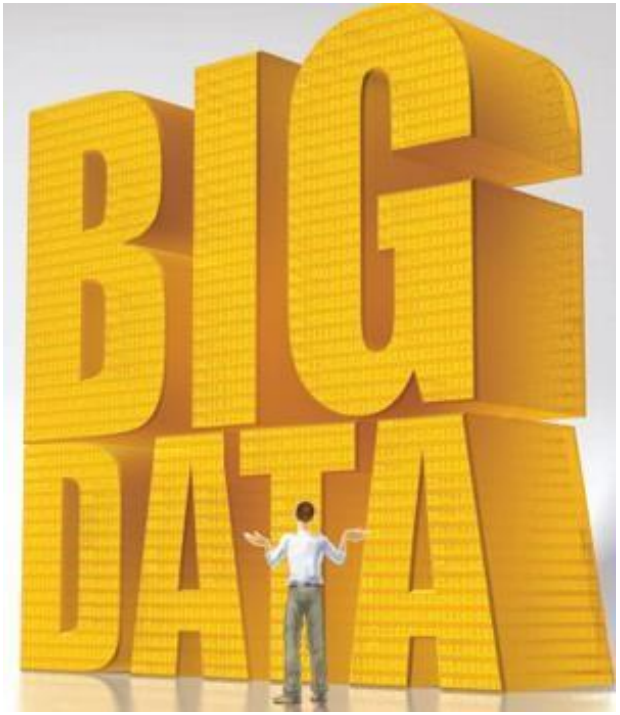

BIG
DATA

Outline

- Big Data
  - Structured and Unstructured Data
  - What is Big Data?
  - Big Data Ecosystem and Life Cycle
  - Industries using Big Data
  - Big Data Applications
  - Challenges in Big Data
- Hadoop
  - Hadoop Components
  - How Hadoop Works?
  - Hadoop Distributed File System (HDFS)
  - Hadoop MapReduce
  - Hadoop1 vs Hadoop2

Simple to start

- What is the maximum file size you have dealt so far?
  - Files
  - Movies
  - Streaming video

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}

Structured & Unstructured Data

- Structured Data
  - Any data that resides in a fixed field within a record or a file.
  - Includes data contained in relational databases and spreadsheets.
- Unstructured Data
  - Information that doesn’t reside in a traditional row-column database.
  - Includes text and multimedia content like e-mail, word processing documents, videos, audios, ppts, webpages, etc.

Structured & Unstructured Data

Data Management

Structured	Unstructured
SQL - a programming language created for managing and querying data in relational database management systems.	Big Data tools – Hadoop
	Business Intelligence Software
	Data Integration tools
	Document Management Systems
	Information Management solutions
	Search & Indexing tools

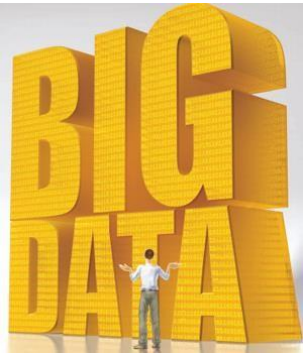
Semi-Structured Data

- Information that doesn't reside in a relational database but that does have some organizational properties that make it easier to analyze.
- Examples of semi-structured data might include XML documents and NoSQL databases.

What is Big Data?

- Every day, we create 2.5 quintillion bytes of data and 90% of the data in the world today has been created in the recent years alone.
- This data comes from everywhere:
  - sensors used to gather climate information,
  - posts to social media sites,
  - digital pictures and videos,
  - purchase transaction records, and
  - cell phone GPS signals
  - and many more ……

This data is



Big Data – A Growing Torrent

- There are huge volumes of data in the world:
  - From the beginning of recorded time until 2003, we created 5 billion gigabytes (exabytes) of data.
  - In 2011, the same amount was created every two days.
  - In 2014, the same amount of data was created every 10 minutes.

Why is “Big Data” a “Big Deal”?

