Tarbiat Modares University

AI in Healthcare Big Data Analytics

Ali Asghar Safaei

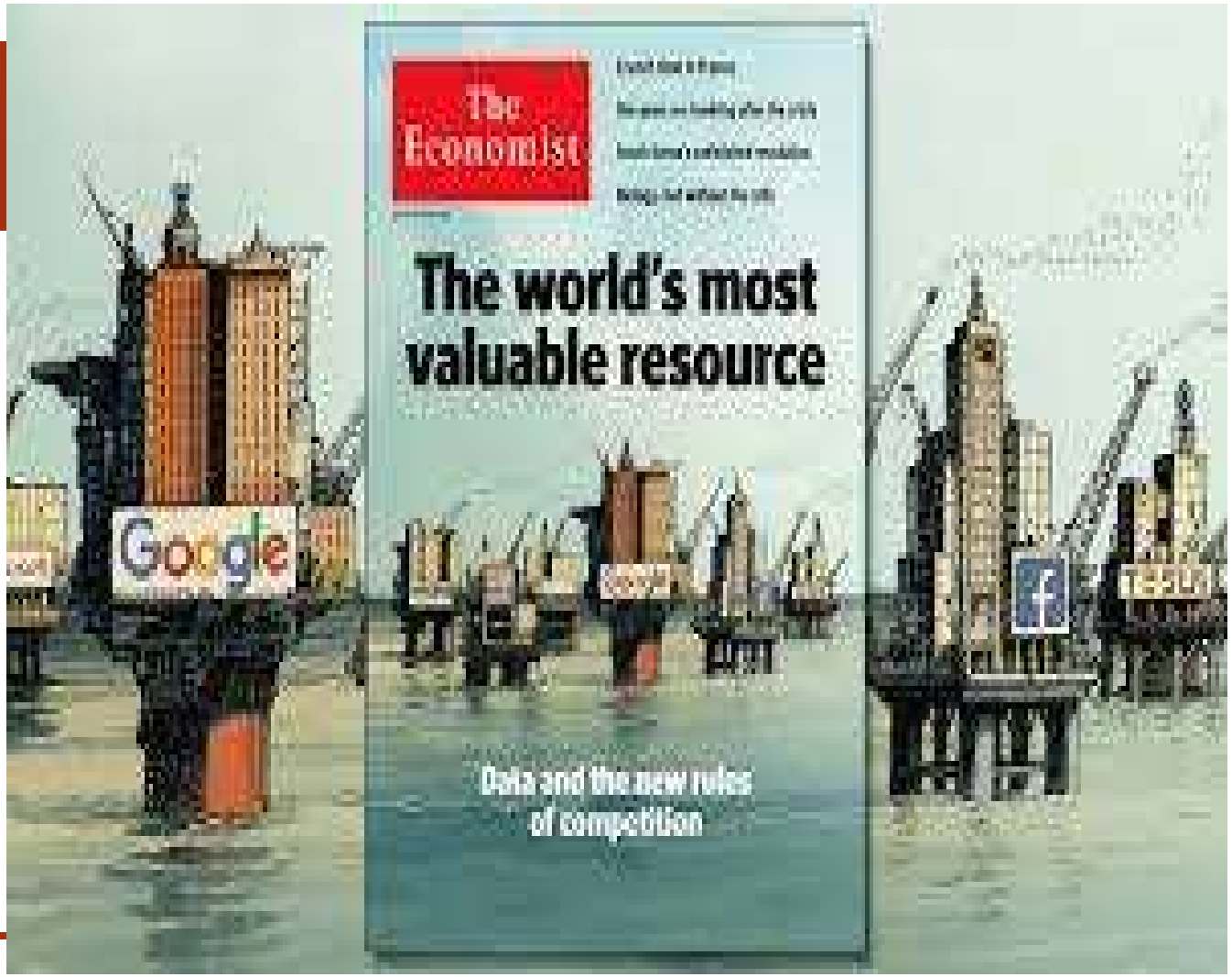
Associate Professor of BioMedical Informatics (Head of department)
Adj. Prof. of Data Sceince (Faculty of Interdisciplinary Sci. and Tech)
Tarbiat Modares University
aa.safaei@modares.ac.ir

OUTLINE

- Introduction
  - The Big Data Problem
- AI for data analytics (Data Mining)
- Data Science applications in Healthcare

INTRODUCTION

A radical rethink is required as data becomes the oil of the digital era. Our cover editorial this week:
http://economist.com/news/leaders/21721656-worlds-most-valuable-resource-no-longer-oil-data-data-economy-demands-new?fsrc=scn/tw/te/bl/ed/regulatingtheinternetgiantsthedataeconomydemandsanewapproachtoantitrust…



This is Big Data.

Every day, 2.5 quintillion bytes (=2,5 exabytes) of data are created.

This data comes from:

digital pictures, videos, posts to social media sites, intelligent sensors, purchase transaction records,

cell phone GPS signals to name a few.

In 2013, estimates reached **4 *zettabytes*** of data generated worldwide (*)

- Mary Meeker and Liang Yu, Internet Trends, Kleiner Perkins Caulfield Byers, 2013, http://www.slideshare.net/kleinerperkins/kpcb-internet-trends-2013.

How Big is Big Data?

1 megabyte = 1,000,000 =10^6 bytes

1 gigabyte = 10^9bytes

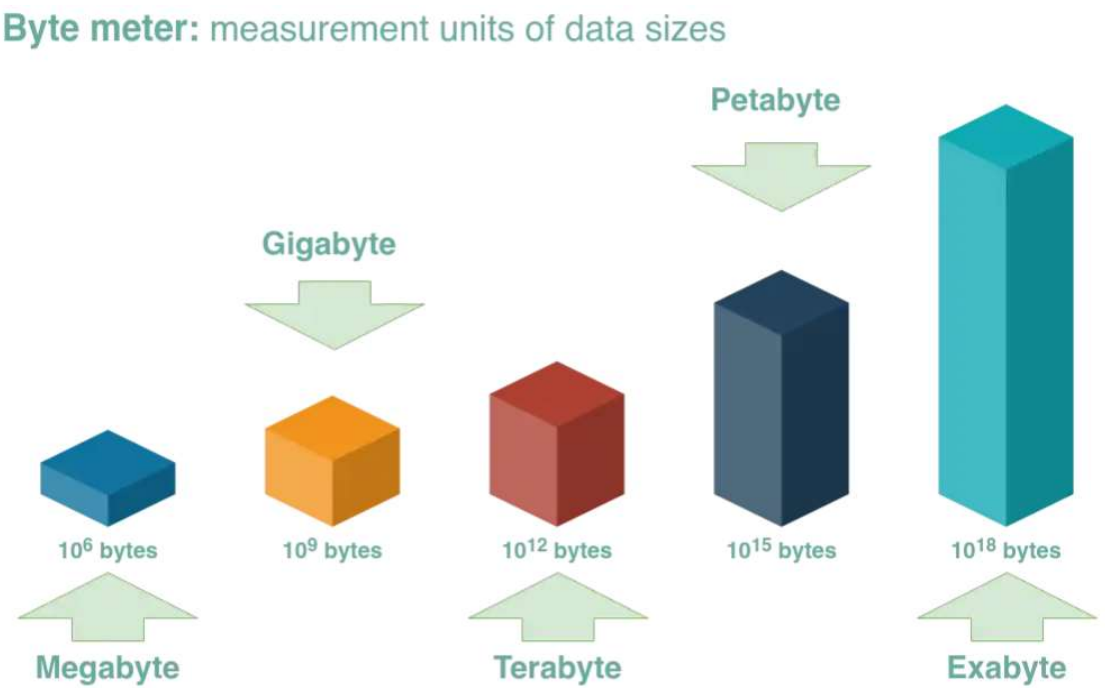
1 terabyte = 1,000,000,000,000 bytes = 10^{12}bytes

