



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



IBDAAI

INSTITUTE FOR BIG DATA ANALYTICS
AND ARTIFICIAL INTELLIGENCE



UNIVERSITI
TEKNOLOGI
MARA

Data Never Sleeps

How much data is generated *every minute?*

Source : Domo.com, as of 2020



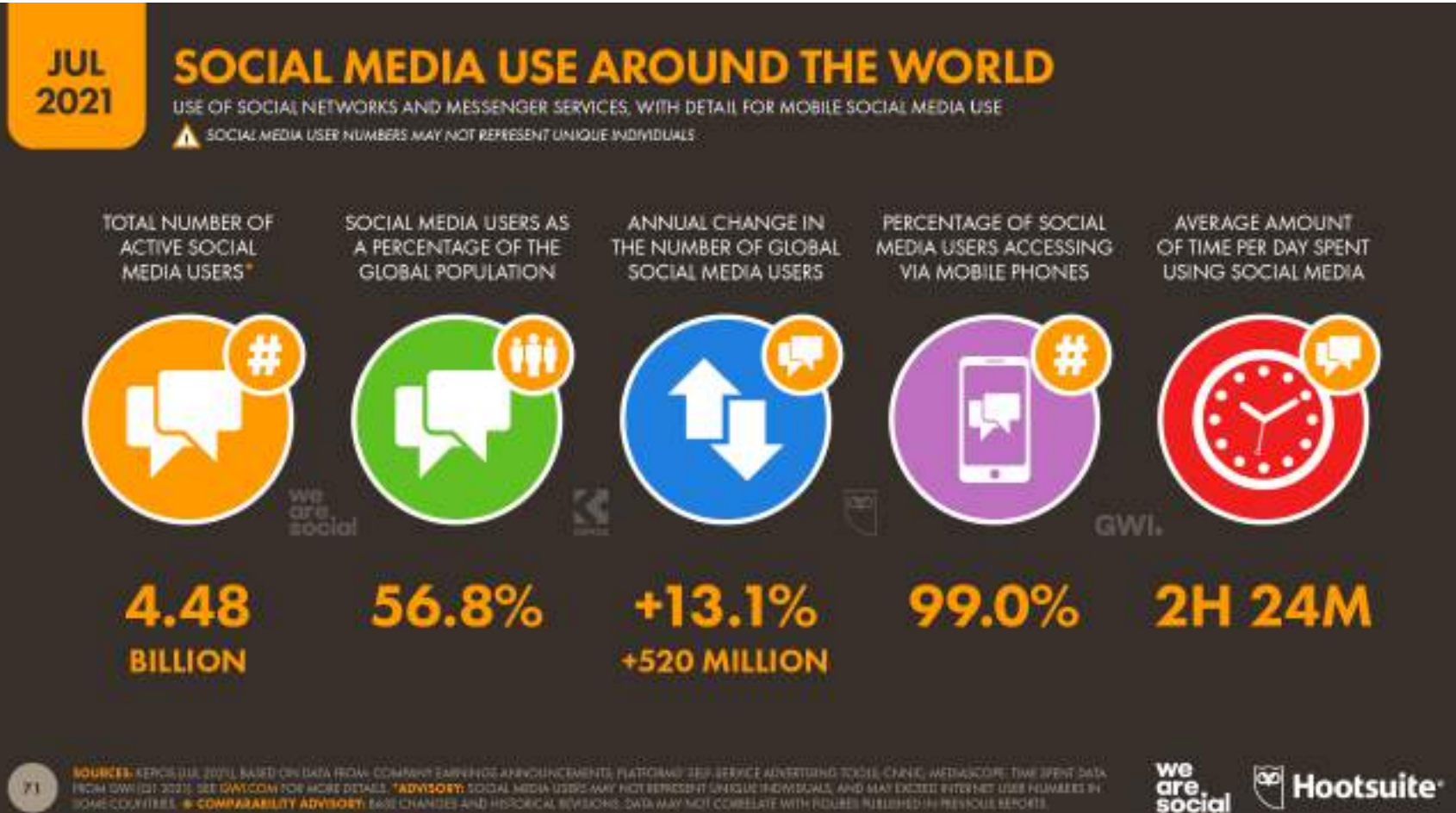
IBDAAI

INSTITUTE FOR BIG DATA ANALYTICS