12+ TBs **of tweet data every day**

? TBs **of data every day**

25+ TBs **of log data every day**

40 billion **RFID tags today (1.3B in 2005)**

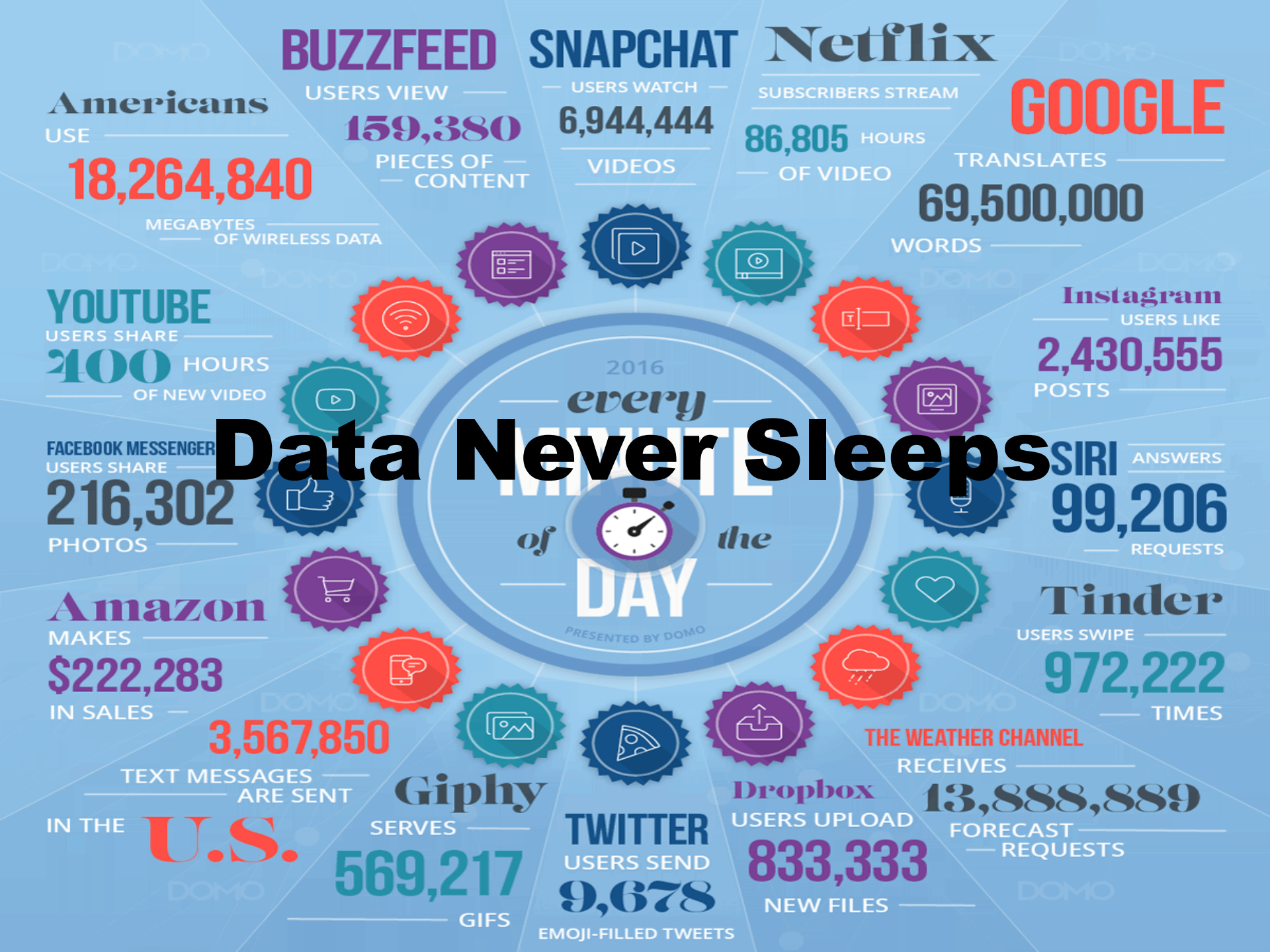
5.5 billion + **camera phones world wide**