---

1 *petabyte* is 1,000 terabytes (TB) =10^{15}bytes

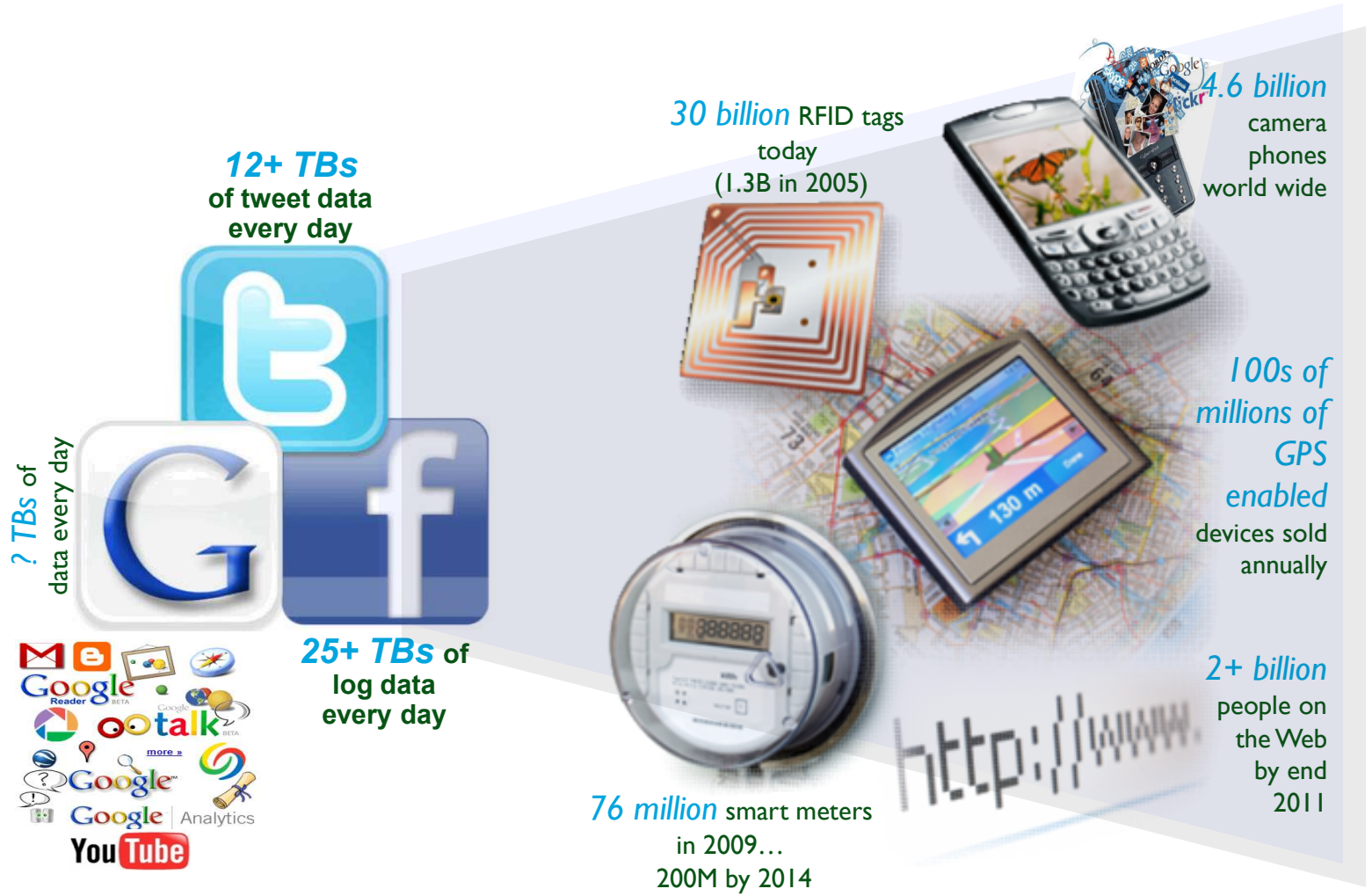
1 exabyte = 10^{18}bytes

1 *zettabyte* is 1,000 000,000,000,000,000,000== 10^{21}bytes

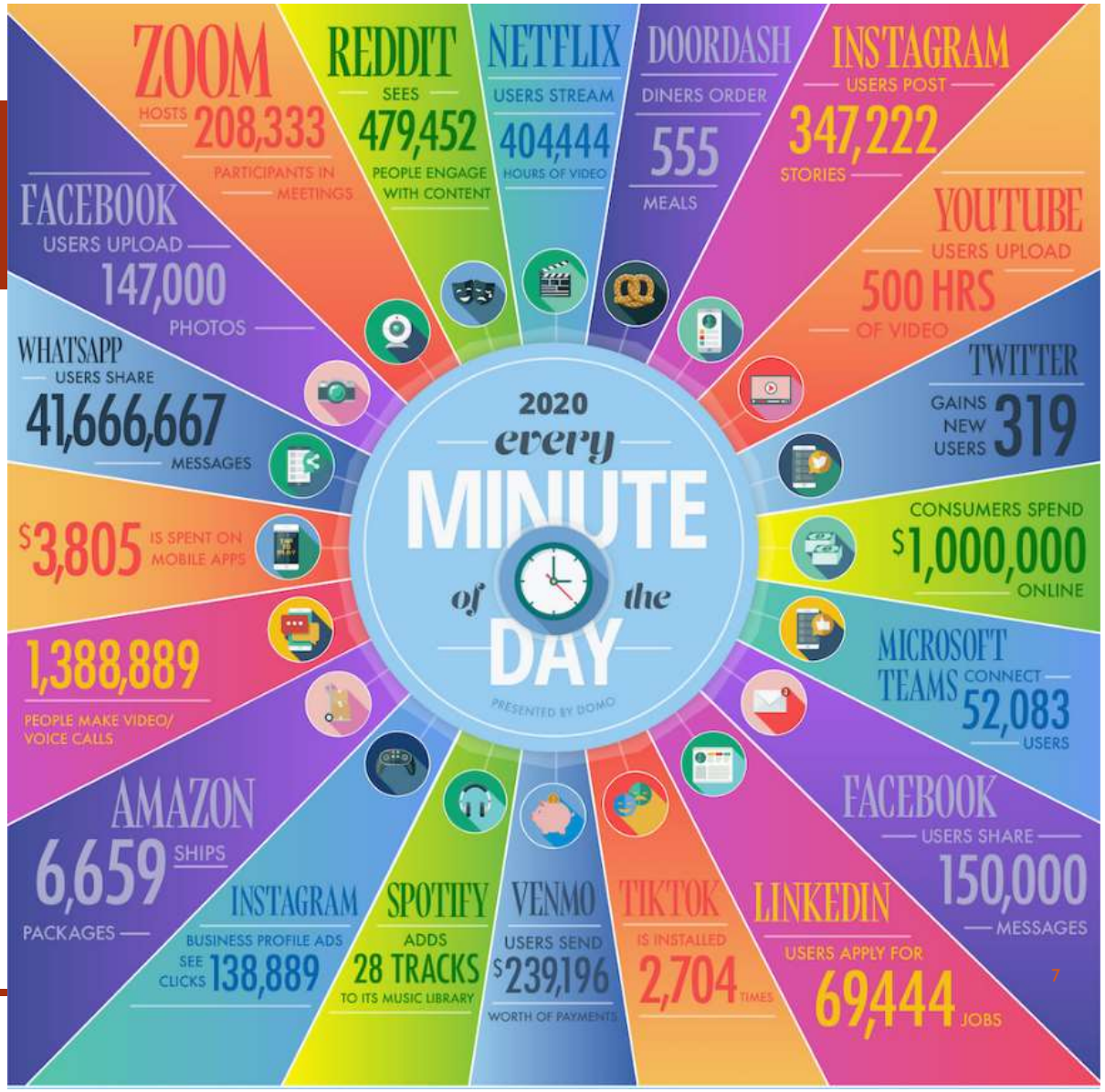


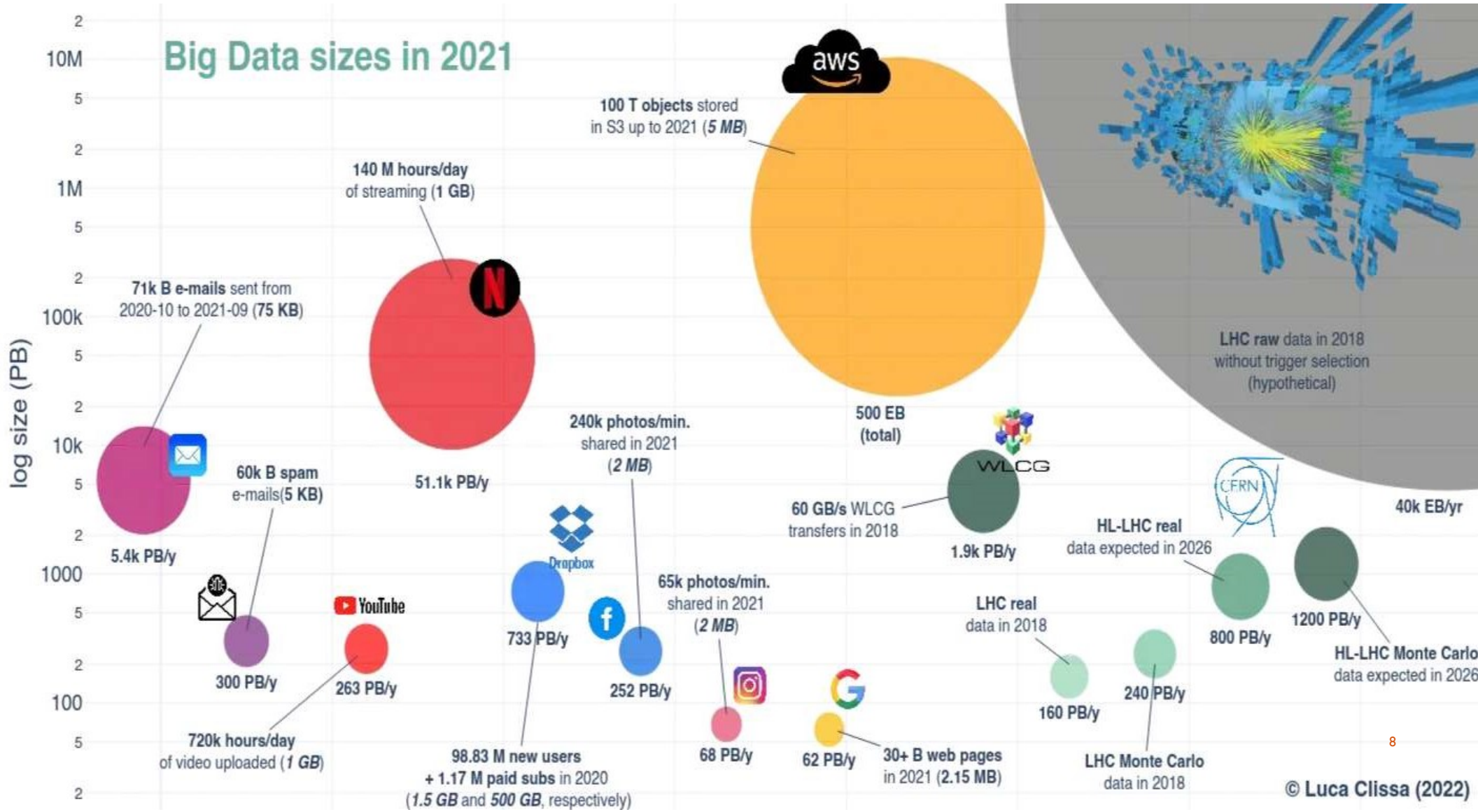
"*Imagine that every person (320,590,000) in the United States took a digital photo every second of every day for over a month.*

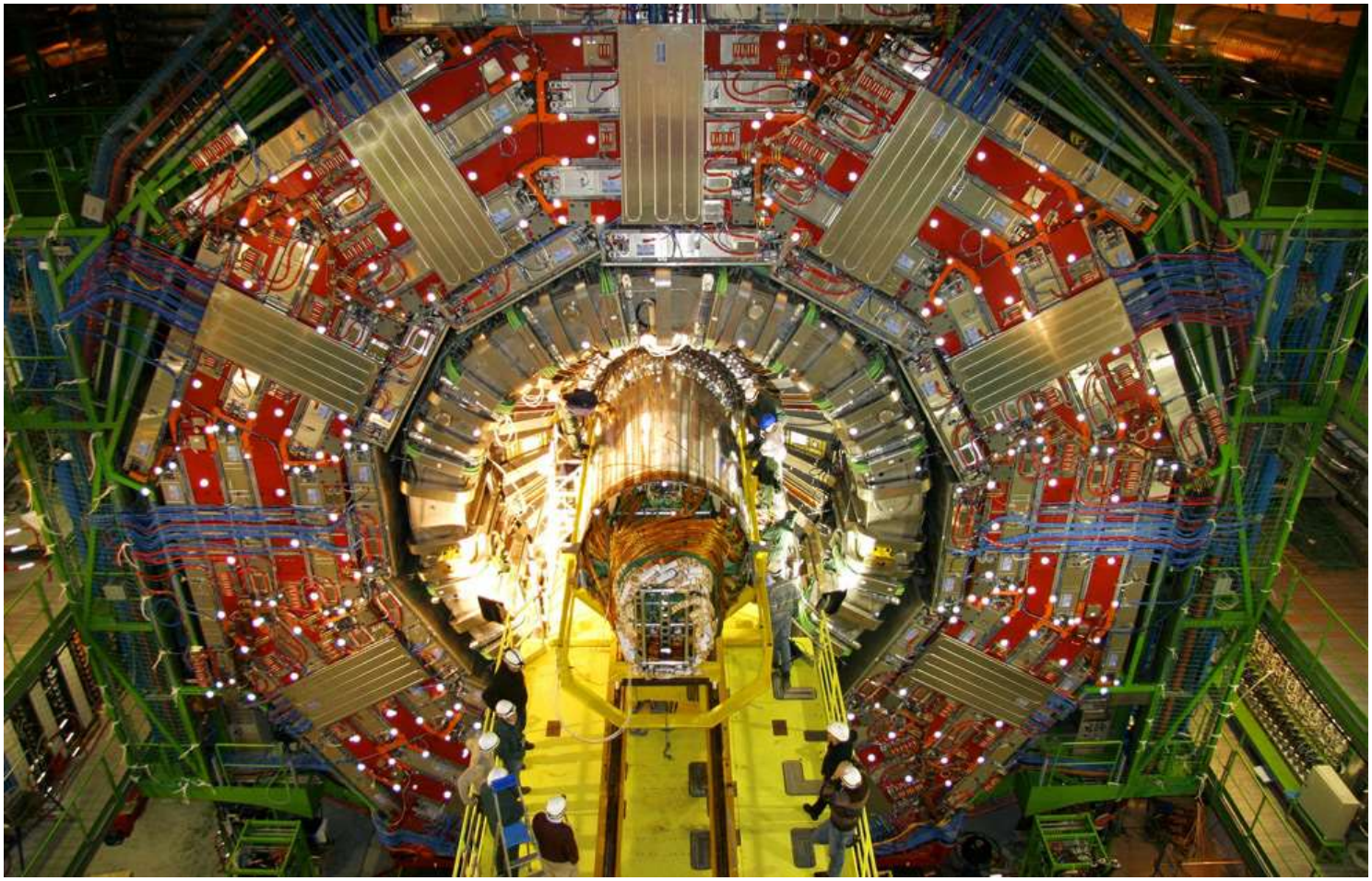
All of those photos put together would equal about ***one zettabyte***" (*)


12+ TBs
of tweet data
every day
? TBs of
data every day
25+ TBs of
log data
every day
30 billion RFID tags
today
(1.3B in 2005)
4.6 billion
camera
phones
world wide
100s of
millions of
GPS
enabled
devices sold
annually
2+ billion
people on
the Web
by end
2011
76 million smart meters
in 2009…
200M by 2014
Google
Reader
ooTalk
more »
Google
Google Analytics
YouTube
http://www.

DATA NEVER SLEEPS




Big Data sizes in 2021
log size (PB)
2
10M
5
2
1M
5
2
100k
5
2
10k
5
2
1000
5
2
100
5
2
100 T objects stored in S3 up to 2021 (5 MB)
500 EB (total)
LHC raw data in 2018 without trigger selection (hypothetical)
40k EB/yr
140 M hours/day of streaming (1 GB)
51.1k PB/y
71k B e-mails sent from 2020-10 to 2021-09 (75 KB)
5.4k PB/y
60k B spam e-mails(5 KB)
300 PB/y
720k hours/day of video uploaded (1 GB)
263 PB/y
240k photos/min. shared in 2021 (2 MB)
733 PB/y
252 PB/y
65k photos/min. shared in 2021 (2 MB)
68 PB/y
98.83 M new users + 1.17 M paid subs in 2020 (1.5 GB and 500 GB, respectively)
62 PB/y
30+ B web pages in 2021 (2.15 MB)
60 GB/s WLCG transfers in 2018
1.9k PB/y
LHC real data in 2018
160 PB/y
240 PB/y
LHC Monte Carlo data in 2018
HL-LHC real data expected in 2026
800 PB/y
1200 PB/y
HL-LHC Monte Carlo data expected in 2026
© Luca Clissa (2022)



the data detected by the electronic equipment of the Large Hadron Collider (LHC) experiments at *CERN*. In one year, the amount of processed information is around 40k ExaBytes (EB), which is 40 thousands billion gigabytes. To get an idea, that could be ten times the amount of data ever stored on Amazon S3!