AND ARTIFICIAL INTELLIGENCE



Reports by
dataportal.com

Digital2021 : Worldwide

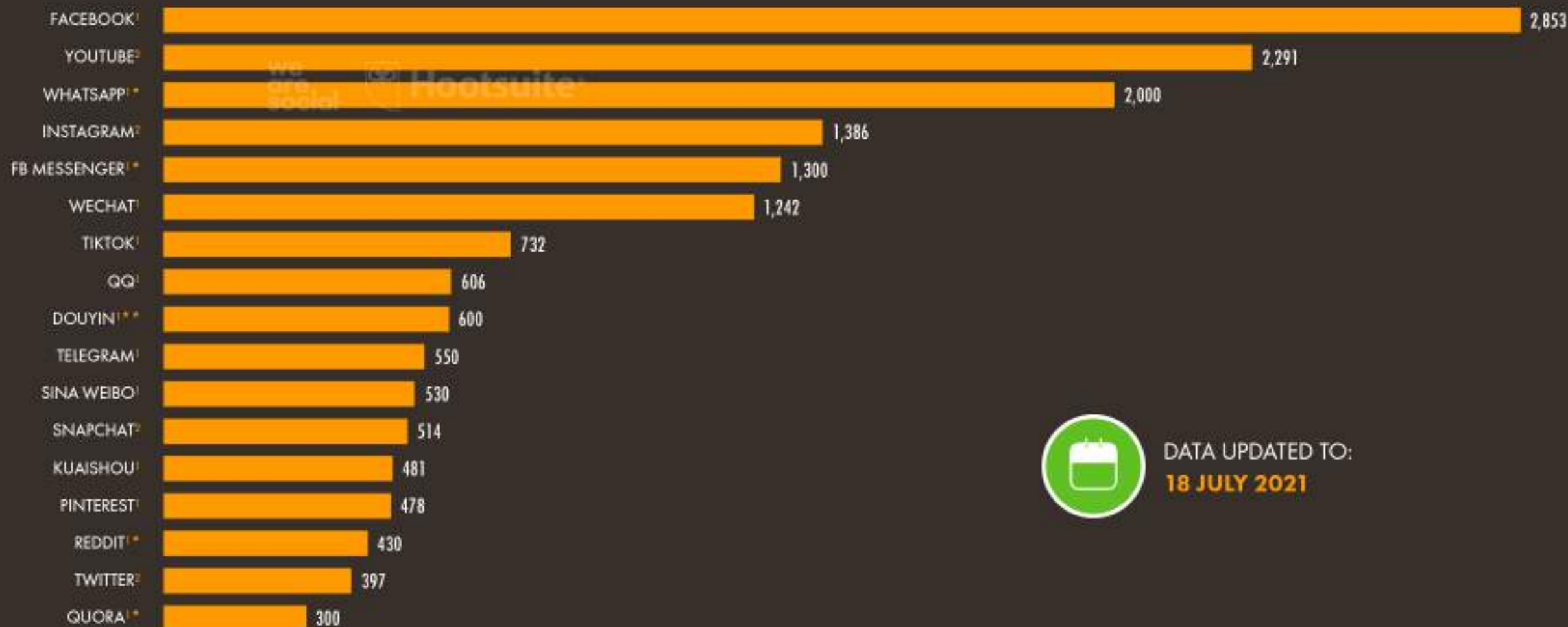


Source : datareportal.com, as of Jul 2021

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

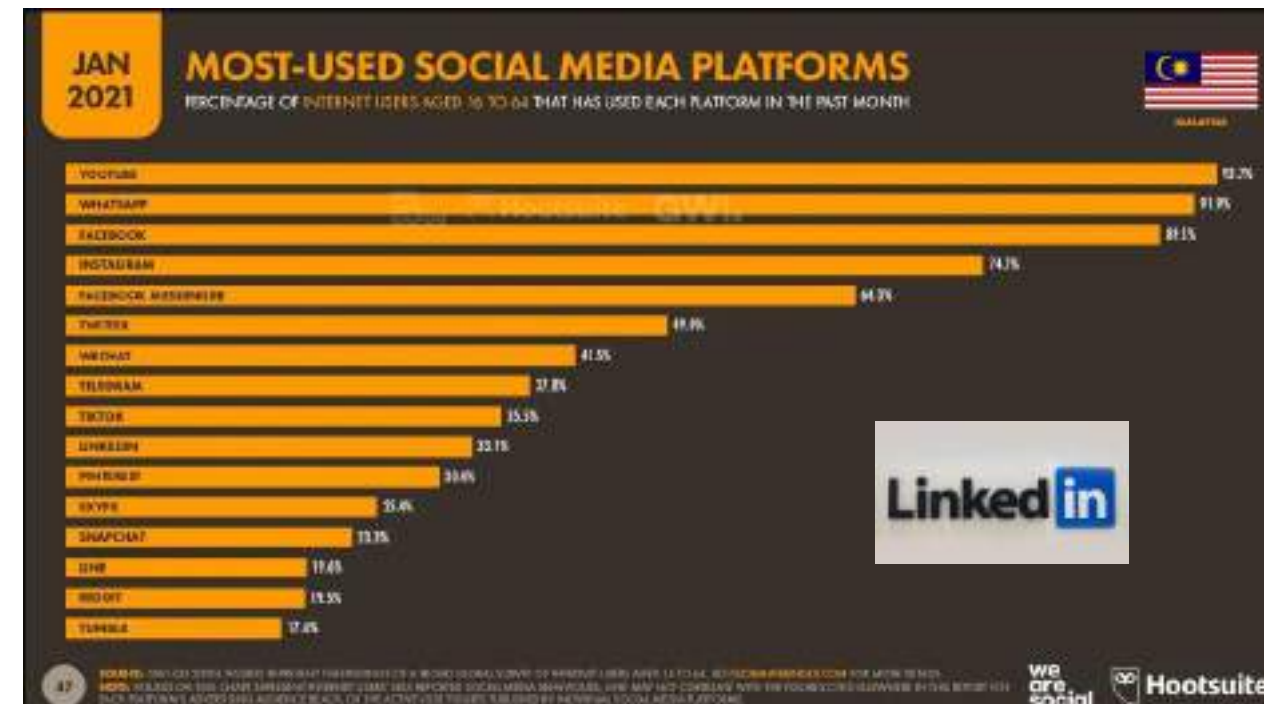
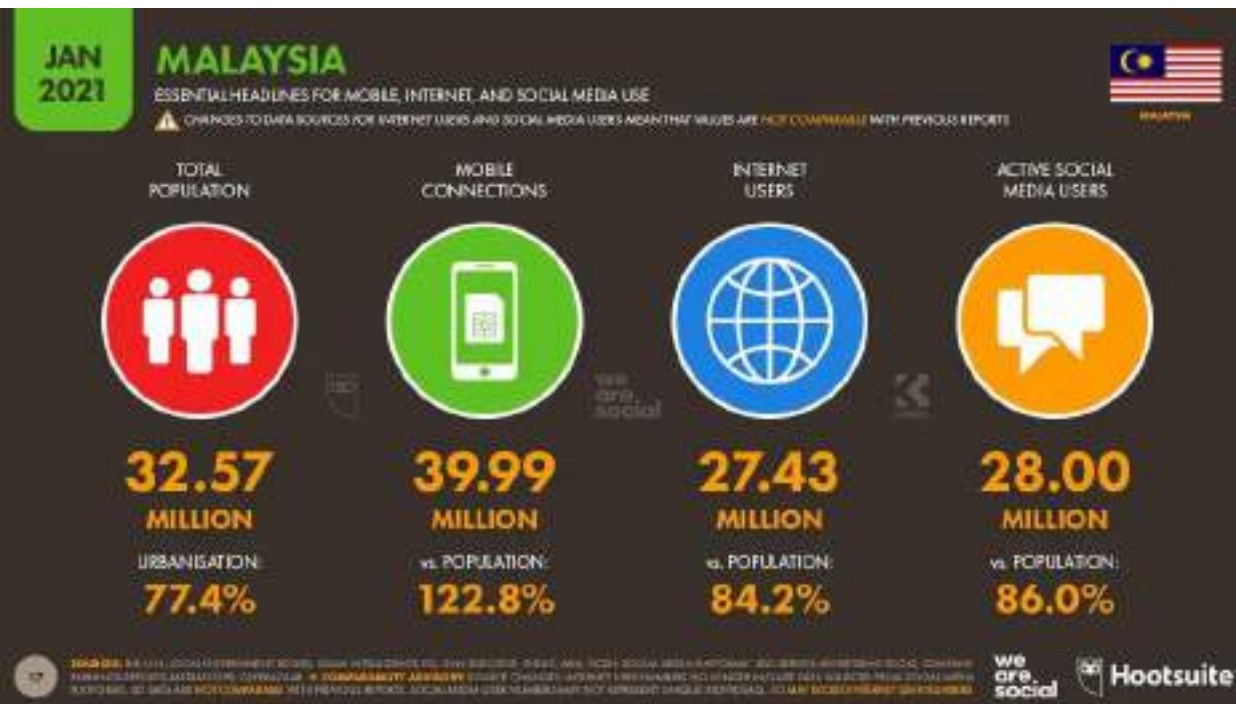


IBDAAI