100s of millions of GPS enabled **devices sold annually**

3.5 billion + **people on the Web by end 2016**

76 million **smart meters in 2009…**

200M + **Now**

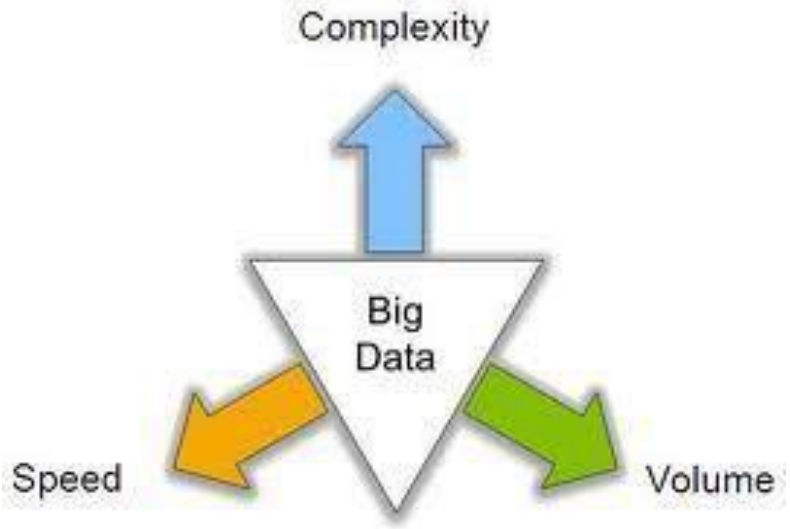


Data Never Sleeps

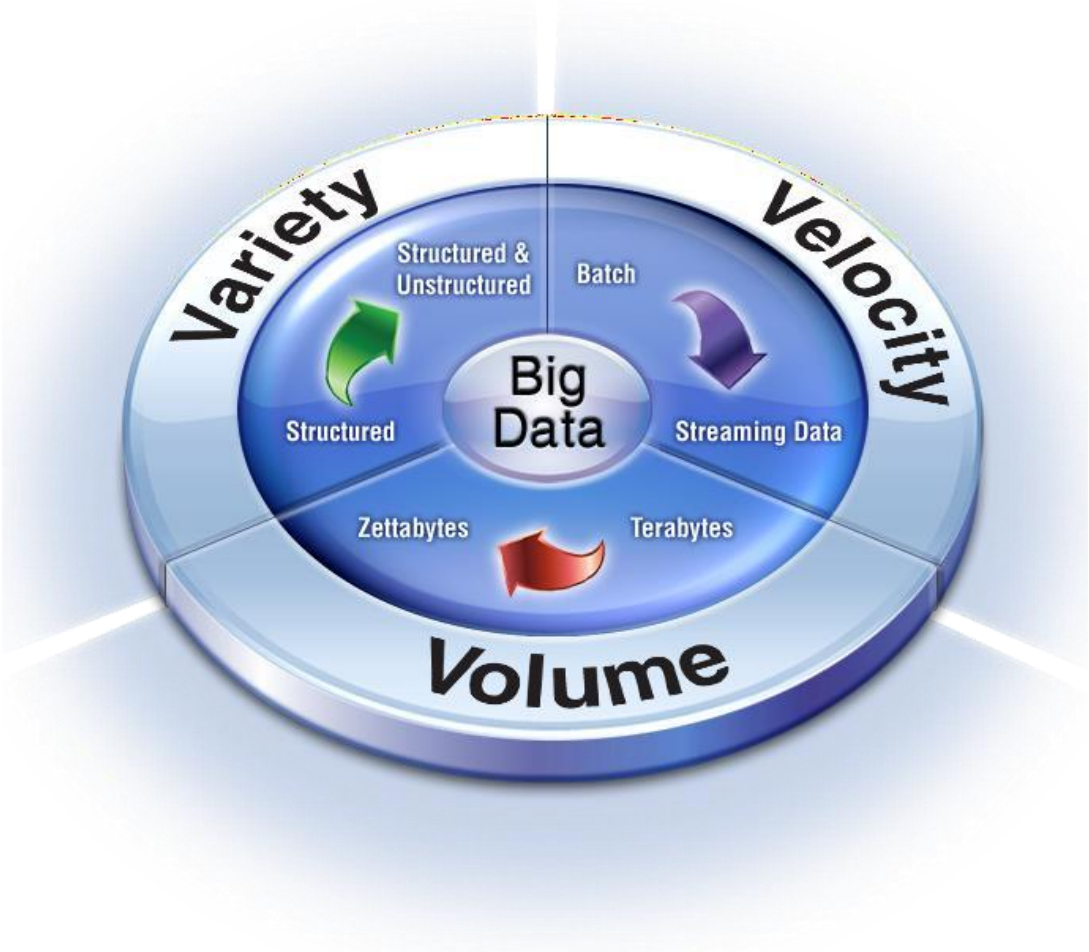
A Formal Definition

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

(http://www.gartner.com/)