A DAY IN DATA

The exponential growth of data is undisputed, but the numbers behind this explosion - fuelled by internet of things and the use of connected devcies – are hard to comprehend, particularly when looked at in the context of one day

500m

tweets are sent every day

Twitter

4PB

of data created by Facebook, including

350m photos

100m hours of video watch time

Facebook Research

320bn

emails to be sent each day by 2021

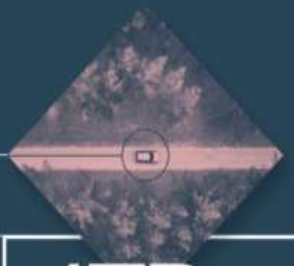
294bn

billion emails are sent

Radicati Group

306bn

emails to be sent each day by 2020



65bn

messages sent over WhatsApp and two billion minutes of voice and video calls made

Facebook

3.9bn

people use emails

4TB

of data produced by a connected car

Intel

DEMYSTIFIYING DATA UNITS

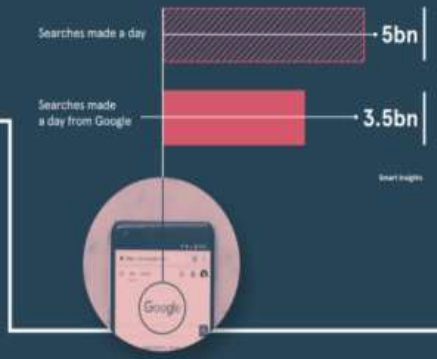
From the more familiar 'bit' or 'megabyte', larger units of measurement are more frequently being used to explain the masses of data

Unit		Value	Size
b	bit	0 or 1	1/8 of a byte
B	byte	8 bits	1 byte
KB	kilobyte	1,000 bytes	1,000 bytes
MB	megabyte	$1{,}000^2$ bytes	1,000,000 bytes
GB	gigabyte	$1{,}000^3$ bytes	1,000,000,000 bytes
TB	terabyte	$1{,}000^4$ bytes	1,000,000,000,000 bytes
PB	petabyte	$1{,}000^5$ bytes	1,000,000,000,000,000 bytes
EB	exabyte	$1{,}000^6$ bytes	1,000,000,000,000,000,000 bytes
ZB	zettabyte	$1{,}000^7$ bytes	1,000,000,000,000,000,000,000 bytes
YB	yottabyte	$1{,}000^8$ bytes	1,000,000,000,000,000,000,000,000 bytes

*A lowercase "b" is used as an abbreviation for bits, while an uppercase "B" represents bytes.

ACCUMULATED DIGITAL UNIVERSE OF DATA

IDC



Smart Insights

463EB

of data will be created every day by 2025

IDC

95m

photos and videos are shared on Instagram

Instagram Business

28PB

to be generated from wearable devices by 2020

Statista

RACONTEUR

- **More than 65 billion messages are being exchanged daily via WhatsApp.**
  - WhatsApp users exchange more than 65 billion messages every day. The app connects more than 1 billion groups, and its users complete **over 55 million video calls daily**. (Source:Mybasis.com)

- **Between 2010 and 2020, data interactions increased by 5000%.**
  - In the last eleven years, the volume of data generated, harvested, copied, and consumed worldwide grew by almost 5000%. The main result is a significant **increase in data usage that went from 1.2 trillion gigabytes to 59 trillion gigabytes.** (Source:Forbes.com)

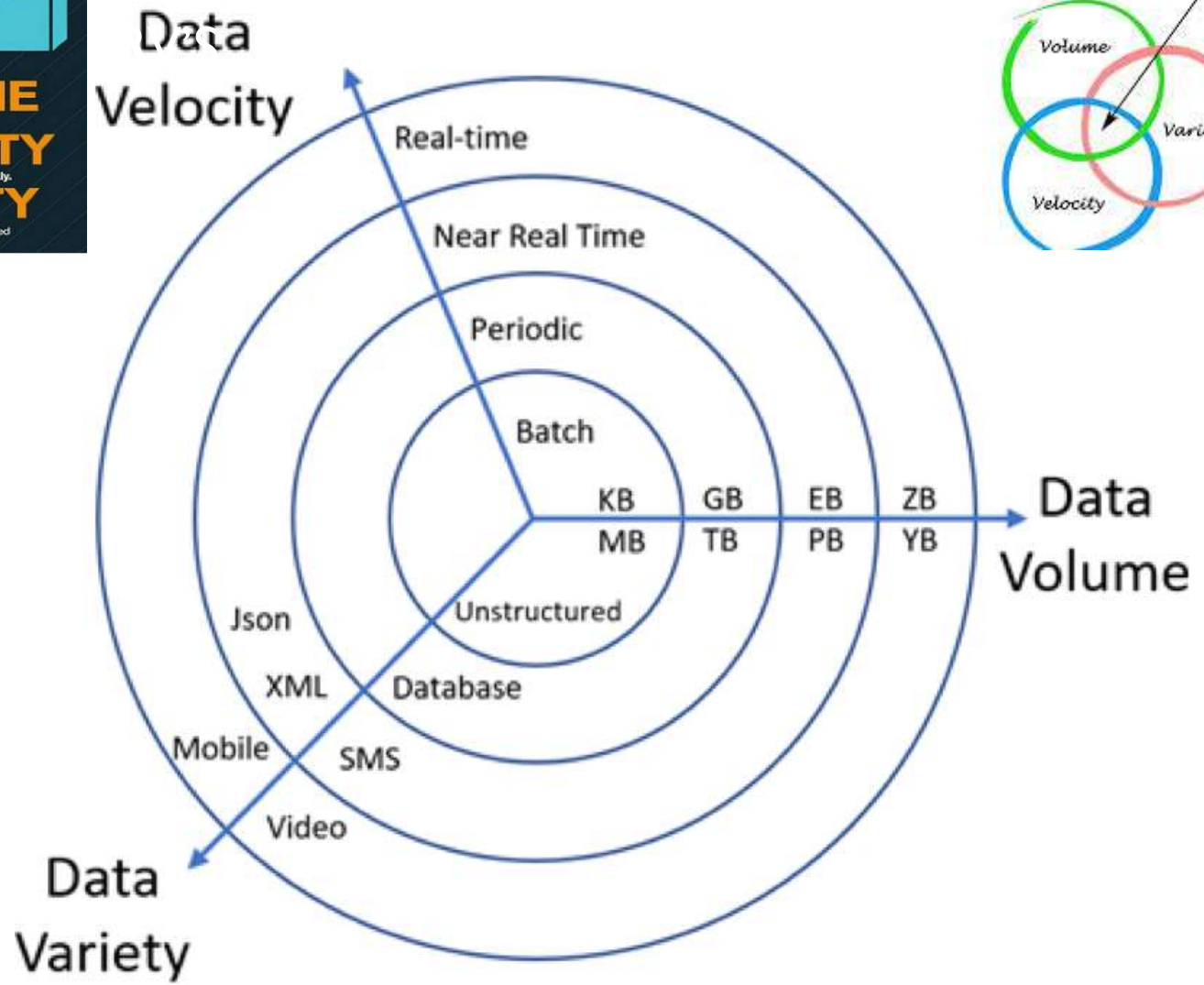
BIG DATA: DEFINITION

COLUMBIA UNIVERSITY

Definition and Characteristics of Big Data

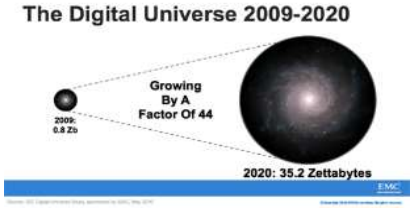
"*Big data is high-**volume**, high-**velocity** and high-**variety** information assets that demand **cost-effective**, **innovative** forms of information processing for **enhanced insight and decision making**.*" -- Gartner

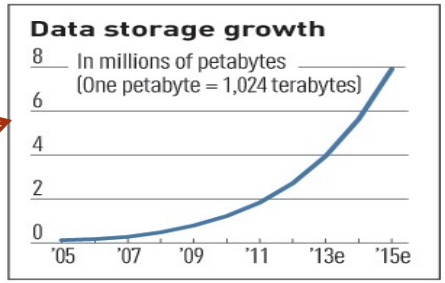

BIG
DATA?
VOLUME
Large amounts of data.
VELOCITY
Needs to be analyzed quickly.
VARIETY
Different types of structured and unstructured data.

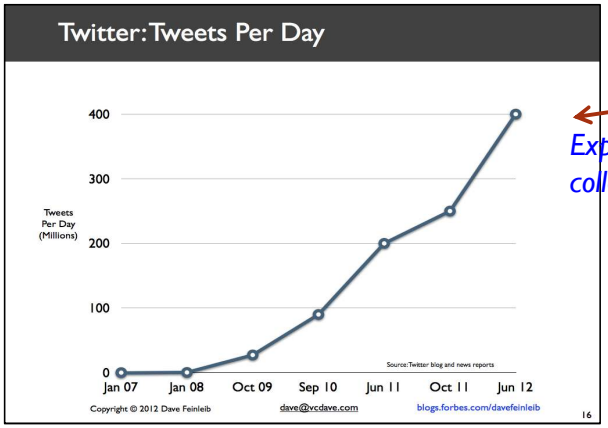

Data Velocity
Real-time
Near Real Time
Periodic
Batch
KB
MB
GB
TB
EB
PB
ZB
YB
Data Volume
Unstructured
Json
XML
Database
Mobile
SMS
Video
Data Variety

Big Data
Volume
Variety
Velocity

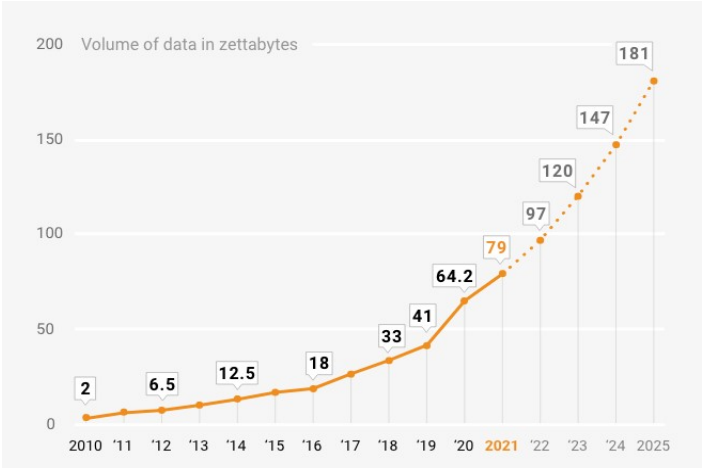
- **Data Volume**
  - 44x increase from 2009 2020
  - From 0.8 zettabytes to 35zb
- Data volume is increasing exponentially







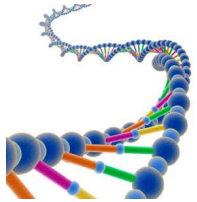
Exponential increase in collected/generated data

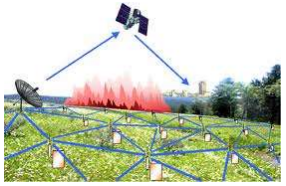
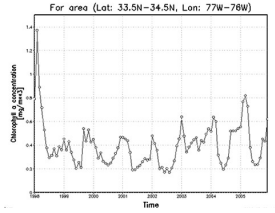


approximately 90% of it is replicated data, with only 10% being genuine, new data.
(Source:Statista.com)

VARIETY (COMPLEXITY)

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
  - Social Network, Semantic Web (RDF), …
- Streaming Data
  - You can only scan the data once
- A single application can be generating/collecting many types of data
- Big Public Data (online, weather, finance, etc)

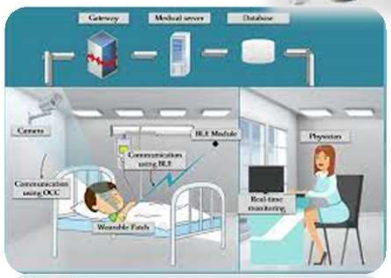




To extract knowledge➔ all these types of data need to linked together

80–90% of the data that internet users generate daily is unstructured. (Source:Cio.com)

