The state of the art and taxonomy of big data analytics: view from new big data framework

Prof Dr Hajah Azlinah Hj Mohamed
Principal Fellow
Institute for Big Data and Analytics (IBDAAI),
Universiti Teknologi MARA
azlinah@uitm.edu.my
016-3388840

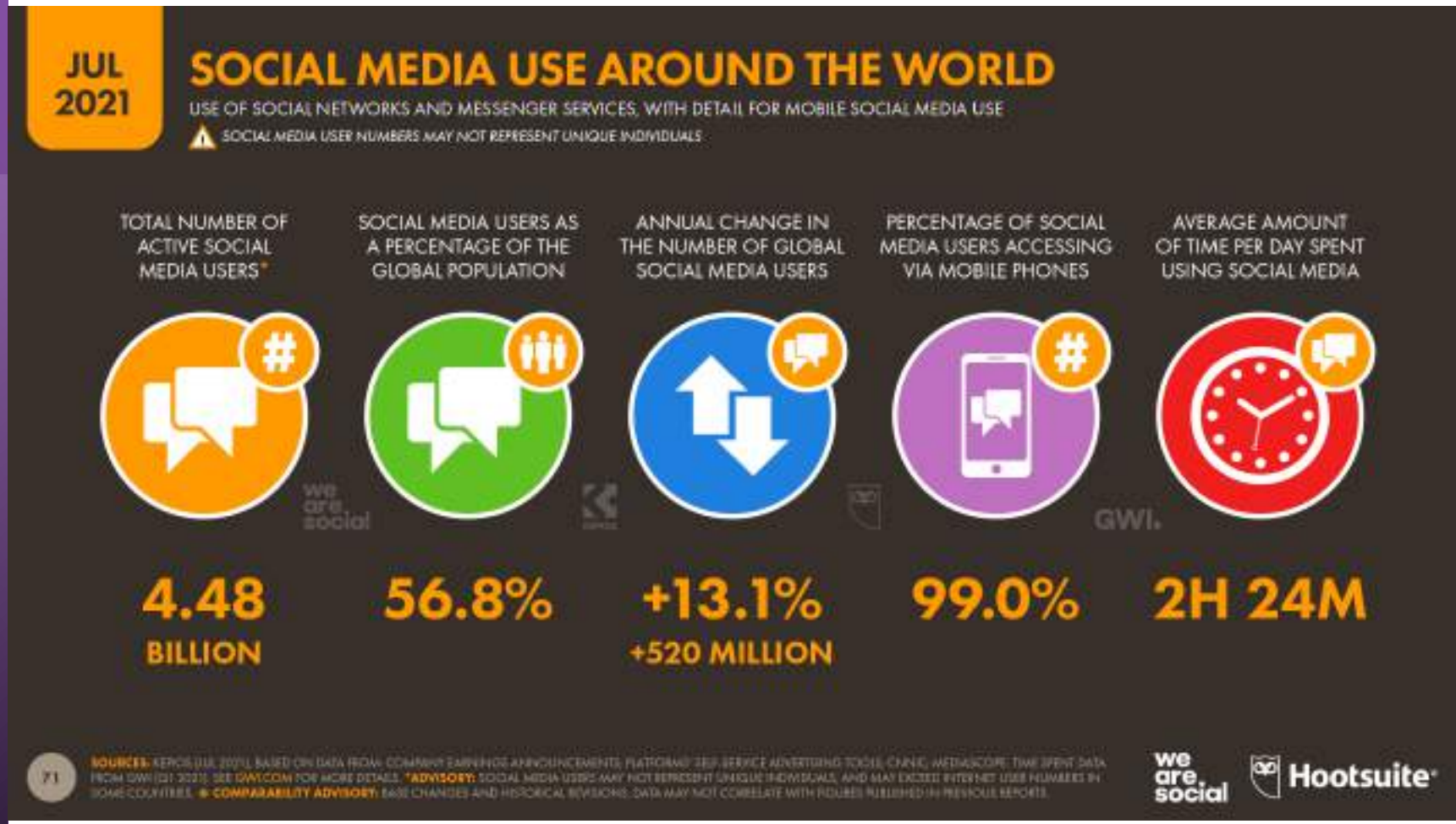
Data Never Sleeps

How much data is generated *every minute*?