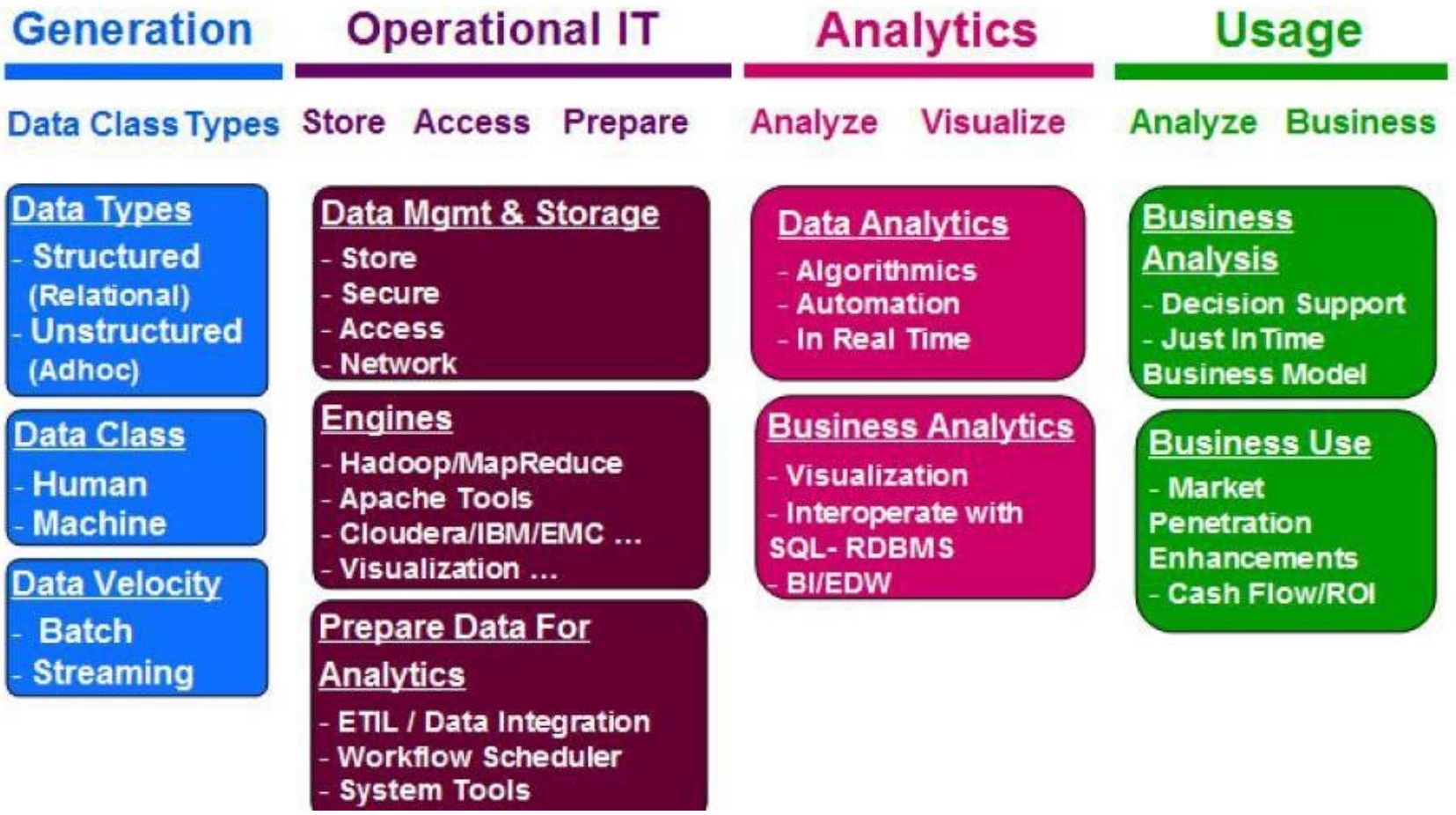
The Three V's of Big Data



Volume, Velocity and Variety

- **Volume:** Enterprises are awash with ever-growing data of all types, easily amassing terabytes—even petabytes—of information.
  - Turn 12 terabytes of Tweets created each day into improved product sentiment analysis
  - Convert 350 billion annual meter readings to better predict power consumption
- **Velocity:** Sometimes 2 minutes is too late. 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
- **Variety:** Big data is any type of data - structured and unstructured data such as text, sensor data, audio, video, click streams, log files and more. New insights are found when analyzing these data types together.
  - Monitor 100's of live video feeds from surveillance cameras to target points of interest
  - Exploit the 80% data growth in images, video and documents to improve customer satisfaction