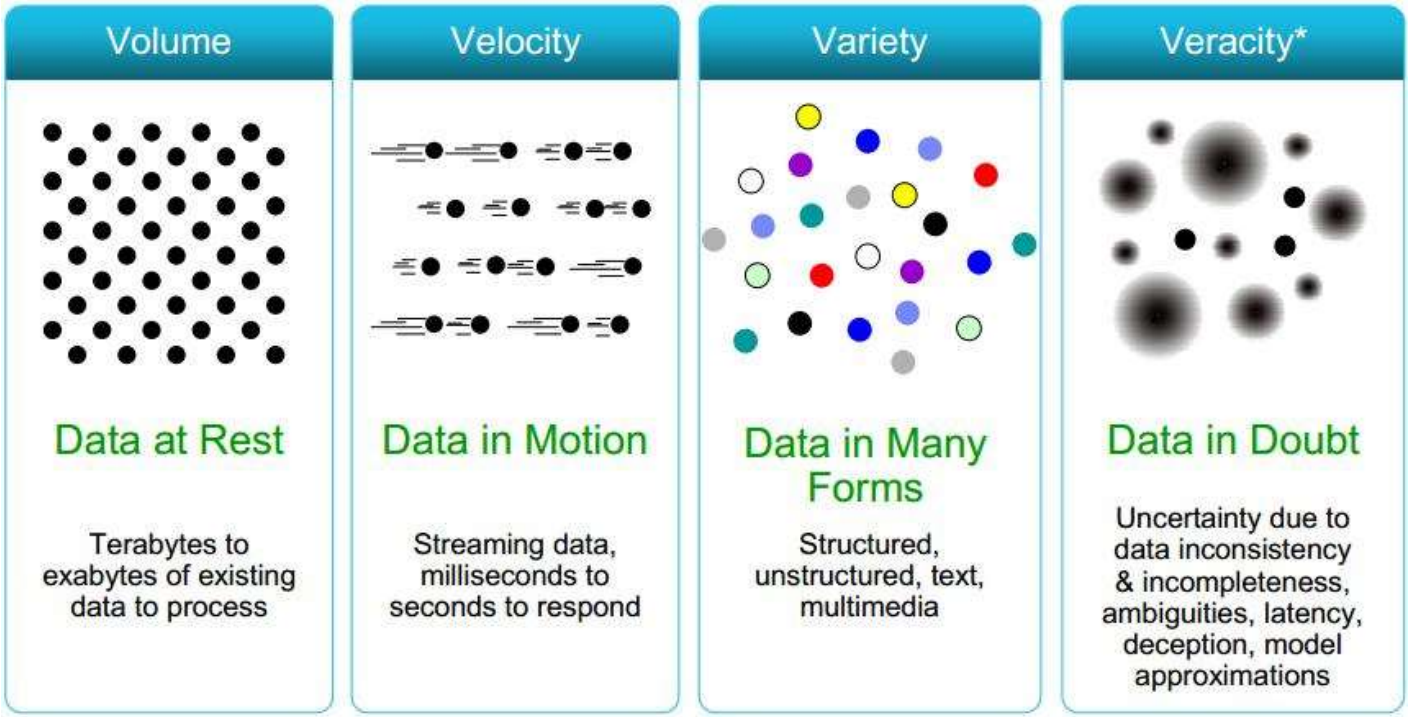
VELOCITY (SPEED)

- Data is begin generated fast and need to be processed fast
- Late decisions ➔ missing opportunities
- **Examples**
  - **E-Promotions:** Based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you
  - **Healthcare monitoring:** sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction



the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion

SOME MAKE IT 4V'S

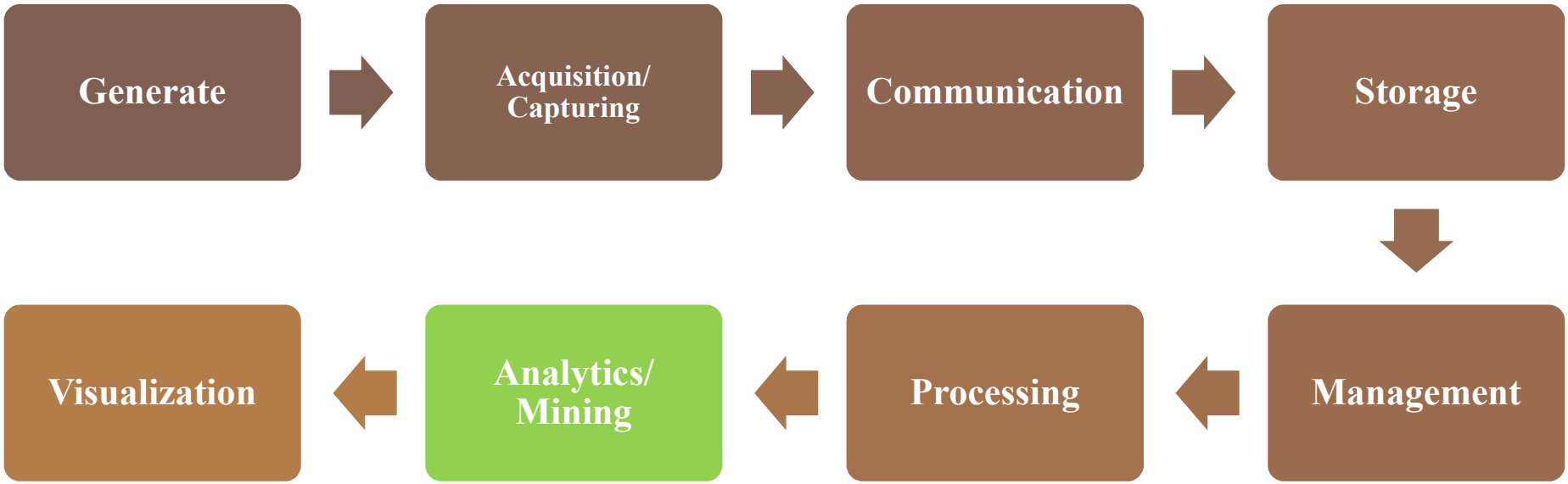


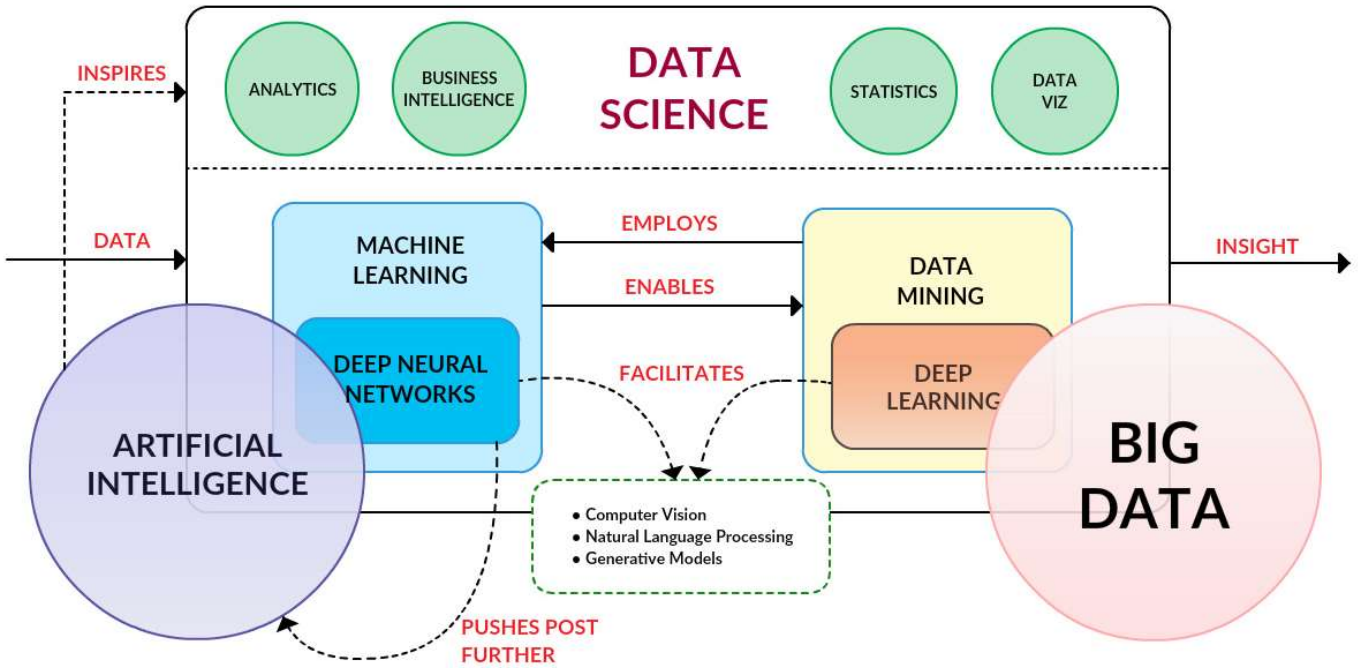
Another Definition of Big Data

"Big Data" *refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze*" (McKinsey Global Institute)

- This definition is Not defined in terms of data size (data sets will increase)
- Vary by sectors (ranging from a few dozen terabytes to multiple petabytes)

BIG DATA LIFECYCLE




INSPIRES
ANALYTICS
BUSINESS INTELLIGENCE
DATA SCIENCE
STATISTICS
DATA VIZ
DATA
MACHINE LEARNING
EMPLOYS
ENABLES
DATA MINING
INSIGHT
DEEP NEURAL NETWORKS
FACILITATES
DEEP LEARNING
ARTIFICIAL INTELLIGENCE
• Computer Vision
• Natural Language Processing
• Generative Models
BIG DATA
PUSHES POST FURTHER

What is Data Science?

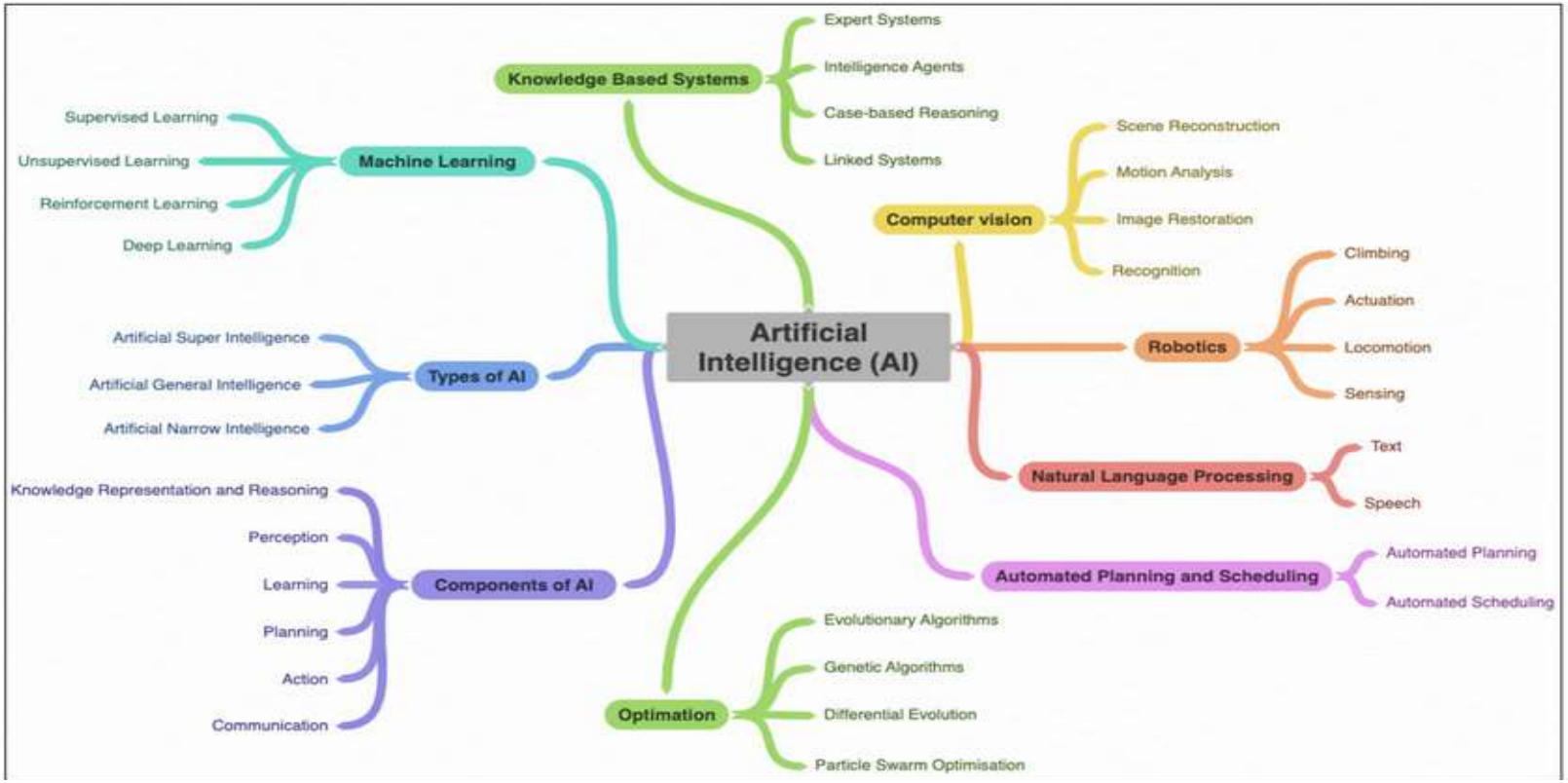
(sourcehttp://datascience.nyu.edu/what-is-data-science/

Data science involves using automated methods to analyze massive amounts of data and to extract knowledge from them.