Source : Domo.com, as of 2020

Reports by dataportal.com

Digital2021 : Worldwide

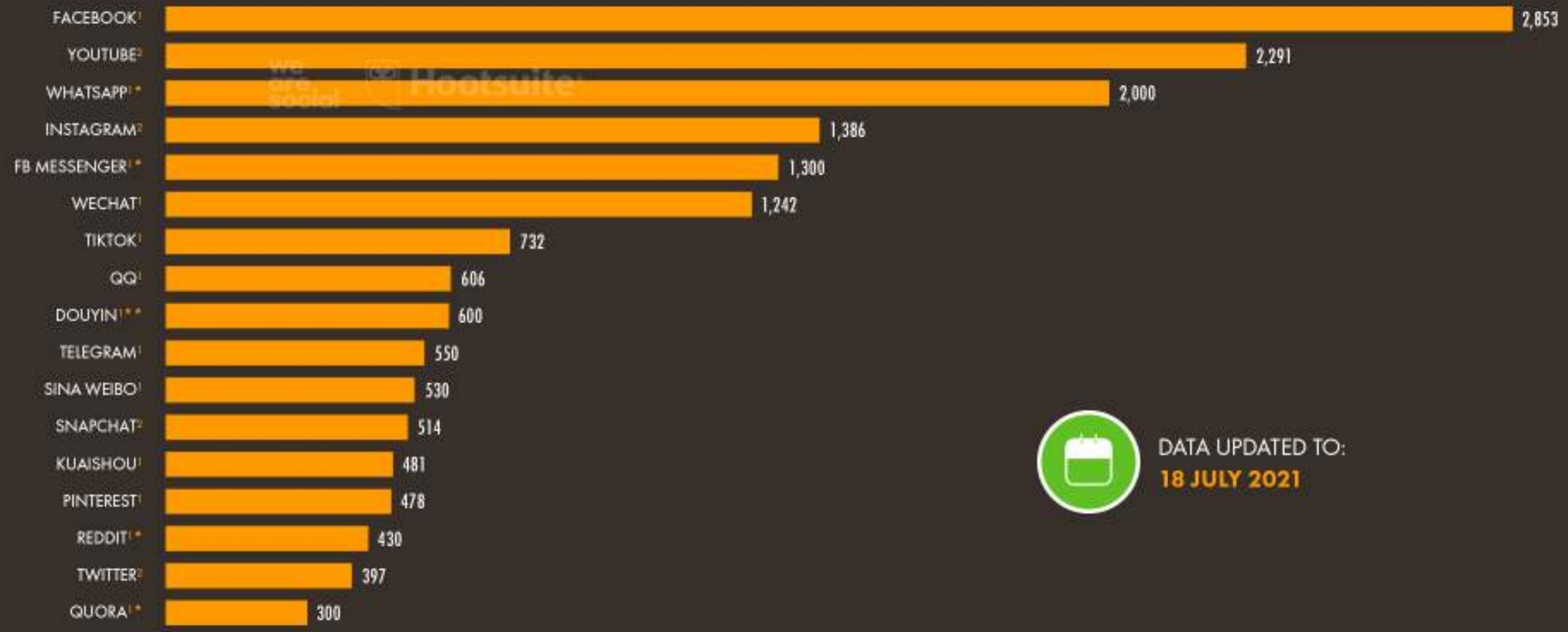


Source : datareportal.com, as of Jul 2021

JUL 2021

THE WORLD'S MOST-USED SOCIAL PLATFORMS

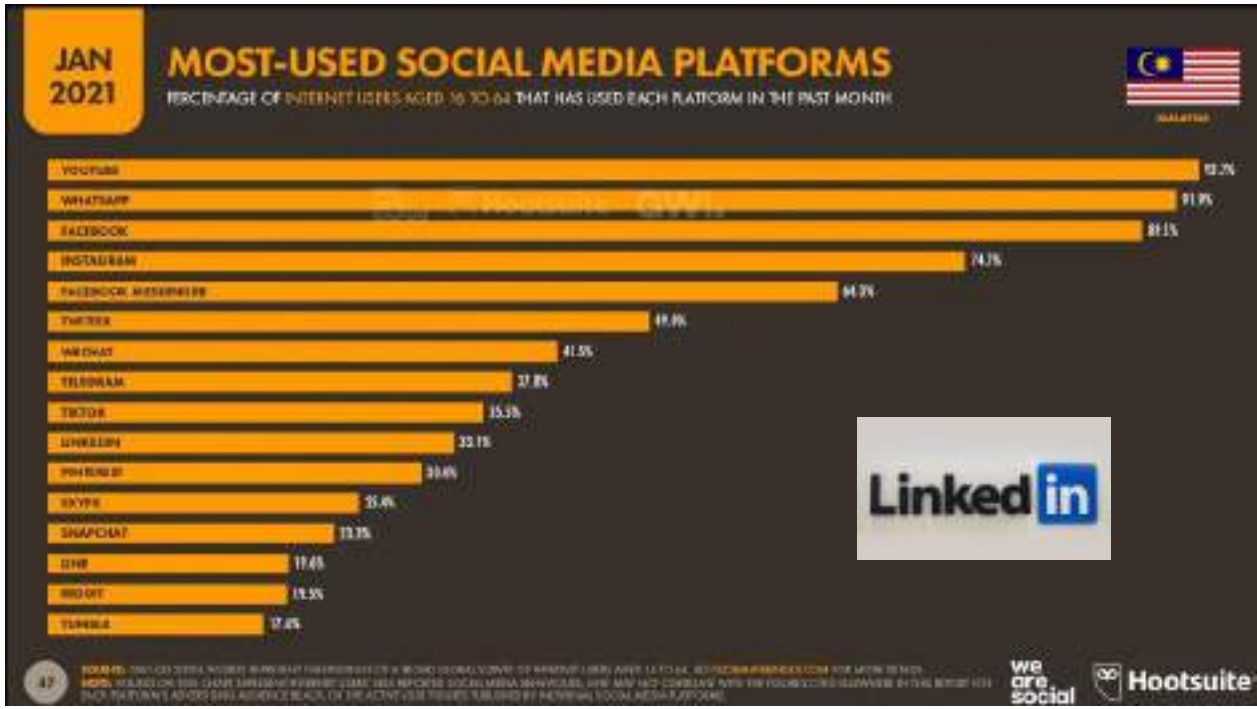
THE LATEST GLOBAL ACTIVE USER FIGURES (IN MILLIONS) FOR A SELECTION OF THE WORLD'S TOP SOCIAL MEDIA PLATFORMS*



DATA UPDATED TO:
18 JULY 2021

SOURCES: KEPIOS ANALYSIS (JUL 2021), BASED ON DATA PUBLISHED IN: (1) COMPANY STATEMENTS AND EARNINGS ANNOUNCEMENTS; (2) PLATFORMS' SELF-SERVICE AD TOOLS.
***NOTES:** 'USERS' MAY NOT REPRESENT UNIQUE INDIVIDUALS. PLATFORMS IDENTIFIED BY (*) HAVE NOT PUBLISHED UPDATED USER NUMBERS IN THE PAST 12 MONTHS, SO FIGURES WILL BE LESS RELIABLE. (**) FIGURE USES THE REPORTED DAILY ACTIVE USER FIGURE, SO MONTHLY ACTIVE USER FIGURE IS LIKELY HIGHER.

Digital2021 : Malaysia



Source : datareportal.com, as of Jan 2021

How does Economist Define New Terms As Commodity : Data, AI, IoT

Data is the new oil

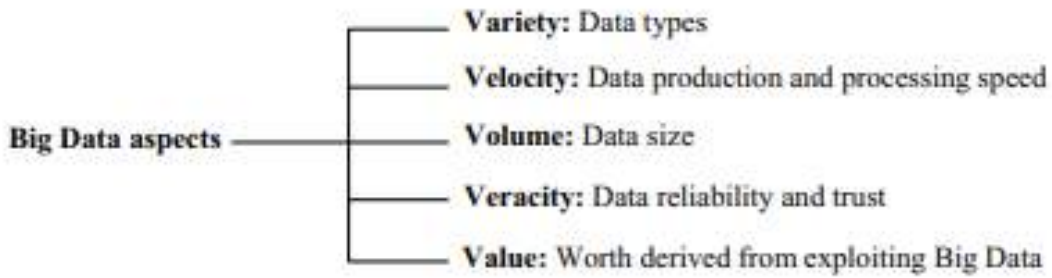
IOT is the new nervous system

AI is the new electricity

WHAT IS *BIG DATA*?

"The definition of big data is data that contains greater variety, arriving in increasing volumes and with more velocity. This is also known as the three Vs.

-Oracle"



(Mohammed,2020)

The V's Evolution of Big Data

9V's (Owais,2016)

Volume
Variety
Velocity
Veracity
Value
Visualization
Variability
Validity
Volatility

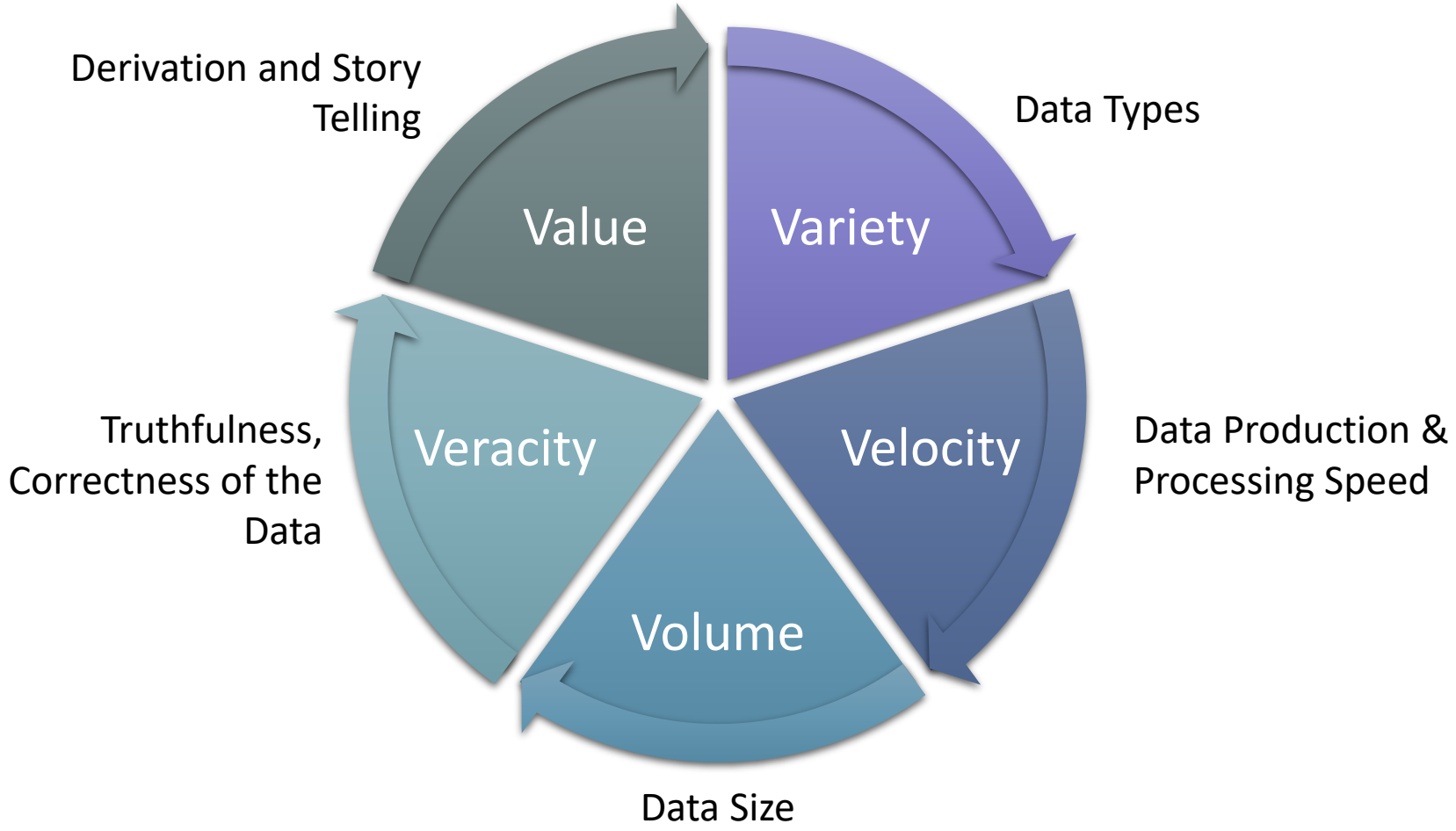
10V's (Data Science Central)