INSTITUTE FOR BIG DATA ANALYTICS
AND ARTIFICIAL INTELLIGENCE



Digital2021 : Malaysia



Source : datareportal.com, as of Jan 2021



IBDAAI

INSTITUTE FOR BIG DATA ANALYTICS
AND ARTIFICIAL INTELLIGENCE



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





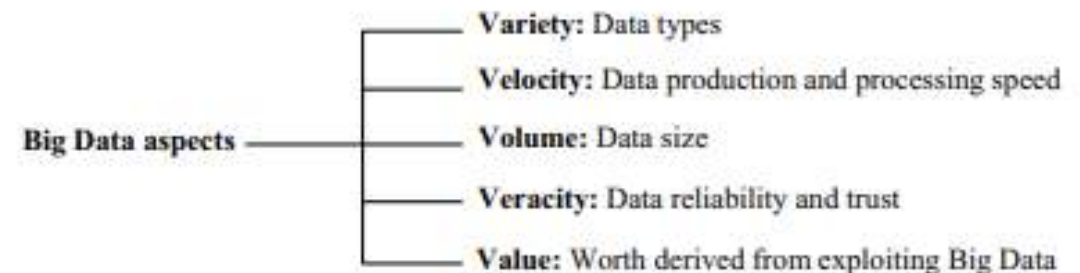
WHAT IS *BIG DATA*?