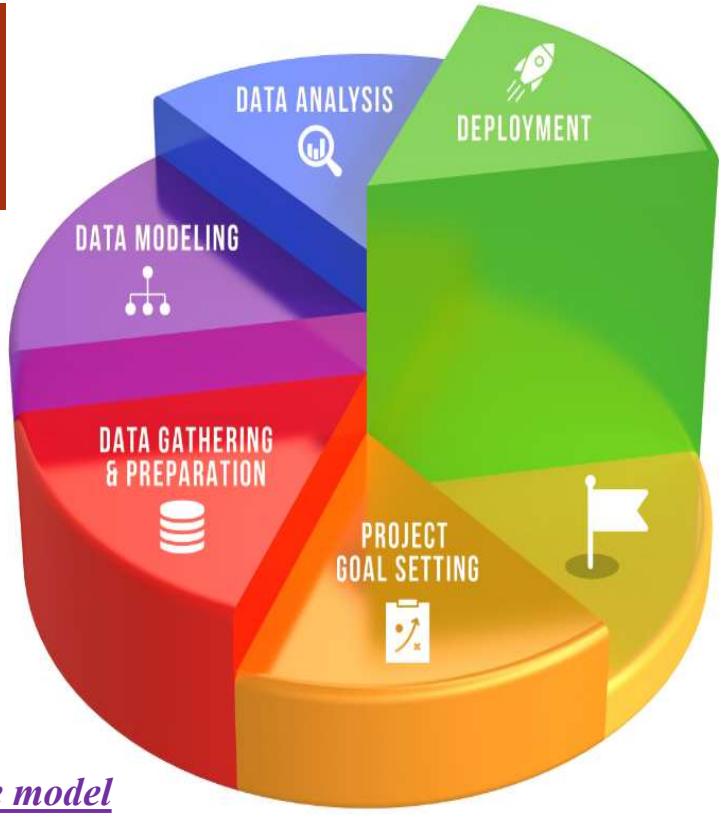
One way to consider data science is as an evolutionary step in interdisciplinary fields like business analysis that incorporate computer science, modeling, statistics, analytics, and mathematics.

ARTIFICIAL INTELLIGENCE (AI)

- "AI is the new electricity. It will transform and improve all areas of human life." Andrew Ng



ANALYTICS (DATA MINING STEPS)



1. **Problem definition and Data understanding/exploration**

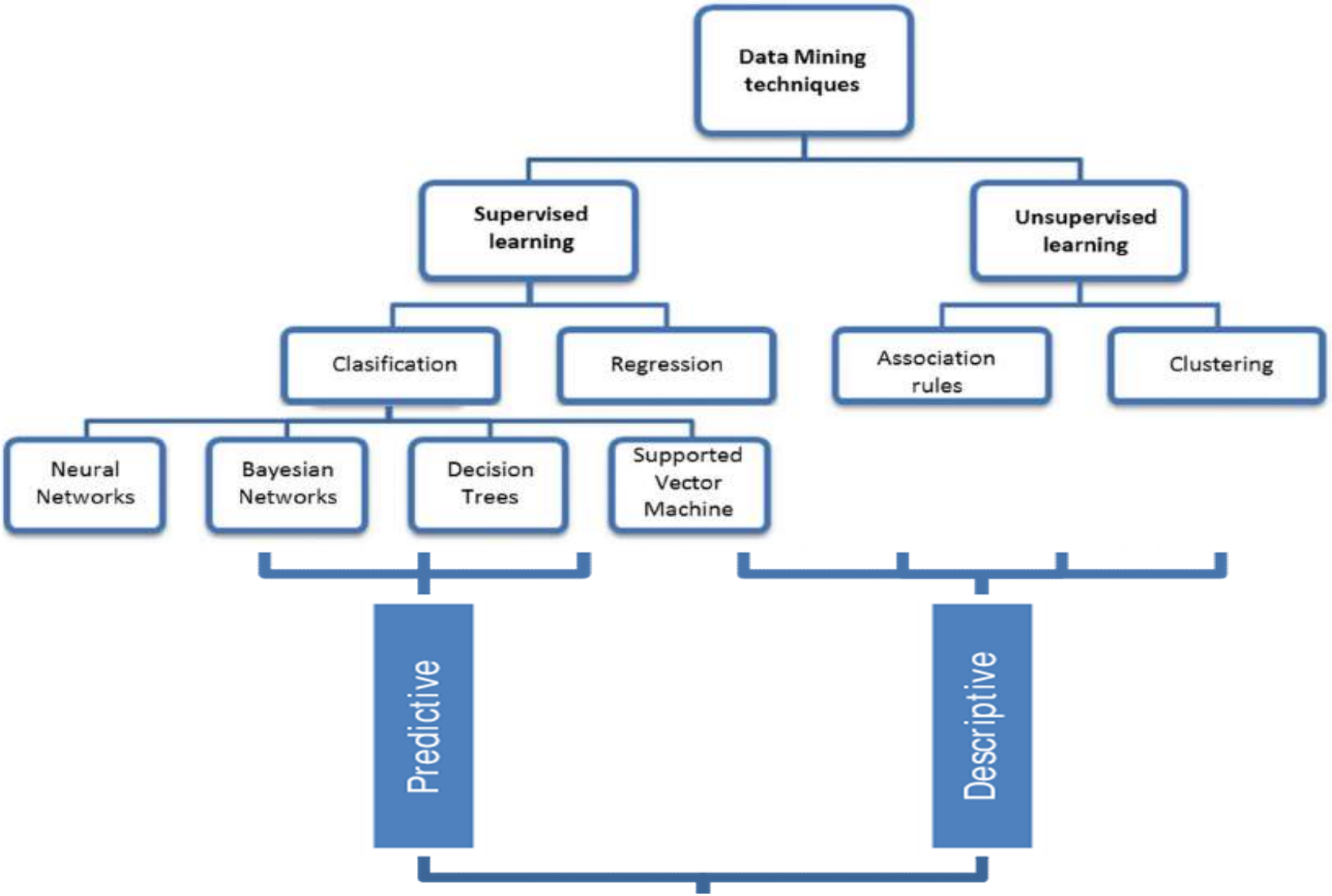
   Then, the **available data** is **collected** and explored. The **quality** of data is also **checked** in this phase.

2. **Pre-processing**
   - **Data preparation**
     - Data Integration,
     - Data Cleansing,
     - Normalization,
     - Feature Selection
   - *Selection* of **Data Mining technique**(s) and proper software **Tools**

3. **Modeling**

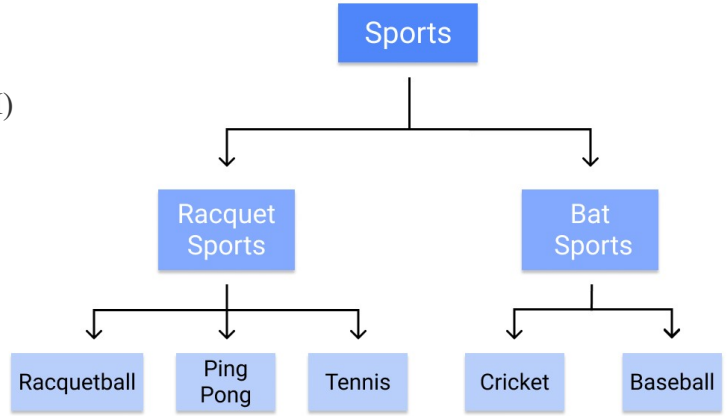
   Running DM techniques on prepared dataset in DM tools to construct the model

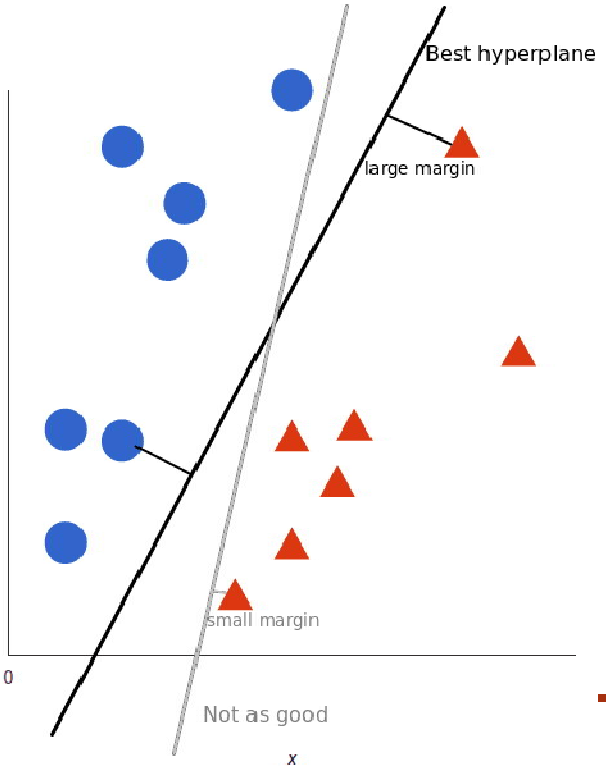
4. **Evaluation of the constructed model**
5. **Deployment**


Data Mining techniques
Supervised learning
Unsupervised learning
Clasification
Regression
Association rules
Clustering
Neural Networks
Bayesian Networks
Decision Trees
Supported Vector Machine
Predictive
Descriptive

DATA MINING (MACHINE LEARNING) TECHNIQUES

- **Classification** – In this items are classified into predefined groups and classes. This method depends upon predictions made using predefined techniques.
  - Decision Tree.
  - Support Vector Machines (SVM)
  - Logistic Regression.
  - Naive Bayes.
  - K-Nearest Neighbors
  - Artificial Neural Network





Naive Bayes

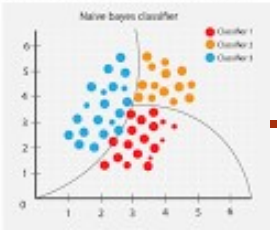
thatware.co

In machine learning, naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

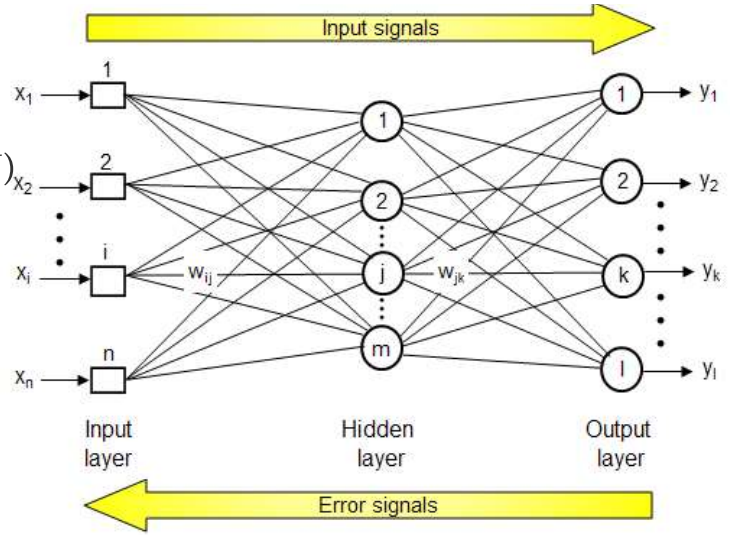
$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

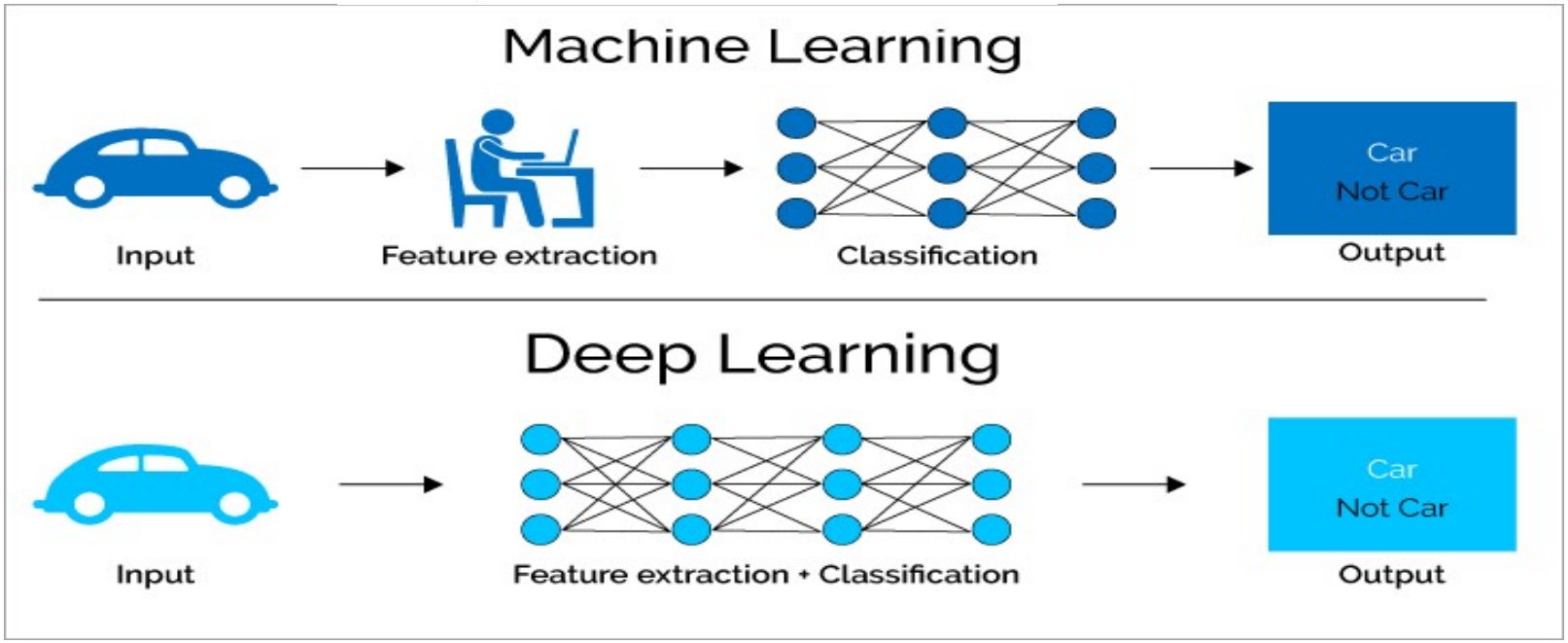
using Bayesian probability terminology, the above equation can be written as

$$\text{Posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$



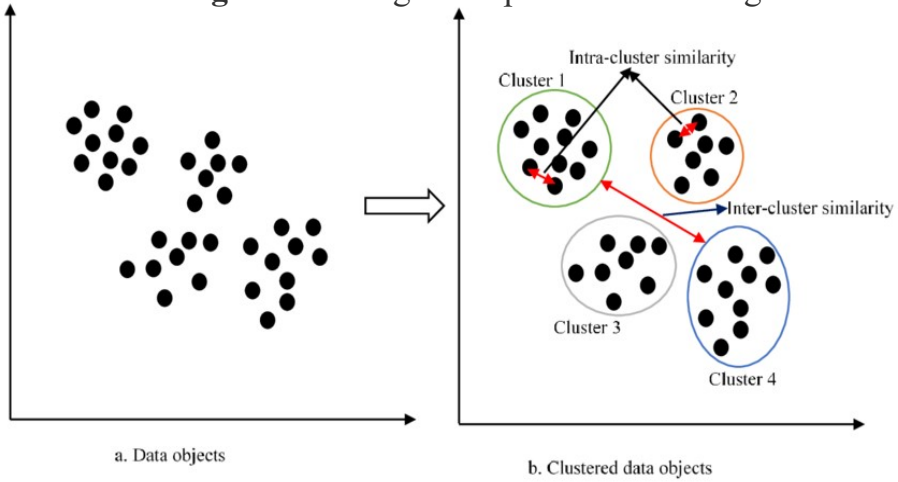
Artificial Neural Network (ANN)