Volume
Variety
Velocity
Veracity
Value
Variability
Validity
Venue
Vocabulary
Vagueness

17V's (Panimalar,2017)

Volume
Variety
Velocity
Veracity
Value
Variability
Validity
Venue
Vocabulary
Vagueness
Volatility
Visualization
Viscosity
Virality
Verbosity
Voluntariness
Versality

5 Main Characteristics (5V's) of Big Data



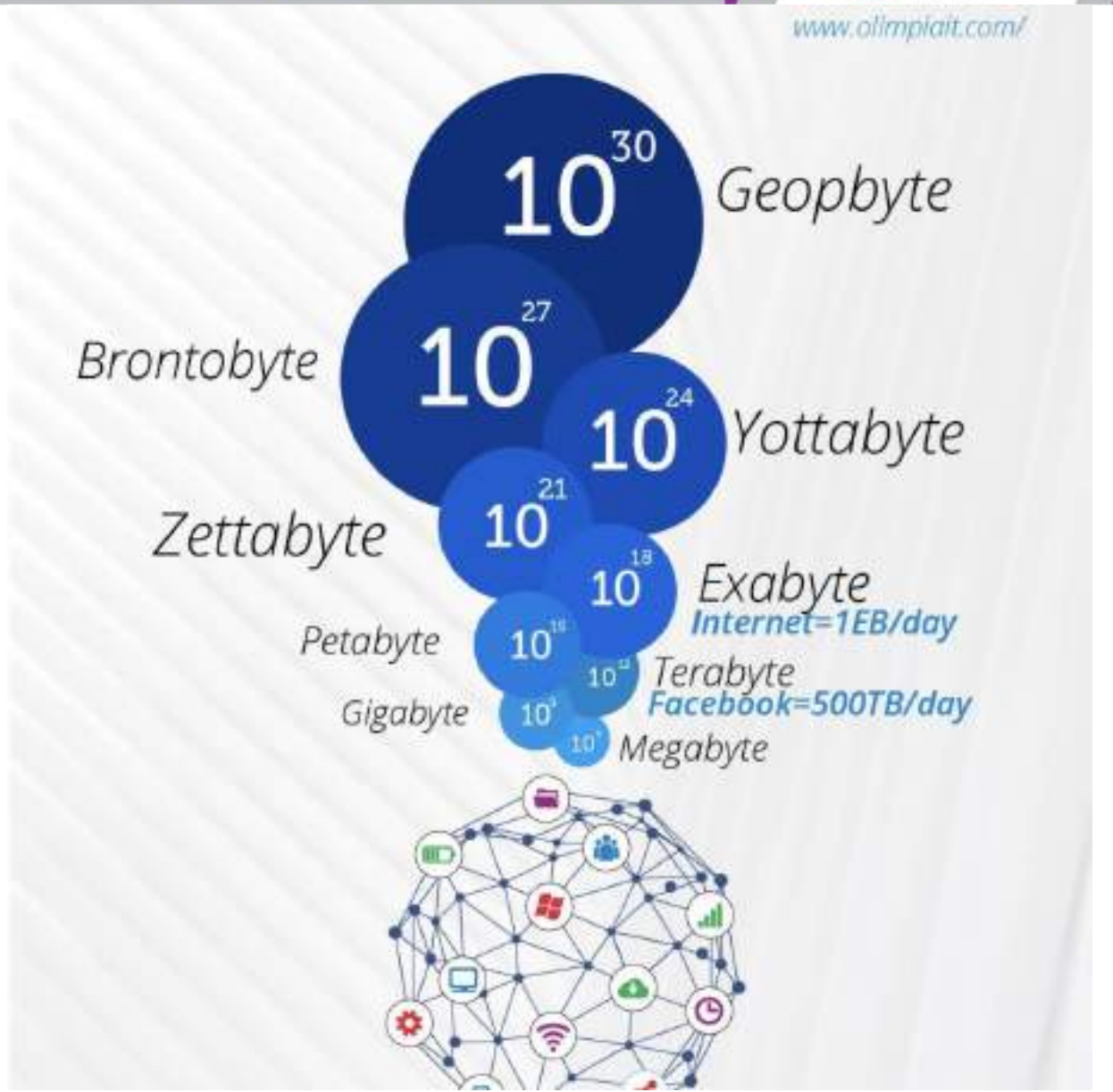
How Big is *Big Data*

From Megabyte (10^6) of data to Brontobyte (10^{27}) and Geopbyte (10^{30}), these measurements will be used to describe the tremendous amount of digital pool formed by the IoT platform.

Cisco-IBSG predicts about more than 50 billion devices connected to the internet by 2020, 75 billion IoT Devices by 2025.

By 2025, it's estimated that 463 exabytes of data will be created each day globally – that's the equivalent of 212,765,957 DVDs per day.