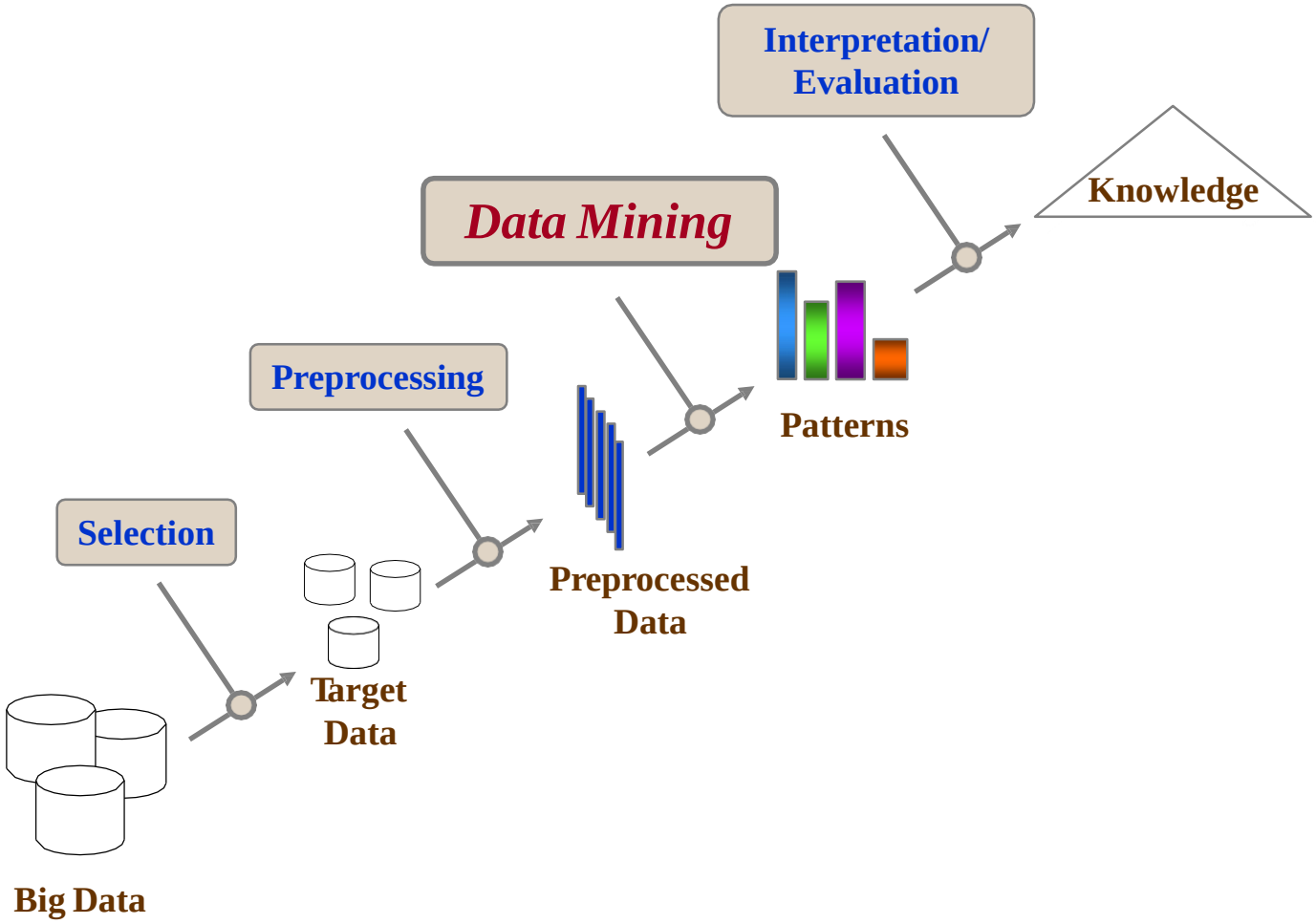
Big Data Ecosystem