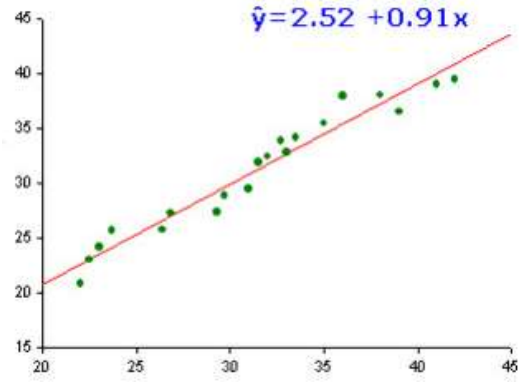




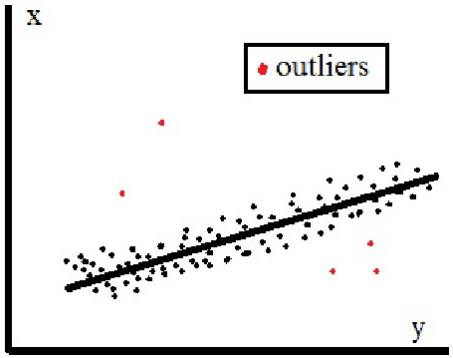
DATA MINING (MACHINE LEARNING) TECHNIQUES

- **Clustering –** Clustering is the process of making a cluster of abstract objects having similar characteristics.





- **Regression-** a technique for investigating the relationship between independent variables or features and a dependent variable or outcome.



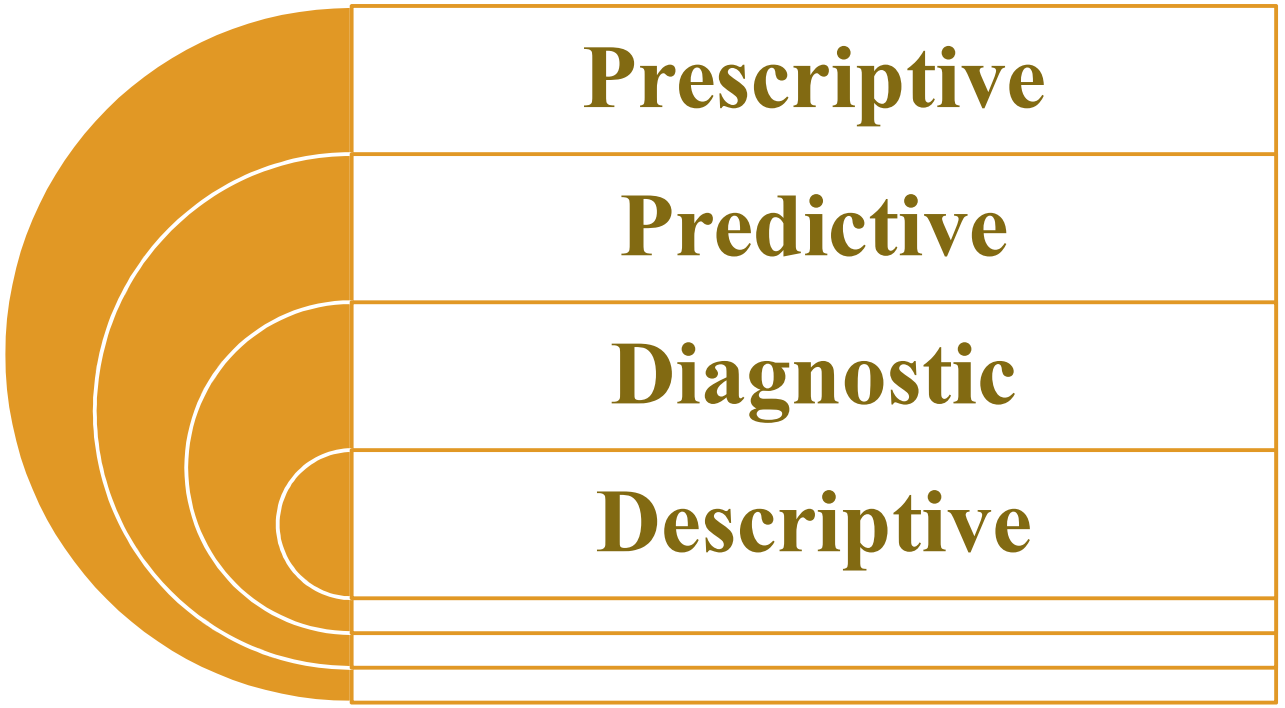
- **Anomaly/Outlier Detection**

DATA MINING (MACHINE LEARNING) TECHNIQUES

- **Association Rule Discovery –** In this technique, a pattern is identified based on the relationship between items of similar proceedings. A customer behavior can be analyzed by an analyst using association technique based on his buying patterns.

- **Sequential Pattern Discovery –** Sequential analysis is a technique that discovers and identifies similar patterns, events, and trends in transactional data over a certain period of time.

LEVELS OF ANALYTICS

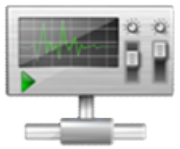


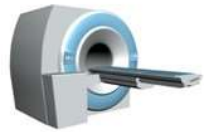
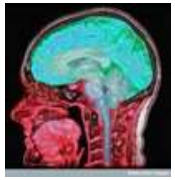
Why Big data analytics in Healthcare?




Healthcare Industry generates a huge amount of data such as

- Clinical data from CPOE
- Clinical decision support systems such as physician's written notes and prescriptions, medical imaging, laboratory, pharmacy, insurance
- Patient data in electronic health records (EHRs)
- Claims data
- Machine generated/sensor data, such as from monitoring vital signs
- Social media posts, including Twitter feeds, status updates on Facebook and other platforms
- Data maintained for regulatory compliance such as Affordable Care Act, HIE, ACO etc.



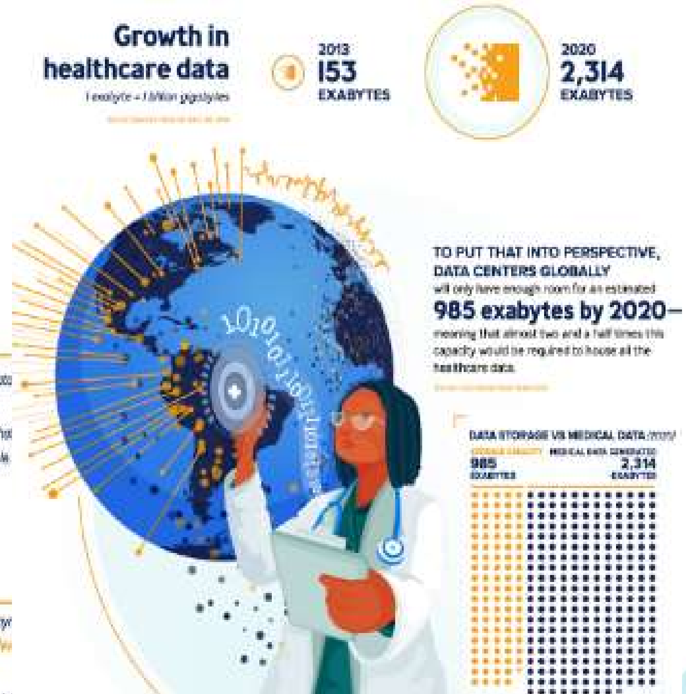



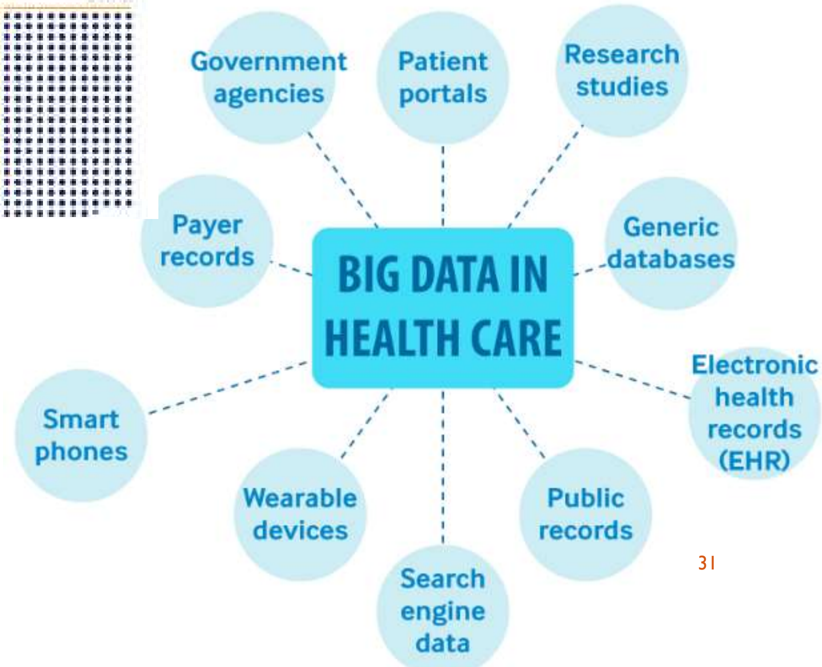
Reports say data from the U.S. healthcare system alone reached, in 2011, 150 Exabytes

At this rate of growth, big data for U.S. healthcare will soon reach the zettabyte (1021 gigabytes) scale and, not long after, the yottabyte (1024 gigabytes)

Industry has faced with unsustainable costs and enormous amounts of under-utilized data, health care needs more efficient practices, research, and tools to harness the full benefits of the big data


Big Data
Big data refers to the mass of structured and unstructured data generated worldwide.
In healthcare, this encompasses everything from electronic medical records,
to internet-connected (IoT) devices.
It spans four dimensions:
VOLUME
VERACITY
Establishing trust and accuracy in big data is imperative in the healthcare industry.
Data assurance includes guaranteeing that data analytics are error-free and credible.
VARIETY
VELOCITY
10:38
HEALTHCARE BUSINESSES
must learn to quickly distill information from masses of data and transform them into actionable insights.
The ability to extract these insights will power the future of health, and become a differentiator for companies to thrive and stay ahead of emerging competitors.