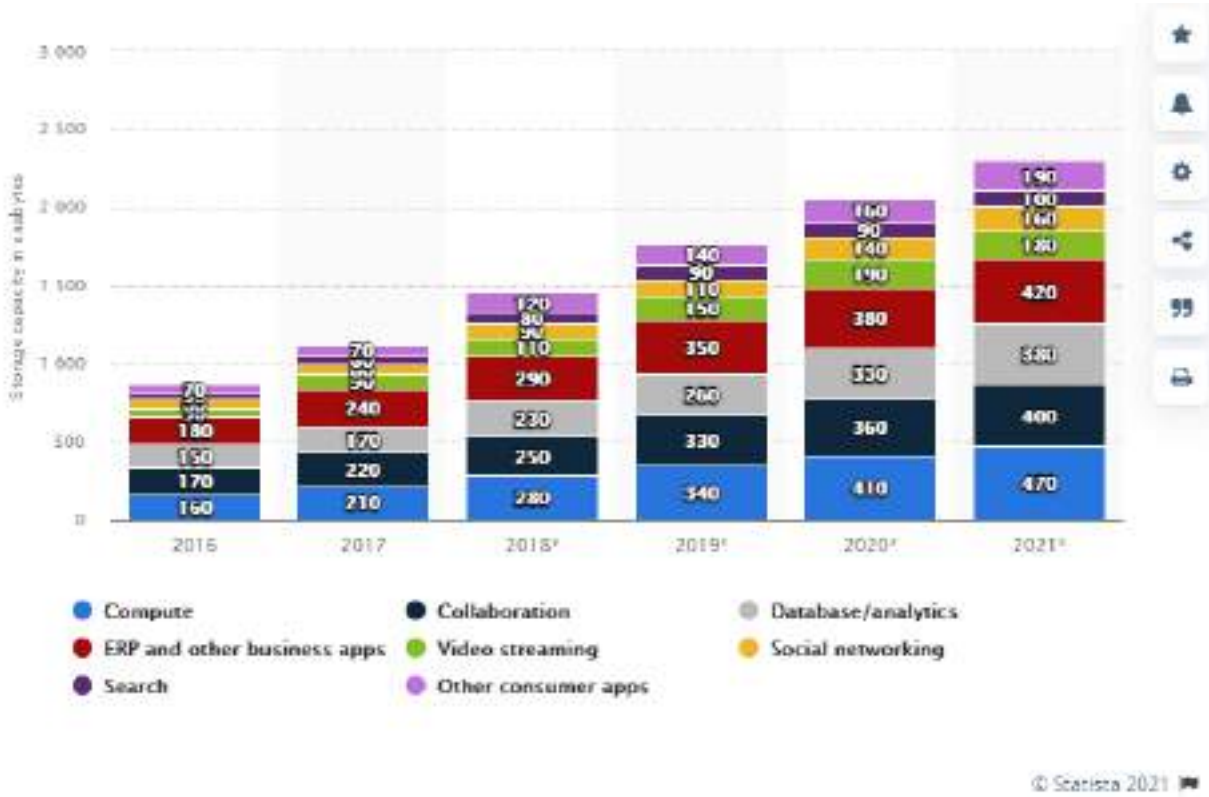
-Almiani,2020



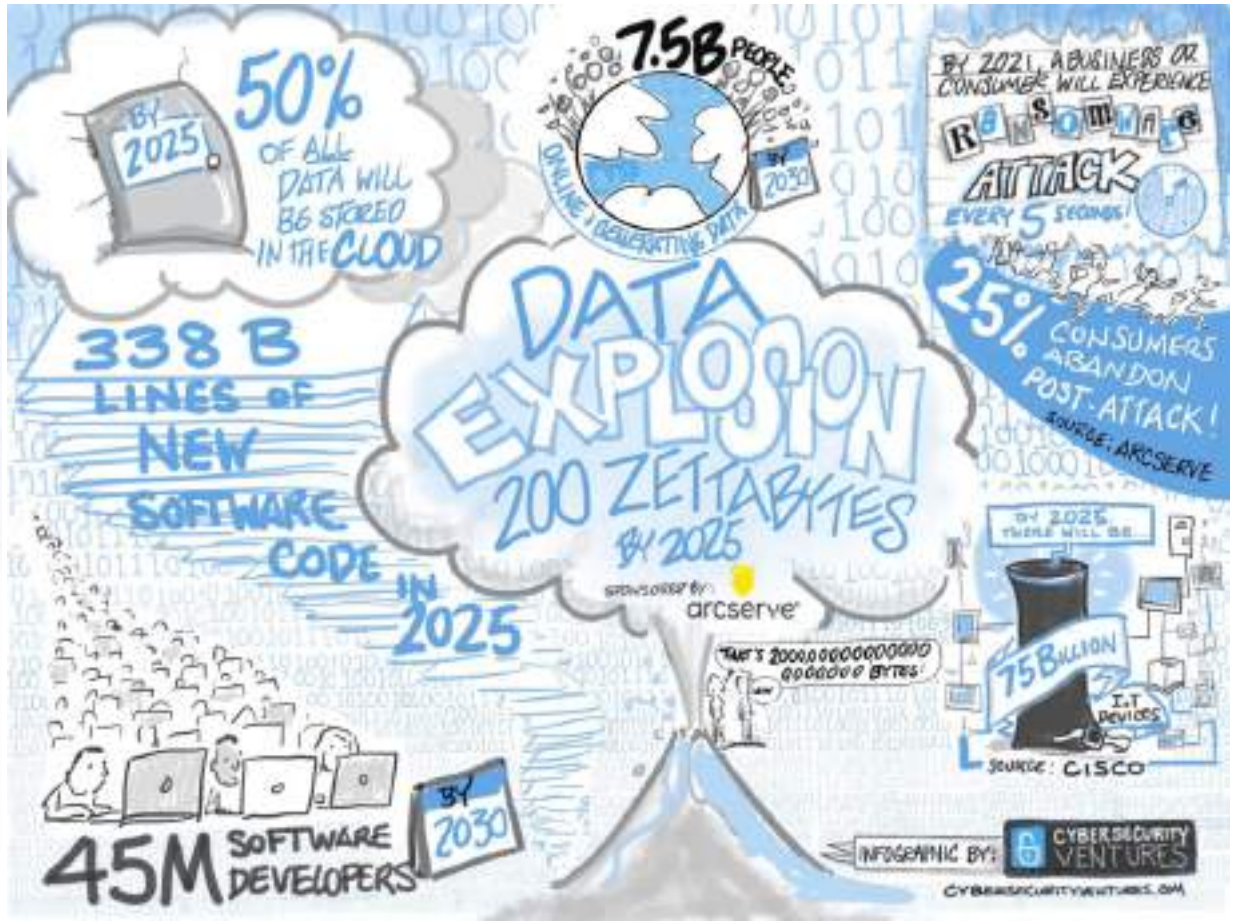
Where is it Now?
Current Big Data User

- NSA (National Security Admission – US Dept. of Defense, Intelligent Agency)
  - House a Yottabyte of Data (R. Herbert)
  - Utah Data Center will intercept, decipher, analyze, and store vast swathes of the world's communications as they zap down from satellites and zip through the underground and undersea cables of international, foreign, and domestic networks (Wired Magazine)
- Google
  - Expected to shift from Petabyte to Exabyte using Big Data Technology

Data Center Storage Capacity Worldwide From 2016 To 2021, By Segment



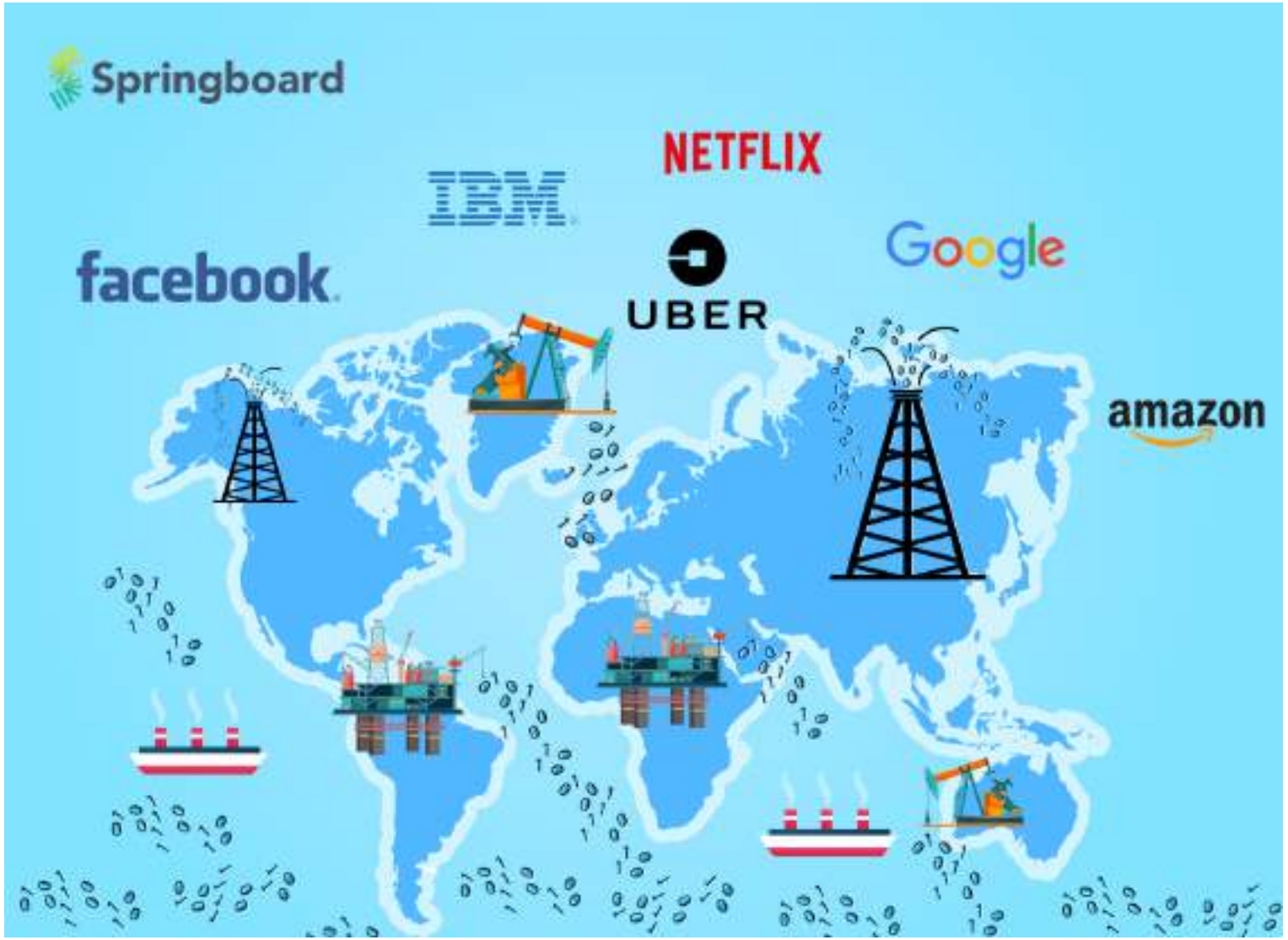
Source : statistica.com, as of 2021



Source : Cybersecurityventures.com, as of June 2020

"Data is the new Oil"

Clive Humby



MONETIZING BIG DATA

Use data to see patterns for business understanding

Use data to optimize the business

Use data to address business challenges

Use data for Storytelling

Use data to gather better data

Adapted : www.gartner.com

Monetizing Data : 3 Big Data Statistics

1 Big data impact on savings and profits

Source: TechJury, Tractica, Entrepreneur, Grazziti

$1 TRILLION
Savings by businesses through IoT by 2020.

