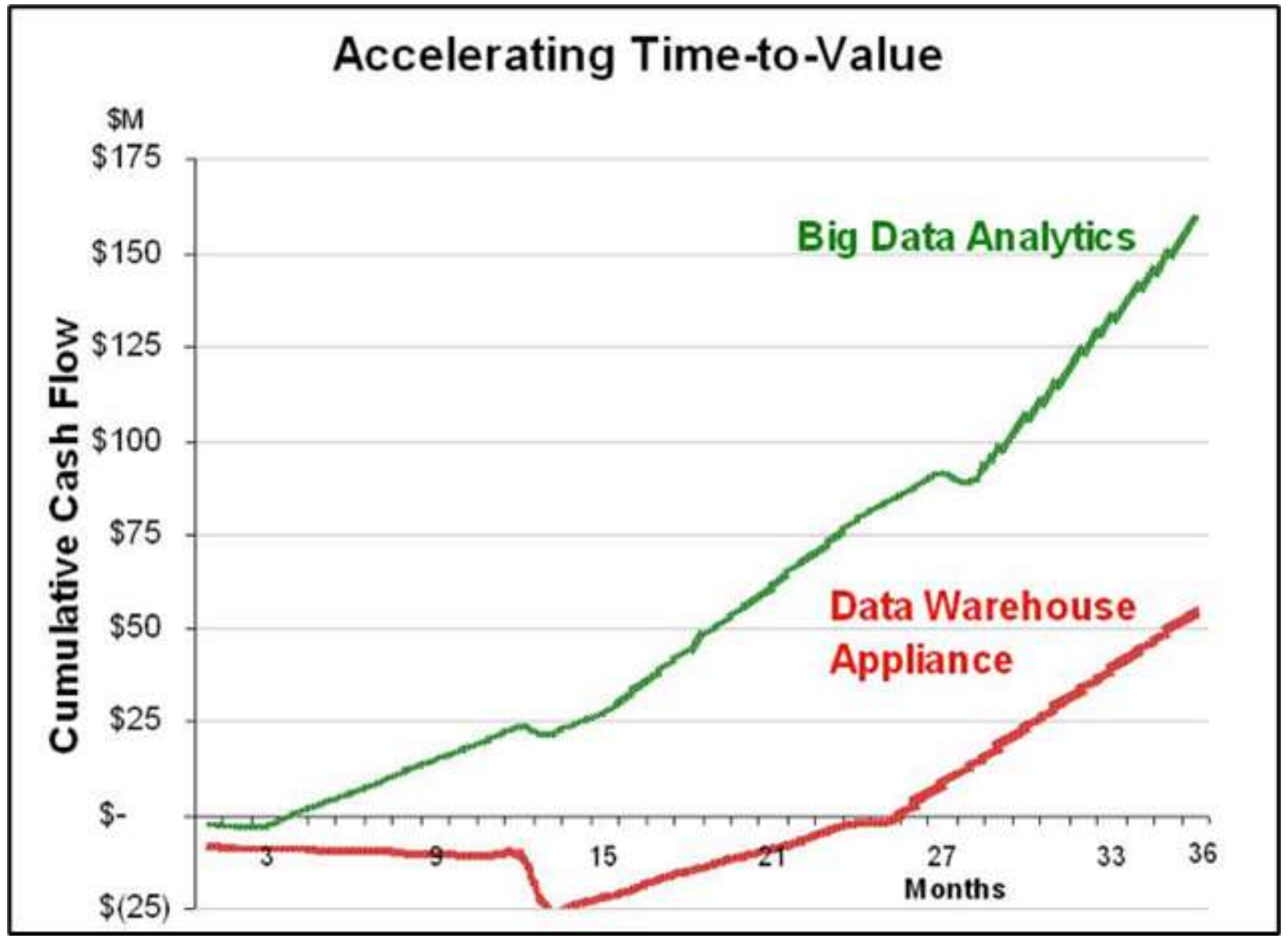


- Optimizations and predictive analytics
- Complex statistical analysis
- All types of data, and many sources
- Very large datasets
- More of a real-time

- Ad-hoc querying and reporting
- Data mining techniques
- Structured data, typical sources
- Small to mid-size datasets

Big Data Vs Small Data

- Big data is more real-time in nature than traditional applications
- Big data architecture
  - Traditional architectures are not well-suited for big data applications (e.g. Exa-data, Tera-data)
  - Massively parallel processing, scale out architectures are well-suited for big data applications



Challenges Ahead

- **The Bottleneck is in technology**
  - New architecture, algorithms, techniques are needed
- **Also in technical skills**
  - Experts in using the new technology and dealing with Big data

Who are the major players in the world of Big data?