Growth in healthcare data
2013
153
EXABYTES
2020
2,314
EXABYTES
TO PUT THAT INTO PERSPECTIVE, DATA CENTERS GLOBALLY
will only have enough room for an estimated
985 exabytes by 2020—
meaning that almost two and a half times this capacity would be required to house all the healthcare data.
DATA STORAGE VS MEDICAL DATA
985
2,314


BIG DATA IN HEALTH CARE
Government agencies
Patient portals
Research studies
Payer records
Generic databases
Electronic health records (EHR)
Smart phones
Wearable devices
Search engine data
Public records

Applications for Big Data in Healthcare

Diagnostics
Data mining and analysis to identify causes of illness

Preventative medicine
Predictive analytics and data analysis of genetic, lifestyle, and social circumstances to prevent disease

Precision medicine
Leveraging aggregate data to drive hyper-personalized care

Medical research
Data-driven medical and pharmacological research to cure disease and discover new treatments and medicines

Reduction of adverse medication events
Harnessing of big data to spot medication errors and flag potential adverse reactions

Cost reduction
Identificaton of value that drives better patient outcomes for longterm savings

Population health
Monitor big data to identify disease trends and health strategies based on demographics, geography, and socio-economics

Applications of Healthcare Big Data Analytics

Benefit #1: Analyzing clinical data to improve medical research

- gathering and analyzing clinical data from various sources. Among the most useful sources of clinical information are EHRs, electronic medical records, personal health records, and public health records.
- Applying data analytics techniques

•Early and Accurate detection of disease
  - Diagnosis
  - Prognosis

• Prevention of unnecessary doctor's visits
•Optimized treatment plan
•Discovery of new drugs

•Etc

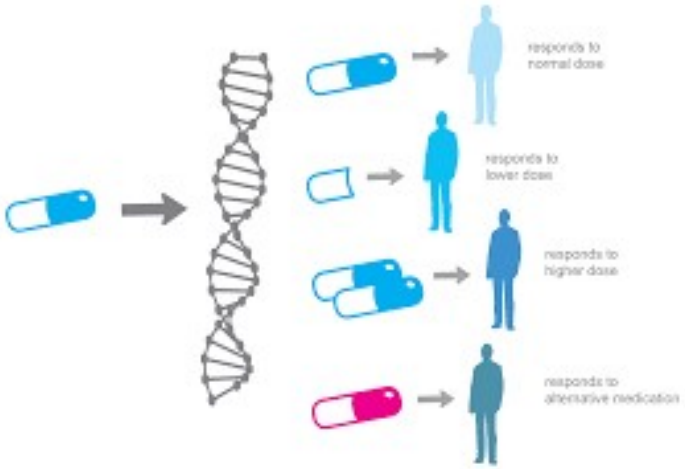
Applications of Healthcare Big Data Analytics

Benefit #2: Using patient data to improve health outcomes

- to treat patients safely and effectively while minimizing the trauma associated with their treatment (improving the Safety and Quality of Healthcare)

  By analyzing patient data, healthcare providers can lower readmission rates, reduce errors, and better identify at-risk populations. The types of patient data used in these analyses include blood sugar level, temperature, blood test results, and the patient's own wishes for care.

  •Personalization of patient care (Personalized Medicine or Tailored /Précised Medicine)



Applications of Healthcare Big Data Analytics

Benefit #3: Gaining operational insights from healthcare provider data

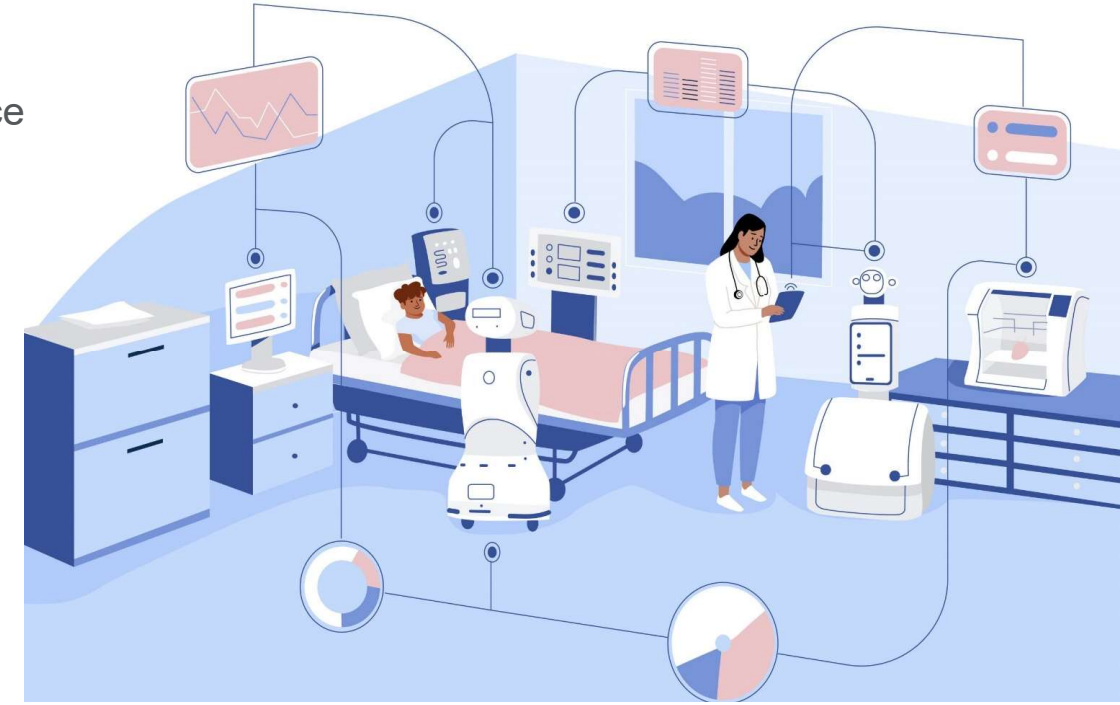
improving the Efficiency (resource utilization)

- Waste reduction: The annual cost of recoverable waste in the U.S. healthcare industry is estimated at $1 trillion
- Increasing hospital capacity. A less-expensive alternative to building new healthcare capacity is to apply analytics techniques to better manage demand for hospital beds and other healthcare resources.

Fraud Detection (e.g., Insurance companies)

- More accurate calculation of health insurance

•Digital Hospitals (IoT-Based Hospitals)



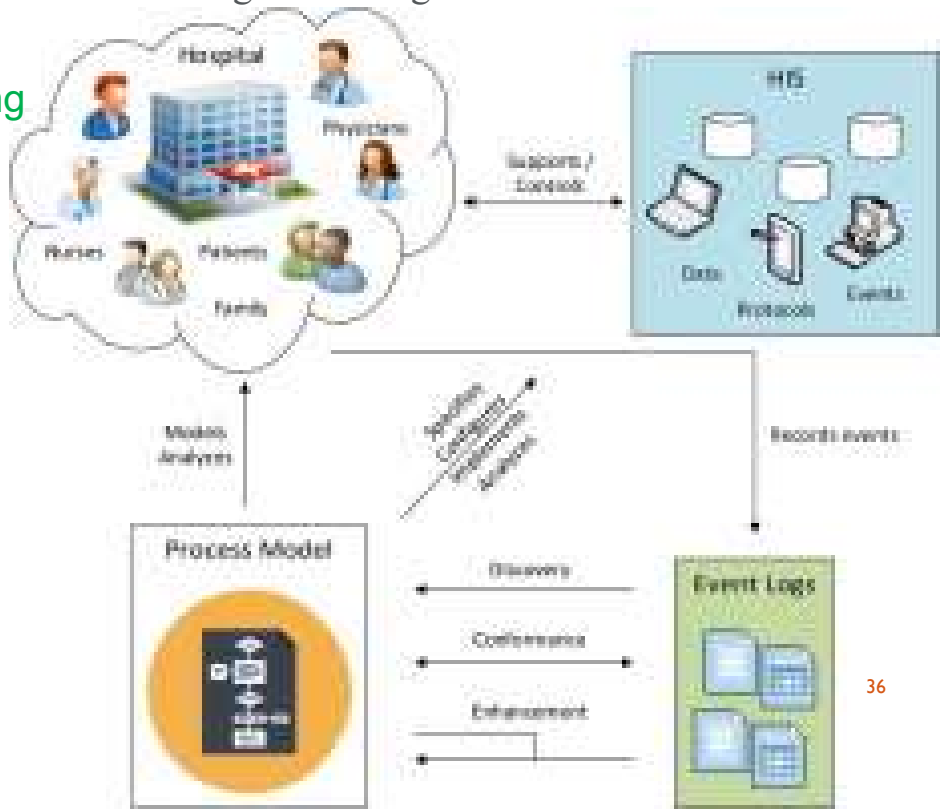
Applications of Healthcare Big Data Analytics

Benefit #4: Improved staffing through health business management analytics

Labor costs account for about 60% of all hospital budgets,

- the increasing demand for healthcare staff is expected to drive salaries and wages even higher in the future.

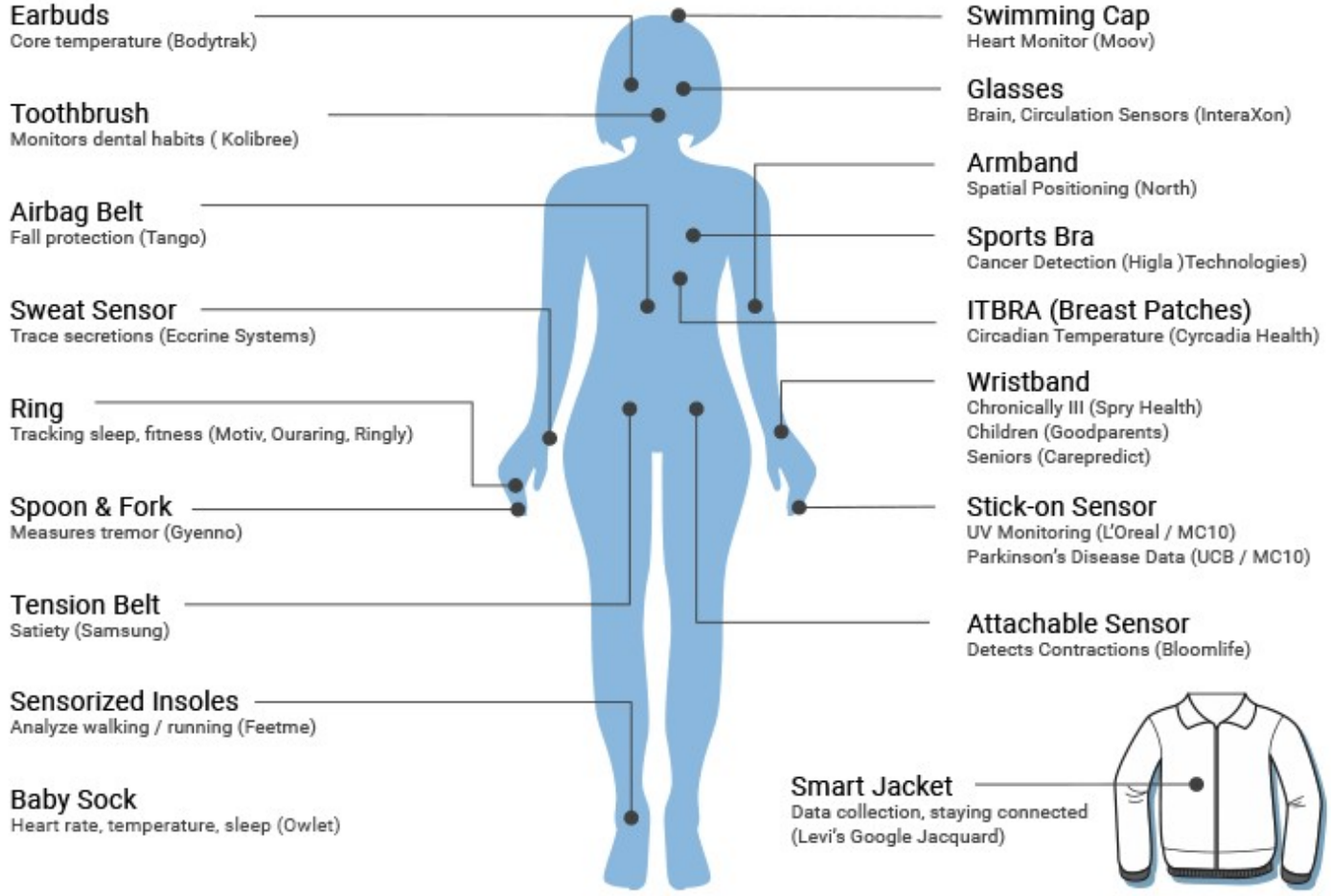
Process Mining



Data analytics is being applied to manage labor costs in healthcare settings while simultaneously improving the quality of care patients receive and the efficiency of service provision.