$1 BILLION
Saved by Netflix using big data to improve customer retention.

8–10 %
Increased profits by businesses that use big data.

$119 BILLION
Big data global revenue by 2025.

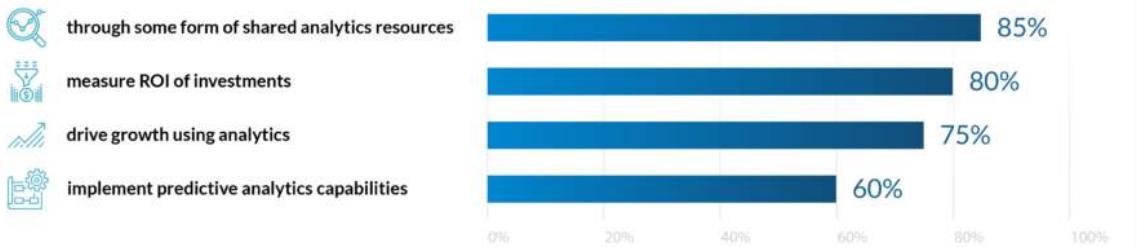
IBM

NETFLIX

Google

2 How do businesses leverage big data?

Source: iView System



3 Top benefits of data analytics

Sources: Chicago Analytics Group

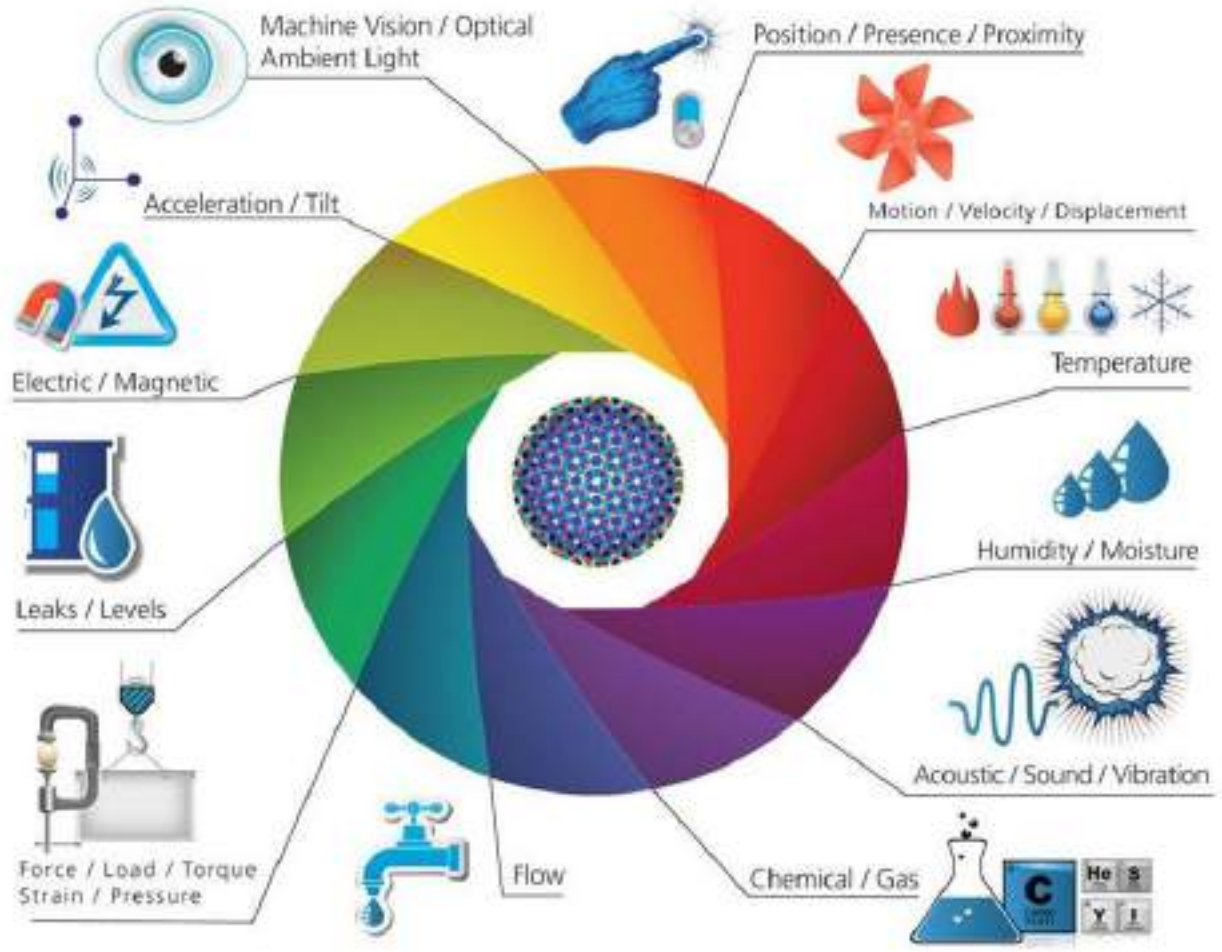
Source : financesonline.com

“AI is the new electricity”

Andrew Ng

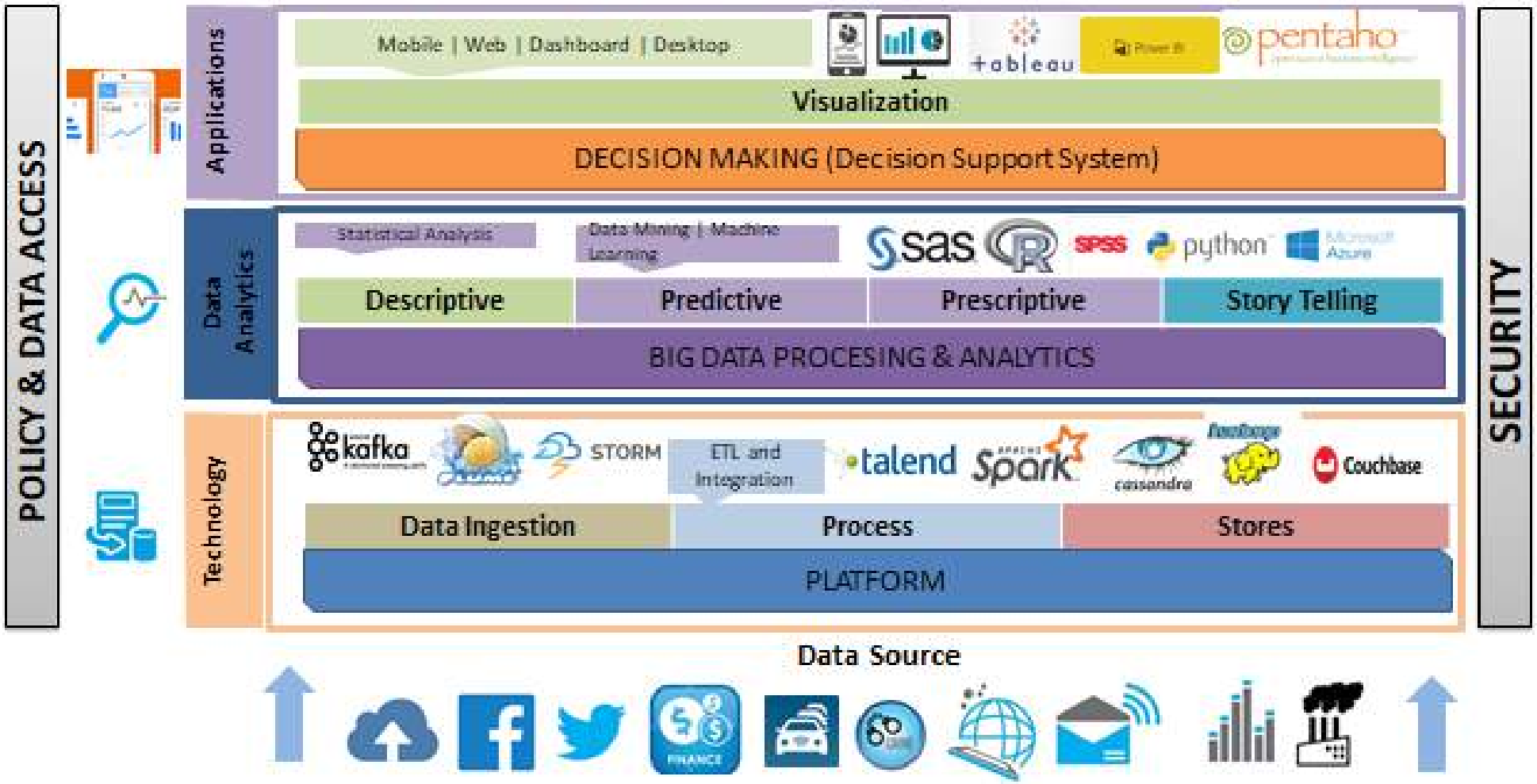
“IoT is the new nervous system”