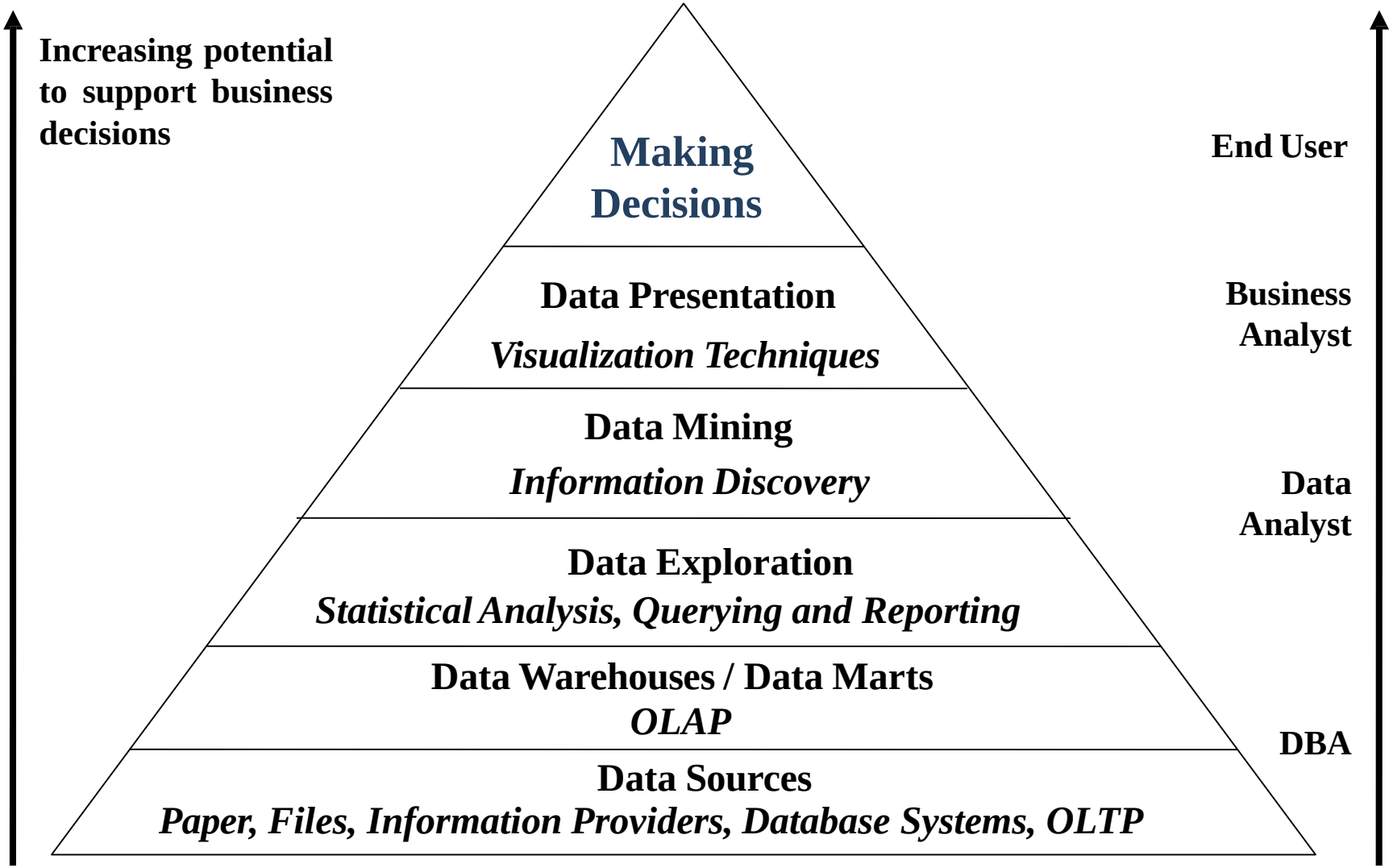
Source : IMEX research-Big Data Industry Report