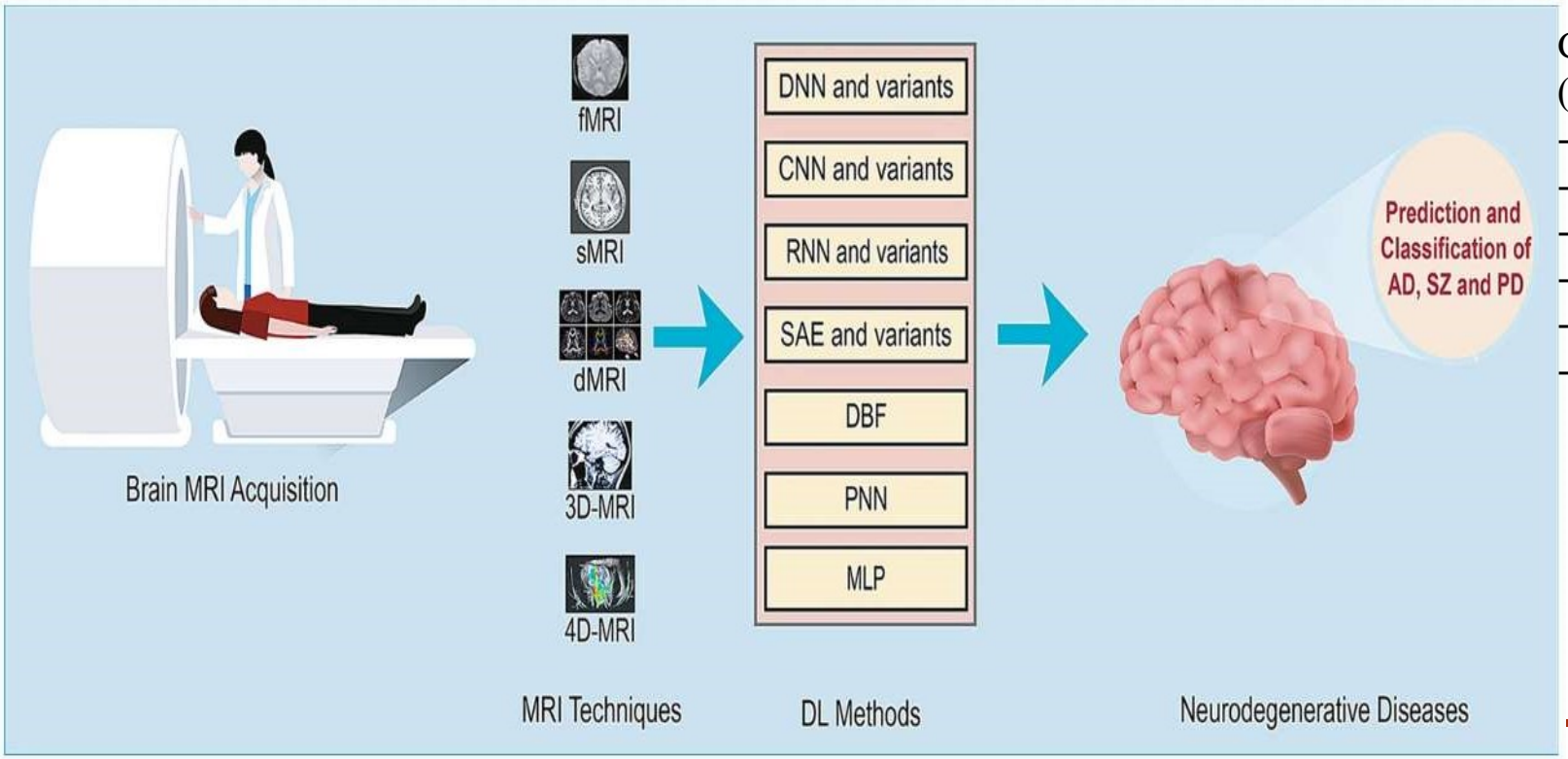
For example, Hawaii Pacific Health (HPH), one of the largest healthcare systems in Hawaii, was able to save $2.2 million over 16 months while maintaining high-quality outcomes by adopting a data-driven approach to labor management.

Humans Expected to Reach 1,400+ Digital Device Interactions by 2020



Source: In Vivo |Pharma Intelligence, Vol. 36 No. 3. Digital Health: Leveraging Data to Power and Personalize the Patient Experience. March 2018.

COGNITIVE DISORDERS

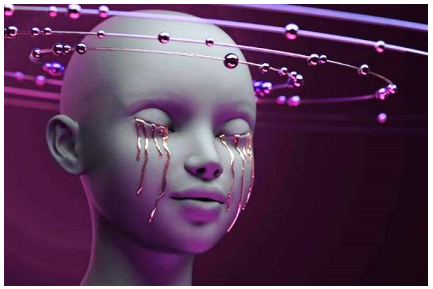


Overview of DL-based prediction and classification pipeline of neurodegenerative disease from different variants of MRI

Cognitive disorders diagnosis (EEG, MRI, fMRI):

- Alzheimer
- ADHD
- Seizure
- Sleep Disorders
- Parkinson
- Etc.

MENTAL HEALTH



The mental health and well-being of the WHO European Region has been hit hard by several large-scale public health emergencies over the past few years, including the COVID-19 pandemic, the war in Ukraine and an increasing cost-of-living crisis

Digital health and mental health are 2 of the 4 flagships under the European Programme of Work 2020–2025,

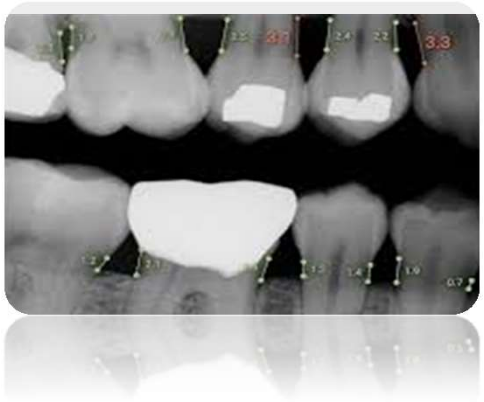
AI and Big Data analytics are seen as novel tools in the planning of mental health services as well as identifying and monitoring mental health problems in individuals and populations

AI-driven tools can harness readily available, real-time data – such as that generated through social media and electronic health records – to effectively plan and allocate resources for mental health services, identify and prevent misinformation related to public health concerns, develop targeted communications to promote behaviour change and predict and intervene early in mental ill health.

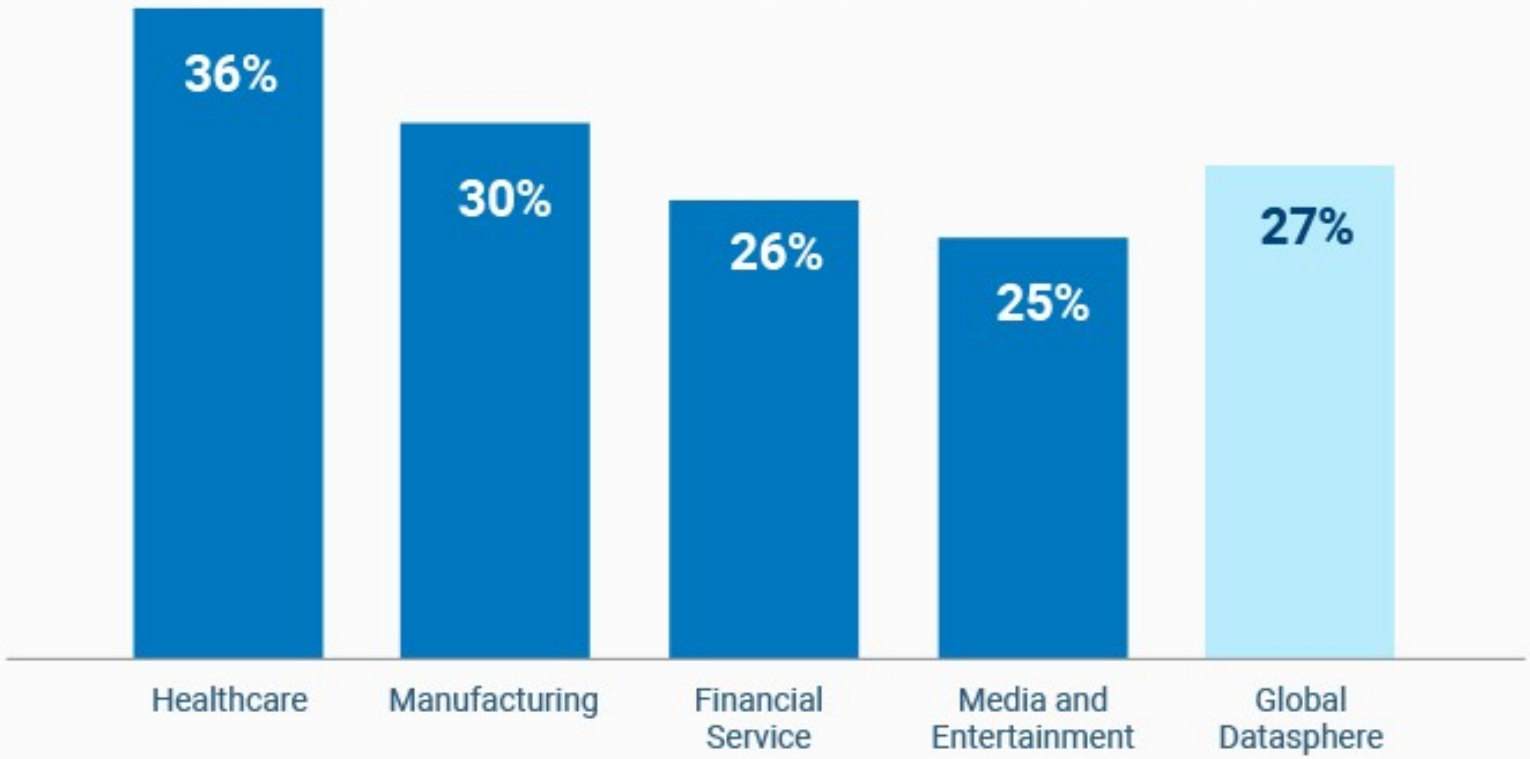
https://www.who.int/europe/news-room/events/item/2022/12/07/default-calendar/big-data-analytics-and-artificial-intelligence-in-mental-health

DENTISTRY

Dental assistant for diagnosis of carries from Bitewing images

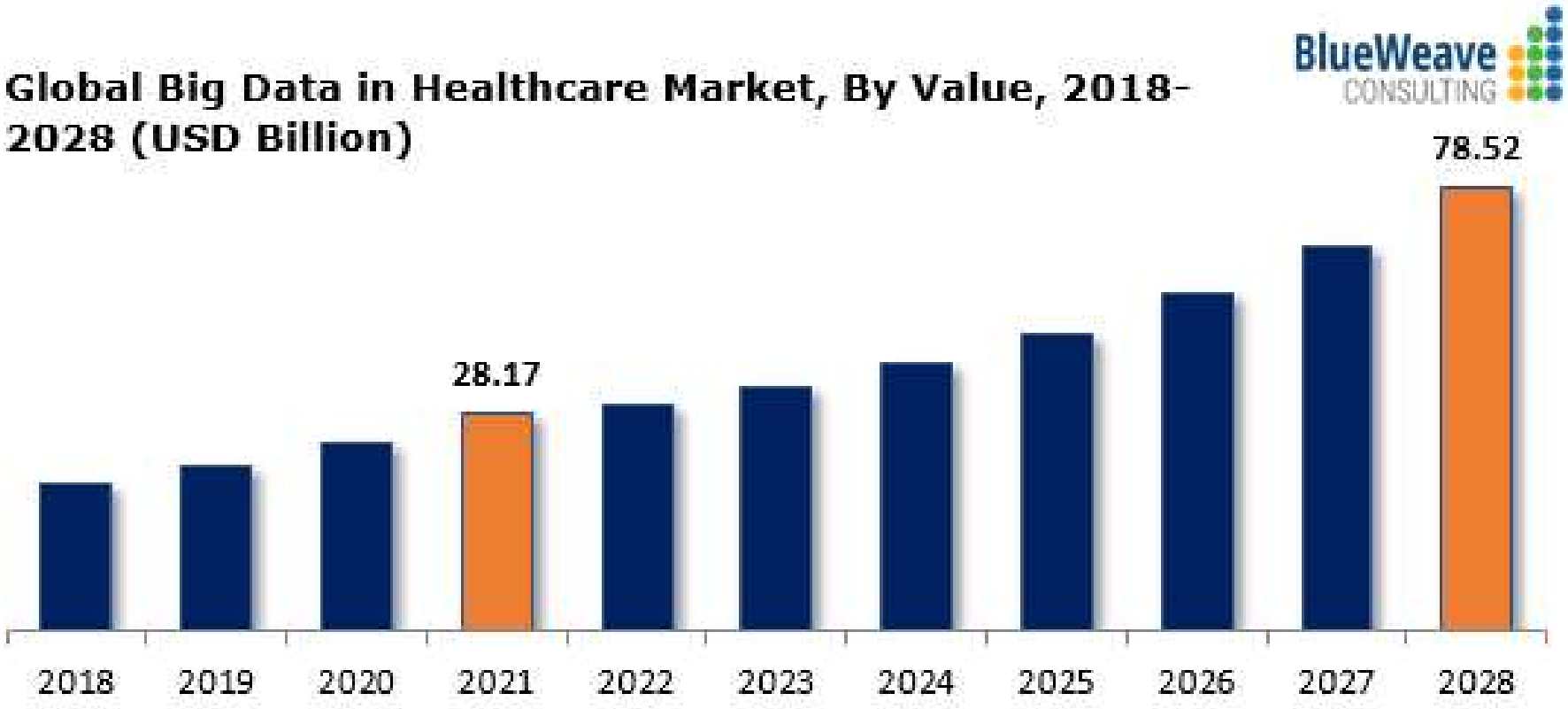


2018-2025 Data – Compound Annual Growth Rate (CAGR)



Source: Coughlin et al Internal Medicine Journal article "Looking to tomorrow's healthcare today: a participatory health perspective". IDC White Paper, Doc# US44413318, November 2018: The Digitization of the World – From Edge to Core".

BIG DATA IN HEALTHCARE MARKET



Source: BlueWeave Consulting

?