“

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)

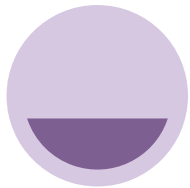


IBDAAI

INSTITUTE FOR BIG DATA ANALYTICS
AND ARTIFICIAL INTELLIGENCE

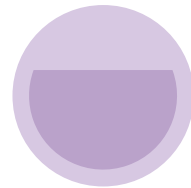


The V's Evolution of Big Data



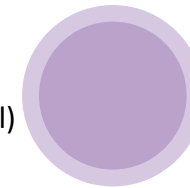
9V's (Owais,2016)

- 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

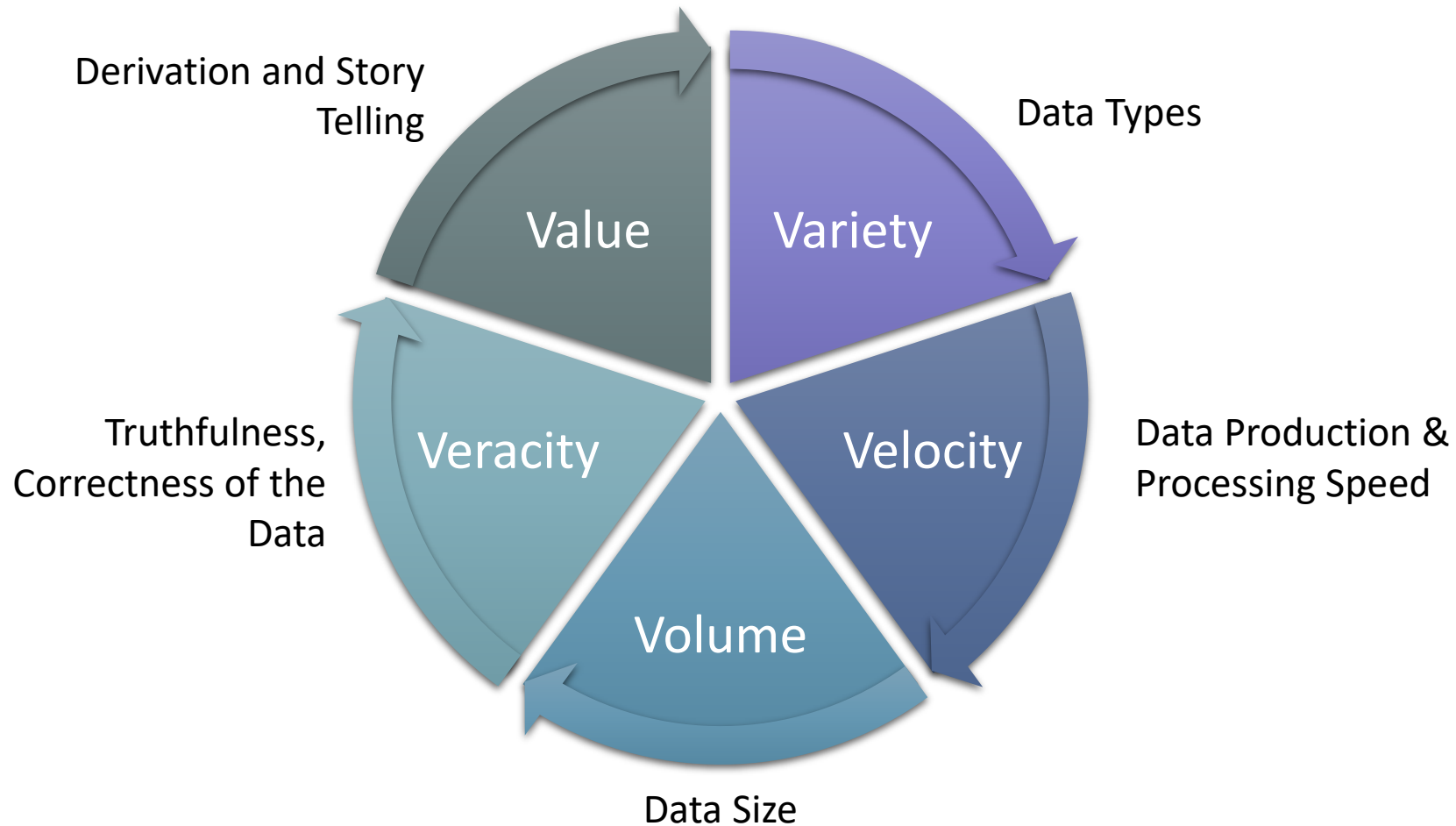


IBDAAI

INSTITUTE FOR BIG DATA ANALYTICS
AND ARTIFICIAL INTELLIGENCE



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





IBDAAI

INSTITUTE FOR BIG DATA ANALYTICS
AND ARTIFICIAL INTELLIGENCE



www.olimpiadit.com/

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

”