Life Cycle of Big Data

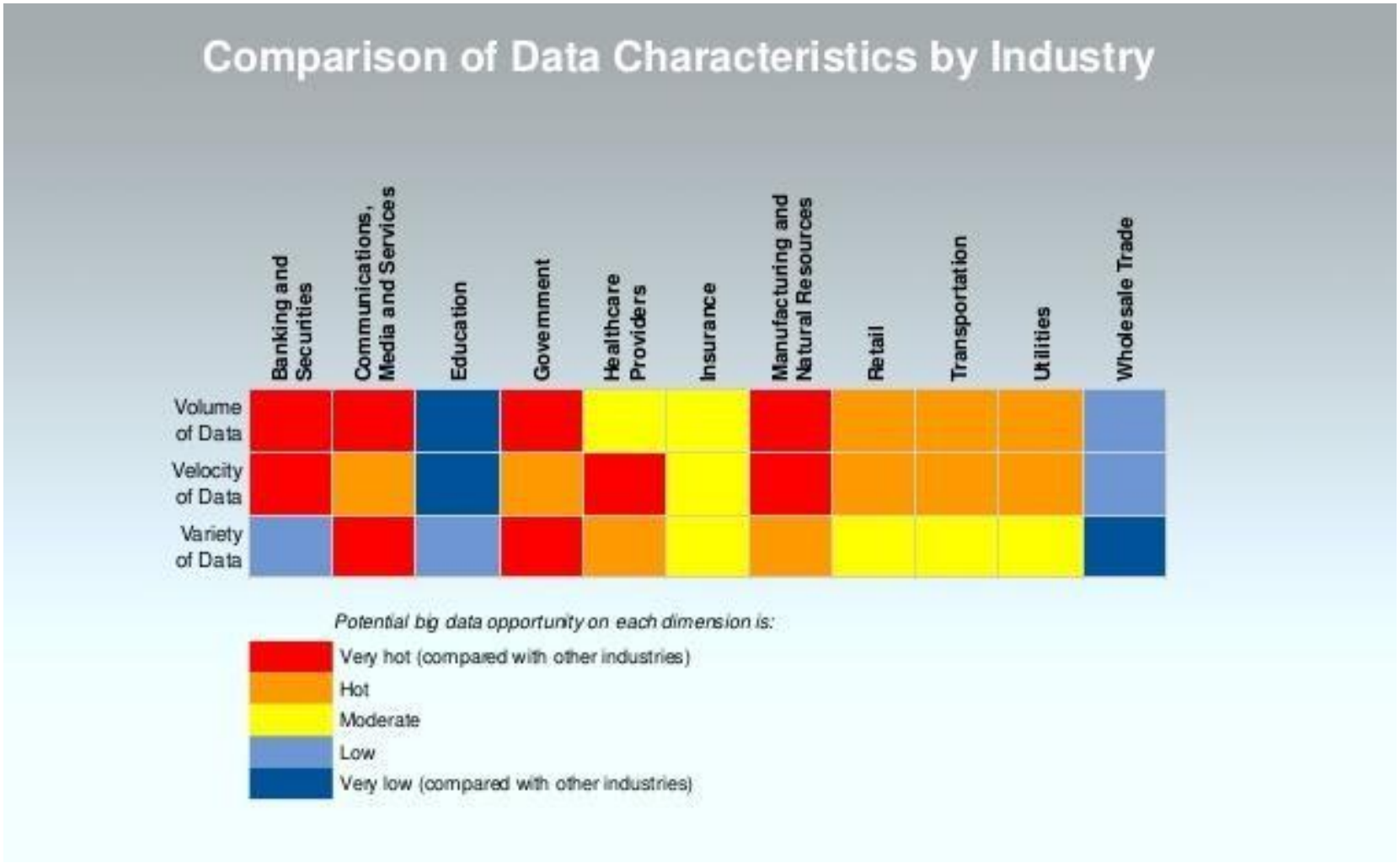
Computational View of Big Data



Where Engineers fit?



Industries using Big Data



Industries using Big Data

- Banking and Securities
  - Securities fraud early warning
  - Card fraud detection and Audit Trails
  - Credit risk reporting
  - Customer data information and Analytics
  - Application
    - The Securities Exchange Commission (SEC) is using big data to monitor financial market activity. They are currently using network analytics and natural language processors to catch illegal trading activity in the financial markets.

Banking

- Card marketing
  - By identifying customer segments, card issuers and acquirers can improve profitability with more effective acquisition and retention programs, targeted product development, and customized pricing.
- Fraud detection
  - Fraud is enormously costly. By analyzing past transactions that were later determined to be fraudulent, banks can identify patterns.
- Predictive life-cycle management
  - DM helps banks predict each customer's lifetime value and to service each segment appropriately (for example, offering special deals and discounts).

Industries using Big Data

- Communication, Media and Entertainment
  - Collecting, analyzing and utilizing customer insights
  - Understanding patterns of real-time media content
  - Application
    - Wimbledon Championships leverages big data to deliver detailed sentiment analysis on the tennis matches to TV, mobile and web users in real-time.

Telecommunication

- Call detail record analysis
  - Telecommunication companies accumulate detailed call records. By identifying customer segments with similar use patterns, the companies can develop attractive pricing and feature promotions.
- Customer loyalty
  - Some customers repeatedly switch providers, or "churn", to take advantage of attractive incentives by competing companies. The companies can use DM to identify the characteristics of customers who are likely to remain loyal once they switch, thus enabling the companies to target their spending on customers who will produce the most profit.

Industries using Big Data

- Retail and Wholesale Trade
  - Unutilized data derived from customer loyalty cards.
  - Data from PoS scanners, RFID, etc.
  - Application
    - Retail industries utilize Big Data for analytics and for other uses including:
      - Optimized staffing through data from shopping patterns, local events etc.
      - Reduced fraud and
      - Timely analysis of inventory