Harbor Research

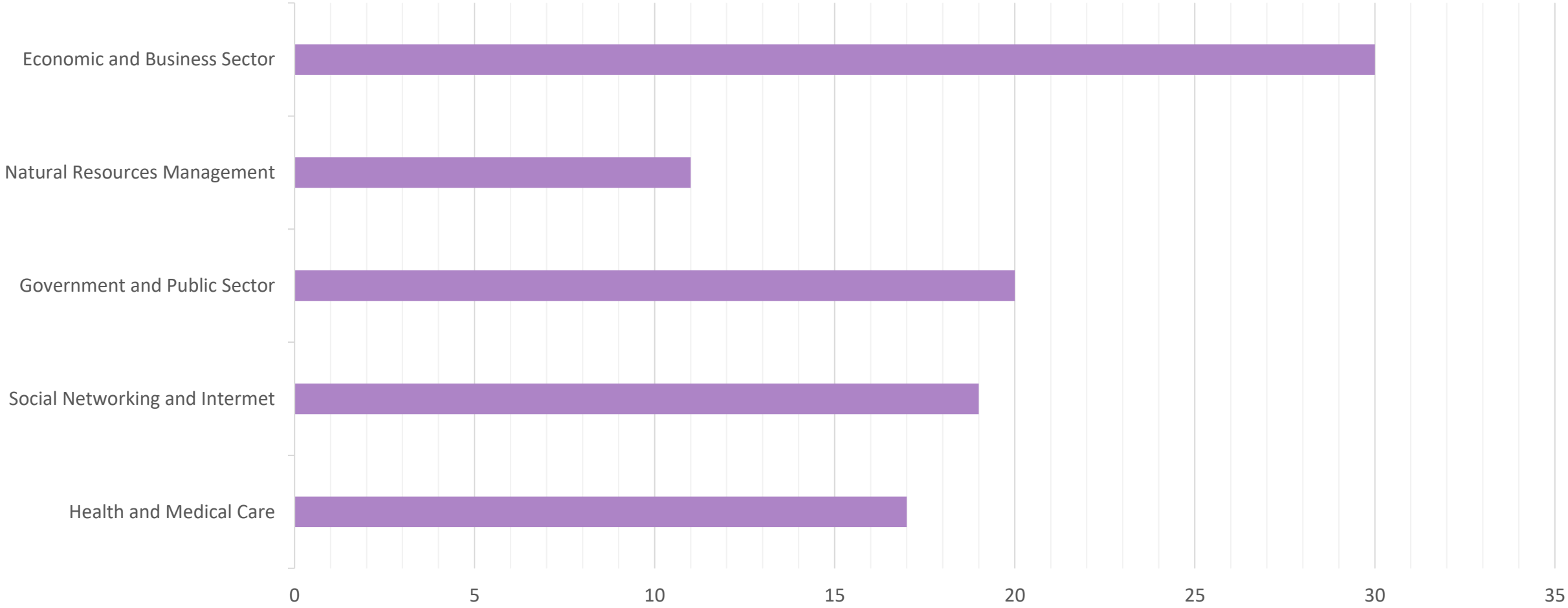


Source : Harborresearch.com as of, 2020

BIG DATA FRAMEWORK©



DOMAIN AREA


Economic and Business Sector
Natural Resources Management
Government and Public Sector
Social Networking and Intermet
Health and Medical Care
0
5
10
15
20
25
30
35

DATA SOURCES & DATA FORMAT

Data Sources

Internet

Social Media/ Multimedia Data

Sensors

Web Services

Other Devices

Data Format

Structured Data
- Enterprise Data
- CRM
- ERP

Semi Structured Data
- XML/JSON
- EDI
- Transaction

Unstructured Data
- Documents
- Videos
- Machine Sensors
- Social Media

BIG DATA TOOLS & TECHNIQUES

DATA PROCESSING & STORAGE

Google

facebook

Batch Processing

- Hadoop / MapReduce
- Apache Pig
- Flume

hadoop

Hybrid Processing

Stream Processing

- Storm
- S4
- Kafka
- Flink
- Spark

Spark

YAHOO!

Alibaba.com

BIG DATA ANALYTICS TECHNIQUES

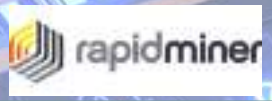
MACHINE LEARNING

1st Generation

WEKA, SAS, R, Python

- Deep Analytics- large collection of serial ML Algo
- Vertical Scaling scalibility
- Single Fault Tolerance

2nd Generation



Mahout, Rapid Miner, Pentaho

- Works over Hadoop Map-reduce
- Shallow analytics on Big Data.
- Small number of ML Algo

3rd Generation

Spark MLlib, Haloop, Apache Giraph, Pregel, Storm

- Deep Analytics on Big Data
- Realize ML Algorithm in real time – use Kafka-Storm integrated environment

CLOUD COMPUTING TECHNOLOGIES

Goal :

1. Explore the large quantity of data and extract useful information or knowledge for future actions.

2. Efficient data processing : Provide reasonable and dynamic computational architecture for large-scale and intricate enterprise applications.

Software-as-a-Service (SaaS)

- SaaS provides users with uninterrupted access to applications

Platform-as-a-Service (PaaS)

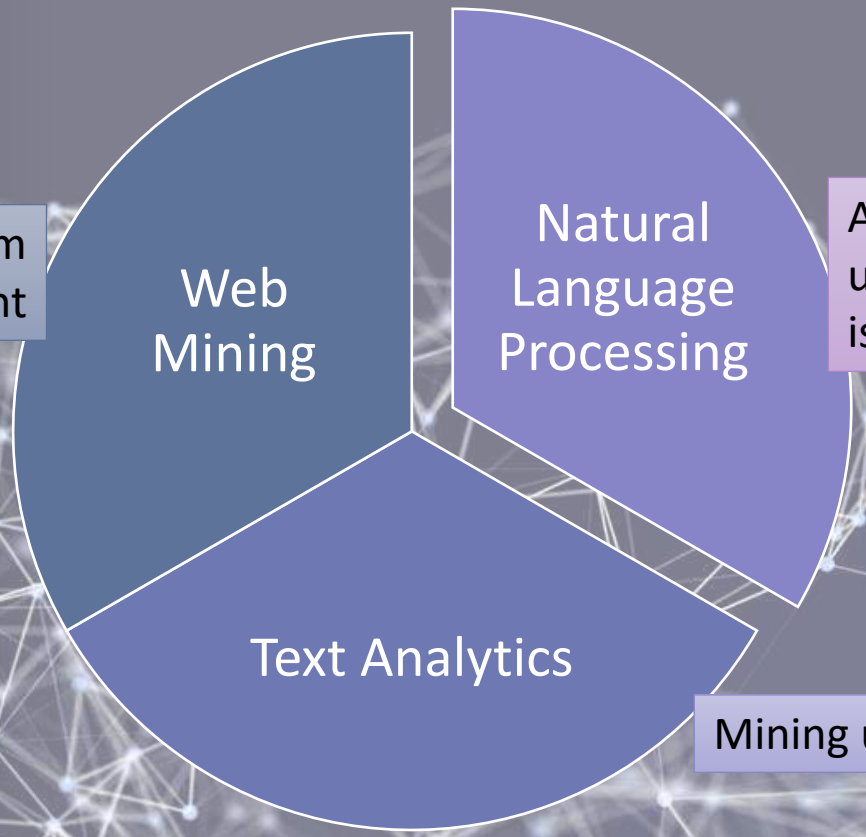
- PaaS helps users to develop, run, and manage applications

Infrastructure-as-a-Service (IaaS)

- IaaS offers users with access to a pool of configurable computing resources like network, storage and servers.

SEMANTIC NETWORK ANALYSIS / WEB MINING

Semantic network analysis is a technique employed to determine a pattern from large network repositories.