IBDAAI

INSTITUTE FOR BIG DATA ANALYTICS
AND ARTIFICIAL INTELLIGENCE



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



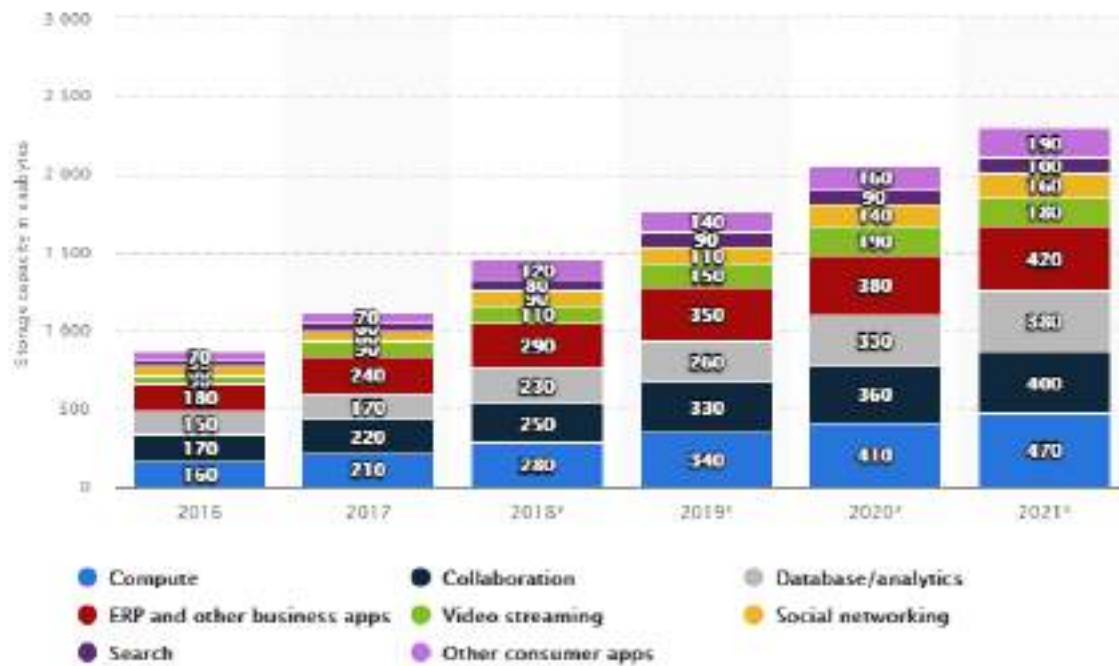


IBDAAI

INSTITUTE FOR BIG DATA ANALYTICS
AND ARTIFICIAL INTELLIGENCE

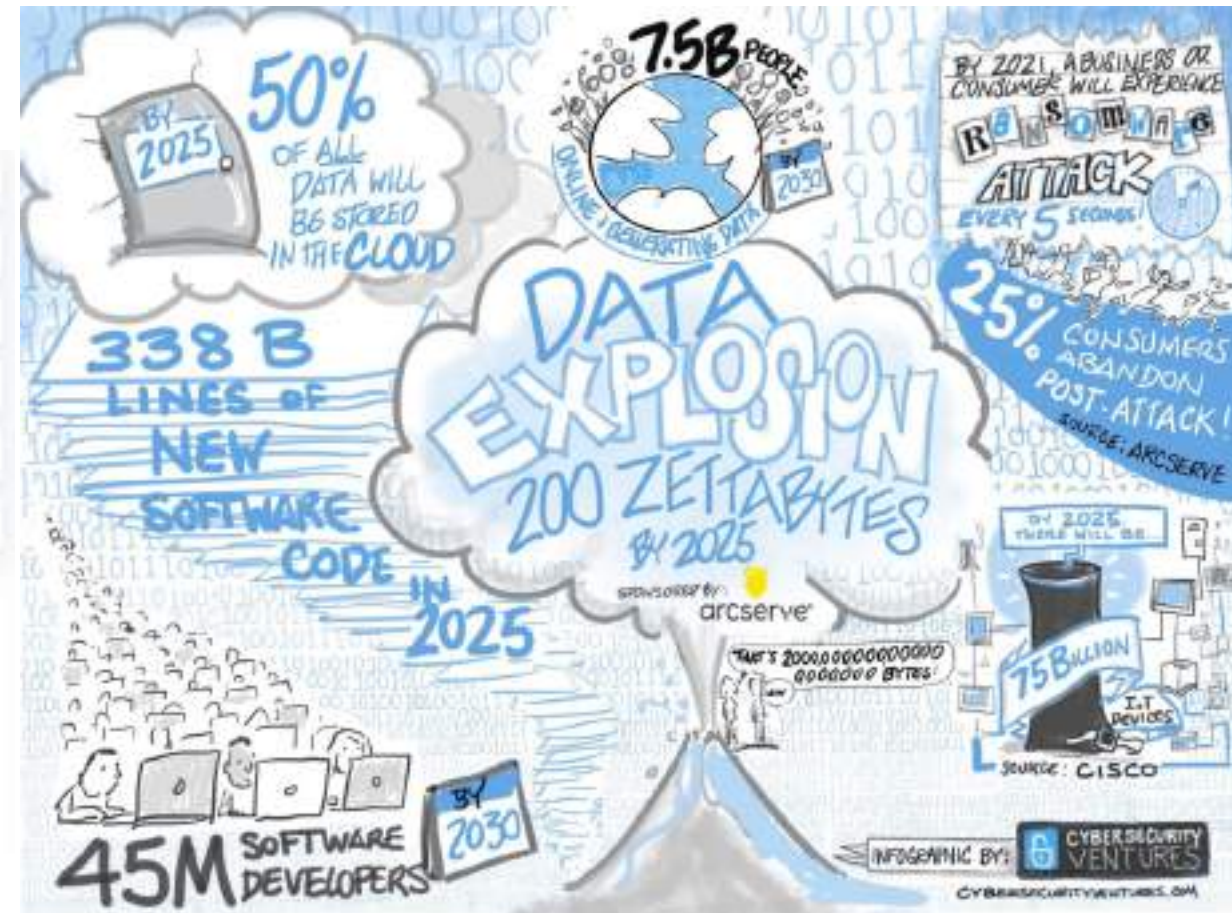