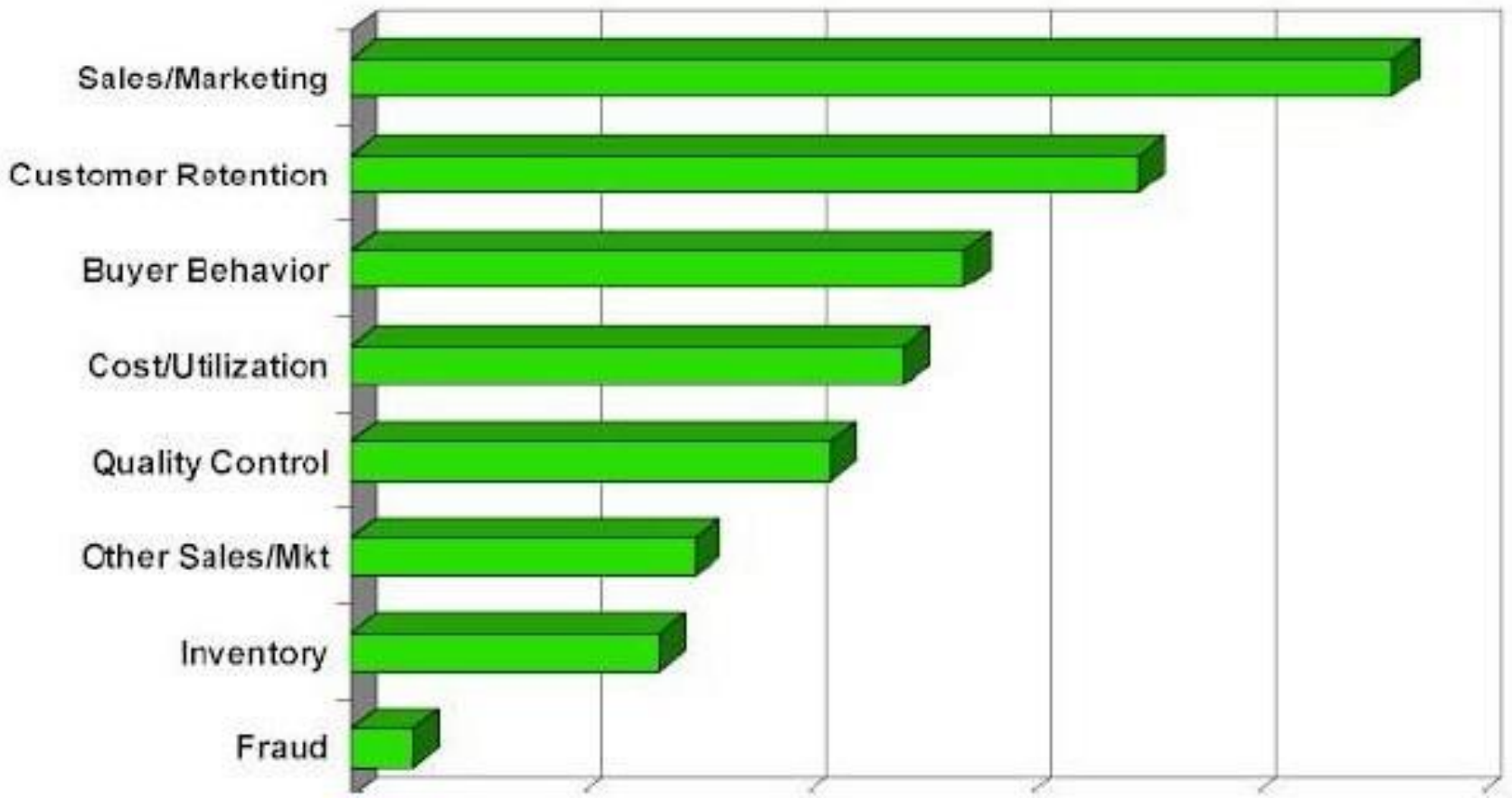
Retail

- Performing basket analysis
  - Which items customers tend to purchase together? Improves stocking, store layout strategies, and promotions.
- Sales forecasting
  - Examining time-based patterns helps retailers make stocking decisions. If a customer purchases an item today, when are they likely to purchase a complementary item?
- Database marketing
  - Profiling of customers with certain behaviors, for example, those who purchase designer labels clothing or those who attend sales. To focus cost–effective promotions.
- Merchandise planning and allocation
  - For new stores, merchandise planning and allocation by examining patterns in stores with similar demographic characteristics. Retailers can also use data mining to determine the ideal layout for a specific store.

Industries using Big Data

- Healthcare Providers
- Education
- Manufacturing and Natural Resources
- Government
- Insurance
- Transportation

Big Data Applications



Fraud Detection

- Applications: Health care, retail, credit card service, telecomm.
  - Auto insurance
    - Ring of collisions
  - Money laundering
    - Suspicious monetary transactions
  - Medical insurance
    - Professional patients
    - Unnecessary or correlated screening tests
  - Telecommunications: phone-call fraud
    - Phone call model: destination of the call, duration, time of day or week. Analyze patterns that deviate from an expected norm
  - Retail industry
    - Analysts estimate that 38% of retail shrink is due to dishonest employees
  - Anti-terrorism

Other Applications

- Customer segmentation
  - All industries can take advantage of DM to discover discrete segments in their customer bases by considering additional variables beyond traditional analysis.
- Manufacturing
  - Through choice boards, manufacturers can customize products for customers. Needs of customers.
- Warranties
  - Manufacturers need to predict the number of customers who will submit warranty claims and the average cost of those claims.
- Frequent flier incentives
  - Airlines can identify groups of customers that can be given incentives to fly more.

Other Applications

- Sports
  - IBM Advanced Scout analyzed NBA game statistics (shots blocked, assists, and fouls) to gain competitive advantage for New York Knicks and Miami Heat
- Internet Web Surf-Aid
  - IBM Surf-Aid applies data mining algorithms to Web access logs for market-related pages to discover customer preferences and behavior pages, analyzing effectiveness of Web marketing, improving Web site organization, etc.

Large Dataset Challenges