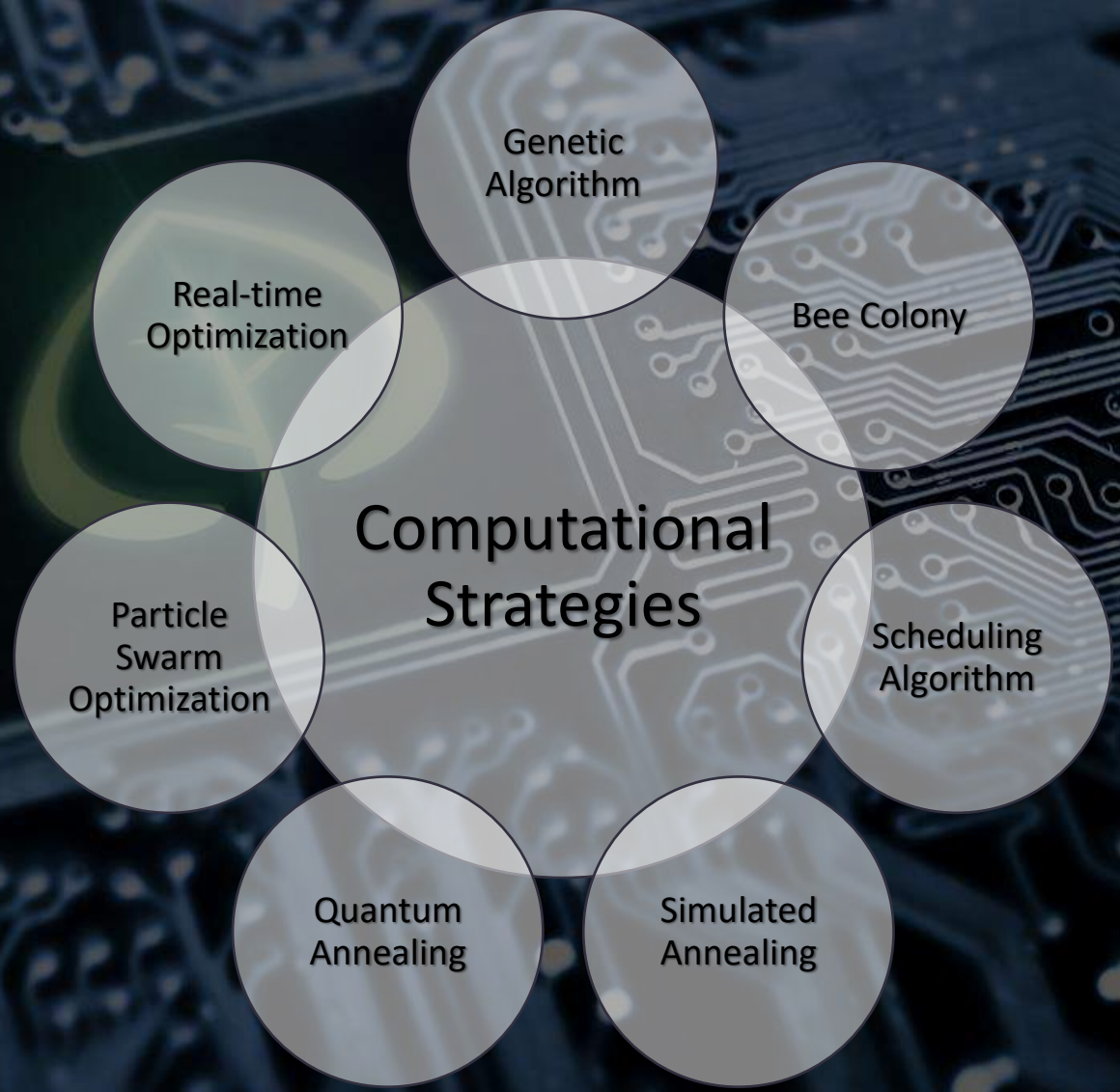
Mine useful information from the web content

Ability of a computer program to understand human language as it is spoken

Mining useful information from text sources.

OPTIMIZATION TECHNIQUES

Solve Quantitative Problems – Multidisciplinary Fields



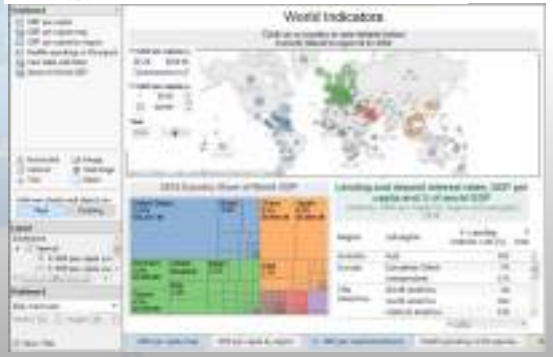
DATA VISUALIZATION

Quantify the effect of future decisions in order to advise on possible outcomes before the decisions are actually made.

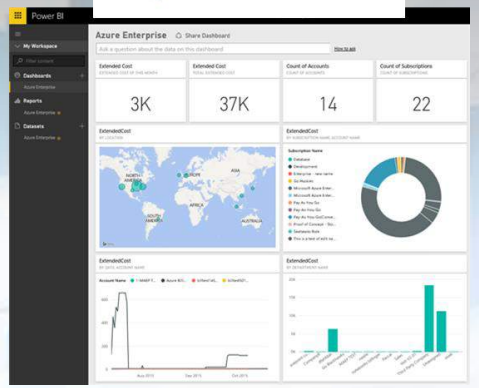
- Narrative + Data = Explain
- Data + Visual = Enlighten
- Visual + Narrative = Engage

ORACLE BI

+ableau



Microsoft Azure



kibana

sas

jupyter

FUSIONEX®
experience.excellence

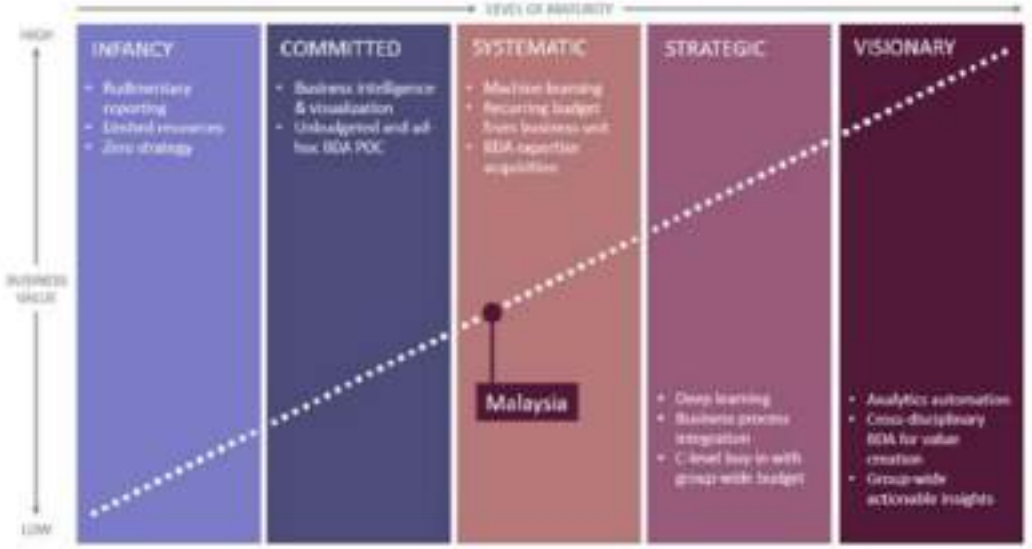
BIG DATA ANALYTICS : MALAYSIA VIEW

The big data analytics (BDA) market in Malaysia is expected to grow to US$1.9 billion (about RM7.85 billion) in 2025, from US$1.1 billion in 2021, according to Malaysia Digital Economy Corp's (MDEC) commissioned study by IDC.

Source : The Edge Markets,as of 17 May 2021

Source : medium.com,as of Feb 2020

MALAYSIA AI BLUEPRINT 2020 ANNUAL REPORT



Malaysia's BDA maturity (overall)

Source :bigittechnology.com, as of 2020

BIG DATA ANALYTICS : MALAYSIA VIEW

TECHNOLOGY DIGITALEDGE

Select Language

DigitalEdge

Cover Story: How smart are our cities?

Tan Zhai Yun / The Edge Malaysia
April 05, 2021 00:00 am +08

"In Malaysia, we can say that in achieving the first and second goals, we are at four out of 10. We're probably at level 2 for evidence-based decision-making using real-time data. For the fourth goal, I think we are at zero," says Hamdan.