Major players

- Google
- Hadoop
- MapReduce
- Mahout
- Apache Hbase
- Cassandra

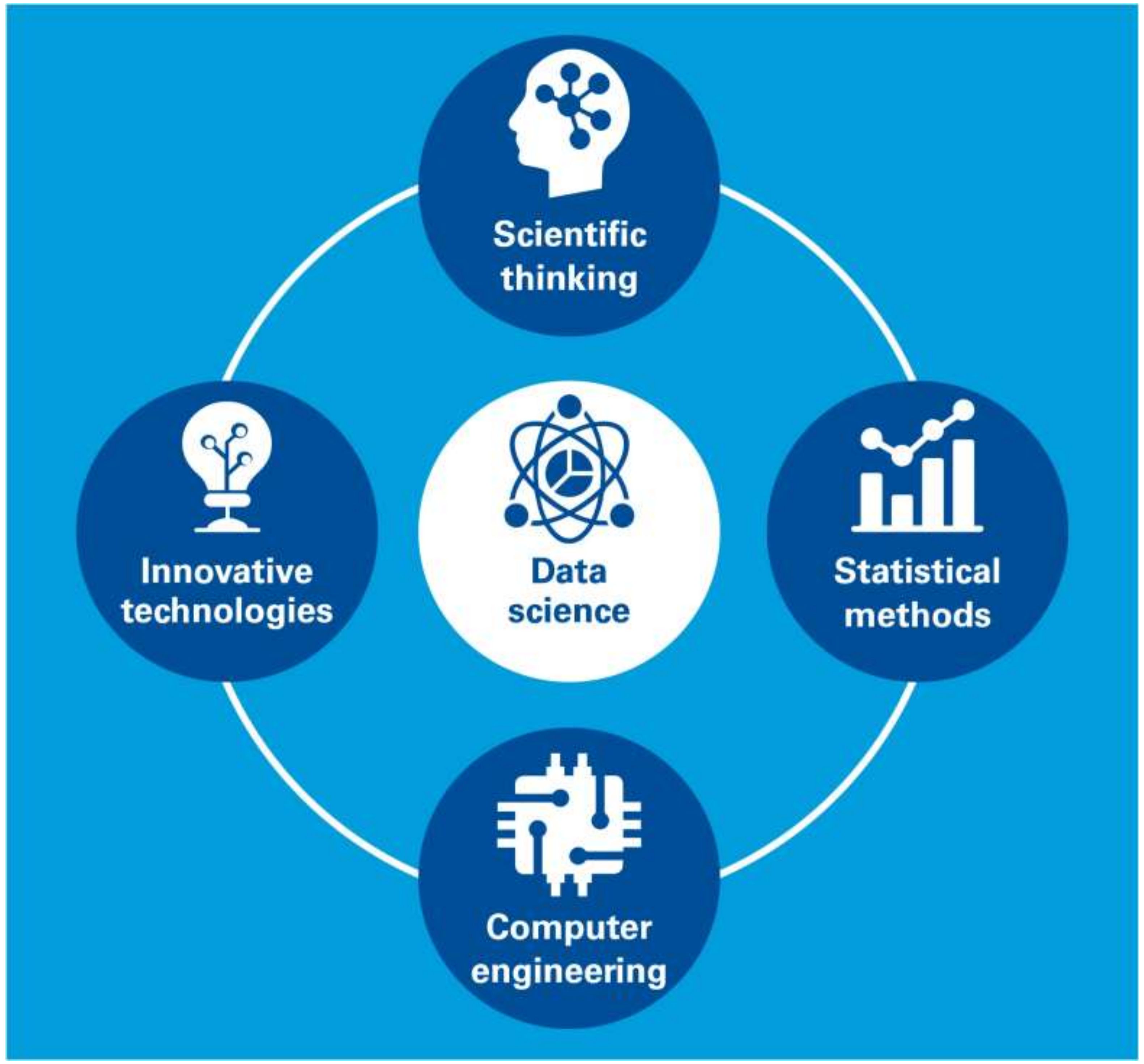
Applications of Big Data

- ✓ Media and entertainment
- ✓ Banking and securities
- ✓ Healthcare
- ✓ Education
- ✓ Energy sectors
- ✓ Retail and wholesale services
- ✓ Government sectors
- ✓ Insurance
- ✓ Cyber security
- ✓ Weather forecasting
- ✓ Travel and tourism sectors
- ✓ Scientific research

Tools available

- **NoSQL**
  - DatabasesMongoDB, CouchDB, Cassandra, Redis, BigTable, Hbase, Hypertable, Voldemort, Riak, ZooKeeper
- **MapReduce**
  - Hadoop, Hive, Pig, Cascading, Cascalog, mrjob, Caffeine, S4, MapR, Acunu, Flume, Kafka, Azkaban, Oozie, Greenplum
- **Storage**
  - S3, HDFS, GDFS
- **Servers**
  - EC2, Google App Engine, Elastic, Beanstalk, Heroku
- **Processing**
  - R, Yahoo! Pipes, Mechanical Turk, Solr/Lucene, ElasticSearch, Datameer, BigSheets, Tinkerpop

Data Science



Data science

Data science is the combination of statistics, mathematics, programming, problem-solving, capturing data in ingenious ways, the ability to look at things differently, and the activity of cleansing, preparing, and aligning data.

Data Science is a multi-disciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data

Overview of the five steps

1. Asking an interesting question
2. Obtaining the data
3. Exploring the data
4. Modeling the data
5. Communicating and visualizing the results

Applications of Data Science

- ✓ Detection of risk in business
- ✓ Healthcare
- ✓ Targeted advertising
- ✓ Internet behavior and searches
- ✓ Advanced Image and voice recognition
- ✓ Gaming

Assessment 1

1. What is Big Data?

Ans : ____

2. What is Data Science?

Ans : ____

References

1. J. E. Hopcroft, J.Motwani and J.D Ullman, “Introduction to Automata Theory, Languages and Computations”, Second Edition, Pearson Education, 2003.

Thank You