Data Center Storage Capacity Worldwide From 2016 To 2021, By Segment



© Statista 2021

Source : statistica.com, as of 2021



Source : Cybersecurityventures.com, as of June 2020



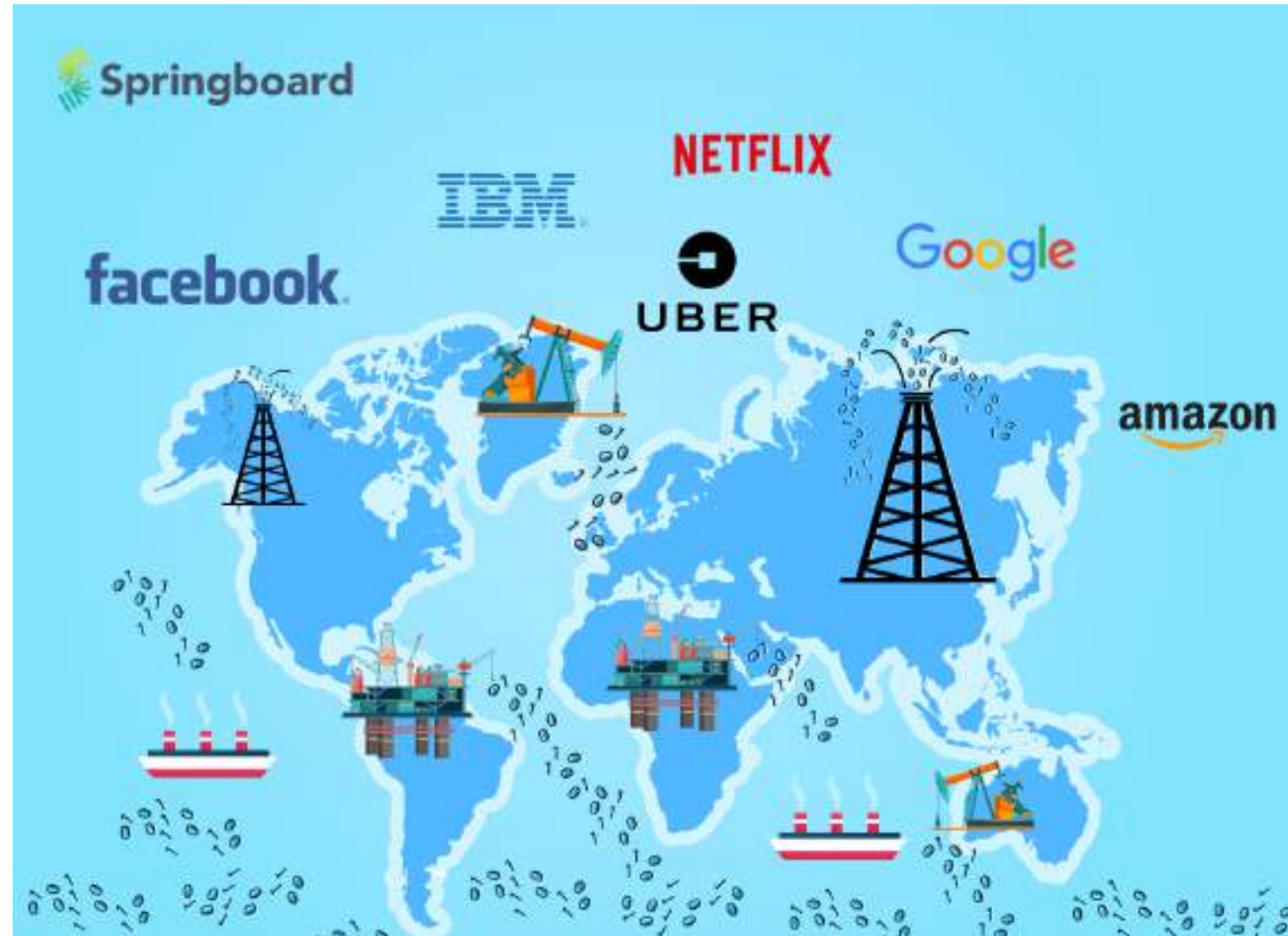
IBDAAI

INSTITUTE FOR BIG DATA ANALYTICS
AND ARTIFICIAL INTELLIGENCE



“Data is the
new Oil”