Better data governance and sharing is crucial for many of these goals to be achieved, he believes. Additionally, "technology can be a great equaliser. There needs to be a broader approach to how smart cities are adopted in the country", he adds.

Think City uses technology to support evidence-based decision-making, whether it is for adaptive use of historical buildings, placemaking or urban analytics. For instance, it created a tech-enabled safety auditing tool to identify zones in the city that have high safety risk. Last month, it released a study measuring temperature trends in Malaysian cities and the contributing factors.

Source : theedgemarket.com, as of Apr 2021

MOH has Introduced Open-Sourcing Our Extremely Rich Data On The Pandemic at GitHub Platform

GitHub
MoH-Malaysia

https://github.com/MoH-Malaysia/covid19-public

With this, we hope that we can broaden our analysis & leverage on intelligence from all corners of society & sites like OurWorldInData pull from our repository & feature our data.

With our vaccination program being one of the fastest in the world, we must look beyond cases. In the repository, you will find anonymous granular data on hospital admissions, critical care (ICU), & even MOH's cluster-based analysis of transmission. In addition, we are also making public data on mobility & contact tracing, powered by MySejahtera. The mobility data even provides insight on the time density of check-ins; something that most other mobility datasets do not contain.

Source : covid-19.moh.gov.my, as of Jul 2021

IBDAAI INITIATIVES

IBDAAI

Social Listening

Health Analytics

Oil and Gas

Data Drifting (Streaming Data)

Leveraging Big Data Analytics

Big Data : Future Direction

Gartner Top 10 Data and Analytics Trends,2020

- **Trend 1: Smarter, faster, more responsible AI**
- **Trend 2: Decline of the dashboard**
- **Trend 3: Decision intelligence**
- **Trend 4: X analytics**
- **Trend 5: Augmented data management**
- **Trend 6: Cloud is a given**
- **Trend 7: Data and analytics worlds collide**
- **Trend 8: Data marketplaces and exchanges**
- **Trend 9: Blockchain in data and analytics**
- **Trend 10: Relationships form the foundation of data and analytics value**

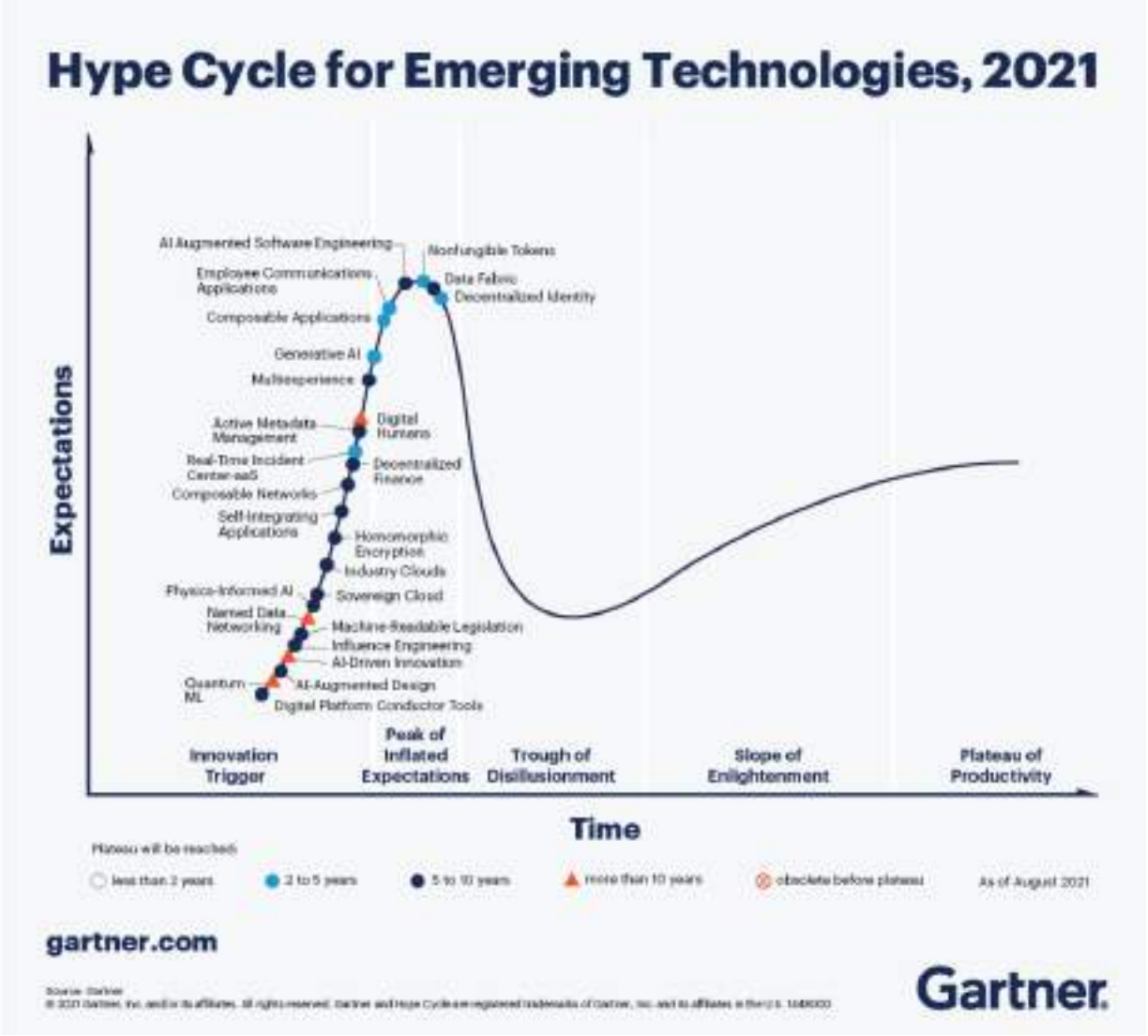
2021 Big Data Trends

“

- **AI drives deeper insights and increasingly sophisticated automation**
- **Rich new ways to explore and interpret data**
- **Hybrid cloud and the edge**
- **The rise of DataOps**

”

Forbes.com



BIG DATA : CHALLENGES TOWARDS DATA VALUE

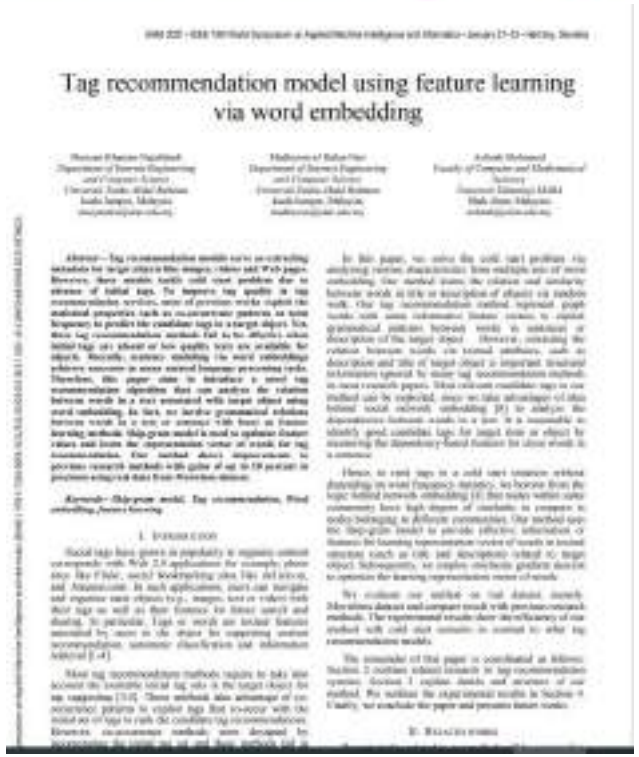
- Choosing Suitable Big Data Tools
- Real-time Data Stream Processing
  - Concept Drift
  - Online Learning
- Volatile Growth Of Data
- Imbalanced Data
- Architecture For Parallel Data Processing
- Quantum Computing for Big Data Analytics

THE 6th INTERNATIONAL CONFERENCE ON

SOFT COMPUTING IN DATA SCIENCE 2021

https://aaec.uitm.edu.my/scds2021/

The state-of-the-art and taxonomy of big data analytics: view from new big data framework



Tag recommendation model using feature learning via word embedding

THANK YOU