Clive Humby





IBDAAI

INSTITUTE FOR BIG DATA ANALYTICS
AND ARTIFICIAL INTELLIGENCE



MONETIZING BIG DATA



Use data to
see patterns
for business
understanding



Use data to
optimize the
business



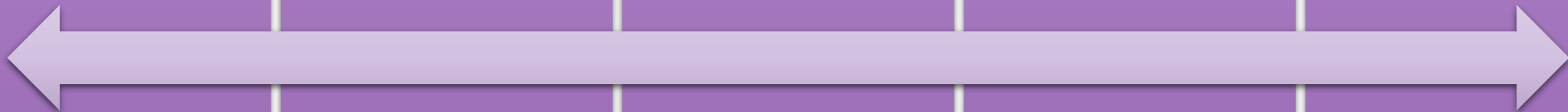
Use data to
address
business
challenges



Use data for
Storytelling



Use data to
gather better
data



Adapted : www.gartner.com



IBDAAI

INSTITUTE FOR BIG DATA ANALYTICS
AND ARTIFICIAL INTELLIGENCE



Monetizing Data : 3 Big Data Statistics

1 Big data impact on savings and profits



\$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.

Source: TechJury, Tractica, Entrepreneur, Grazziti

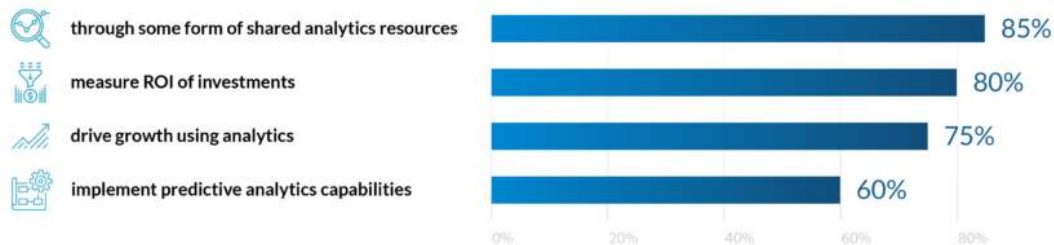
IBM

NETFLIX

Google

2 How do businesses leverage big data?

Source: iView System



Source : financesonline.com

3 Top benefits of data analytics

Sources: Chicago Analytics Group





IBDAAI

INSTITUTE FOR BIG DATA ANALYTICS
AND ARTIFICIAL INTELLIGENCE



“
AI is the new
electricity

Andrew Ng

”





IBDAAI

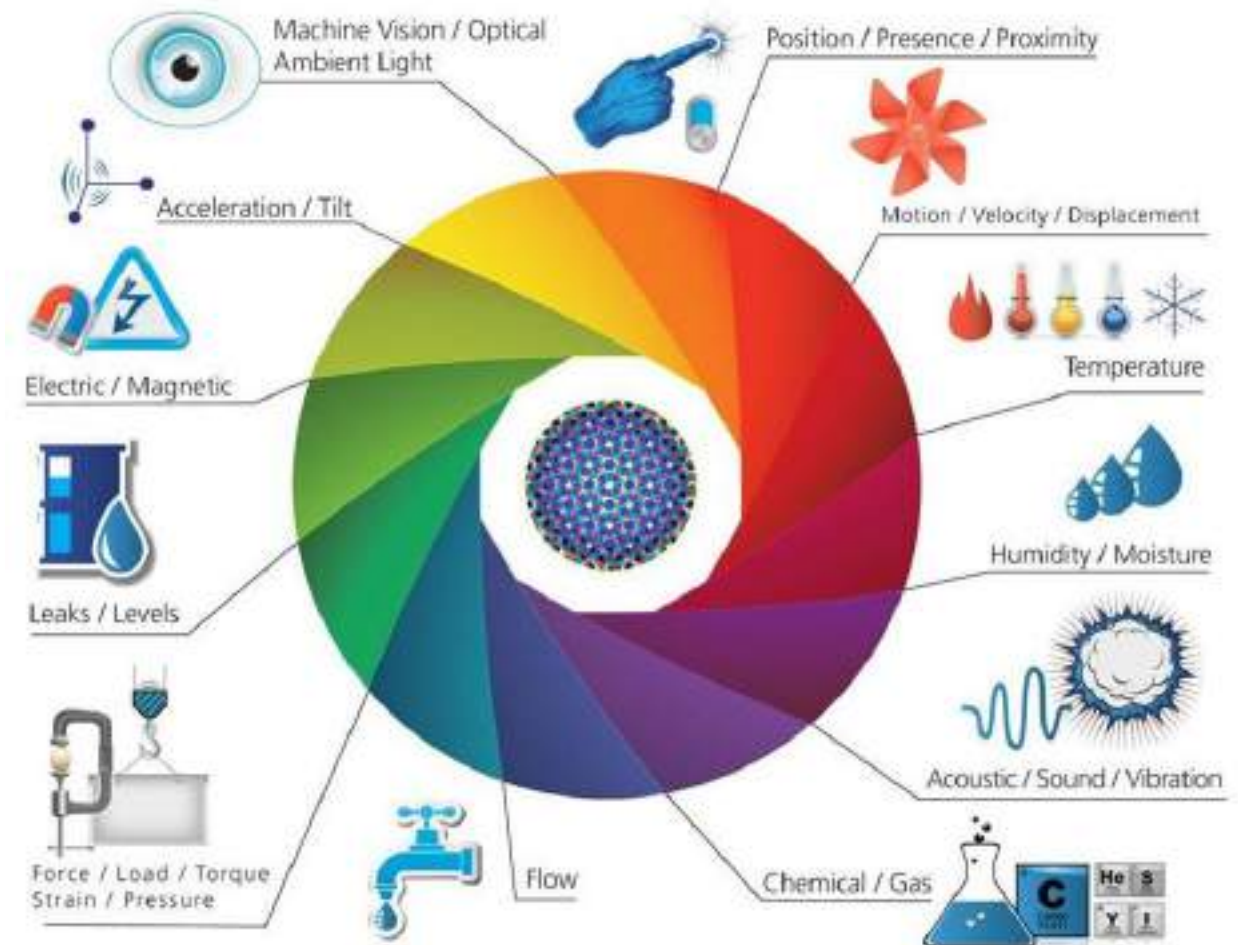
INSTITUTE FOR BIG DATA ANALYTICS
AND ARTIFICIAL INTELLIGENCE



“IoT is the new
nervous
system

Harbor Research

”



Source : Harborresearch.com as of, 2020

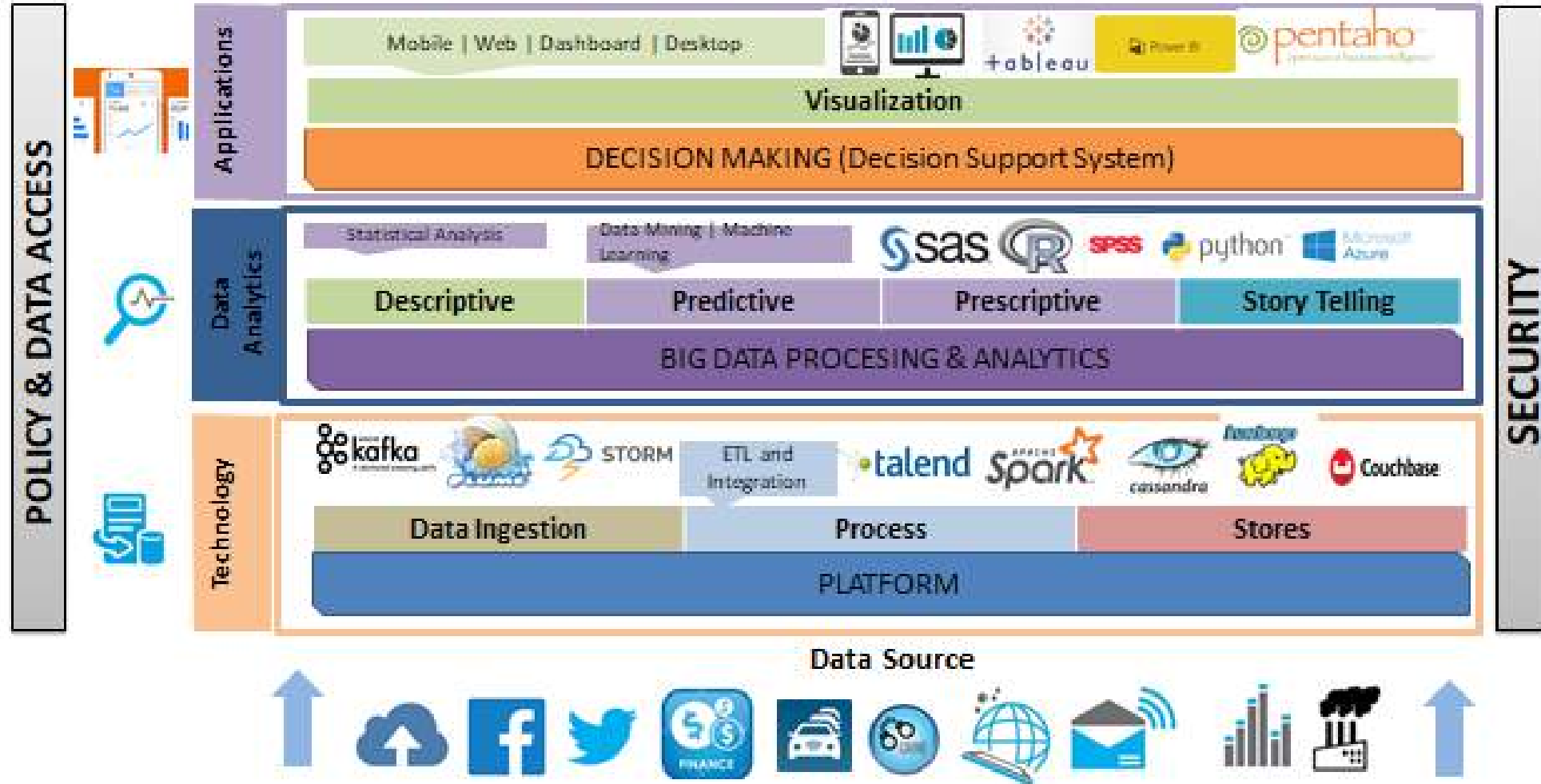


IBDAAI

INSTITUTE FOR BIG DATA ANALYTICS
AND ARTIFICIAL INTELLIGENCE



BIG DATA FRAMEWORK[©]



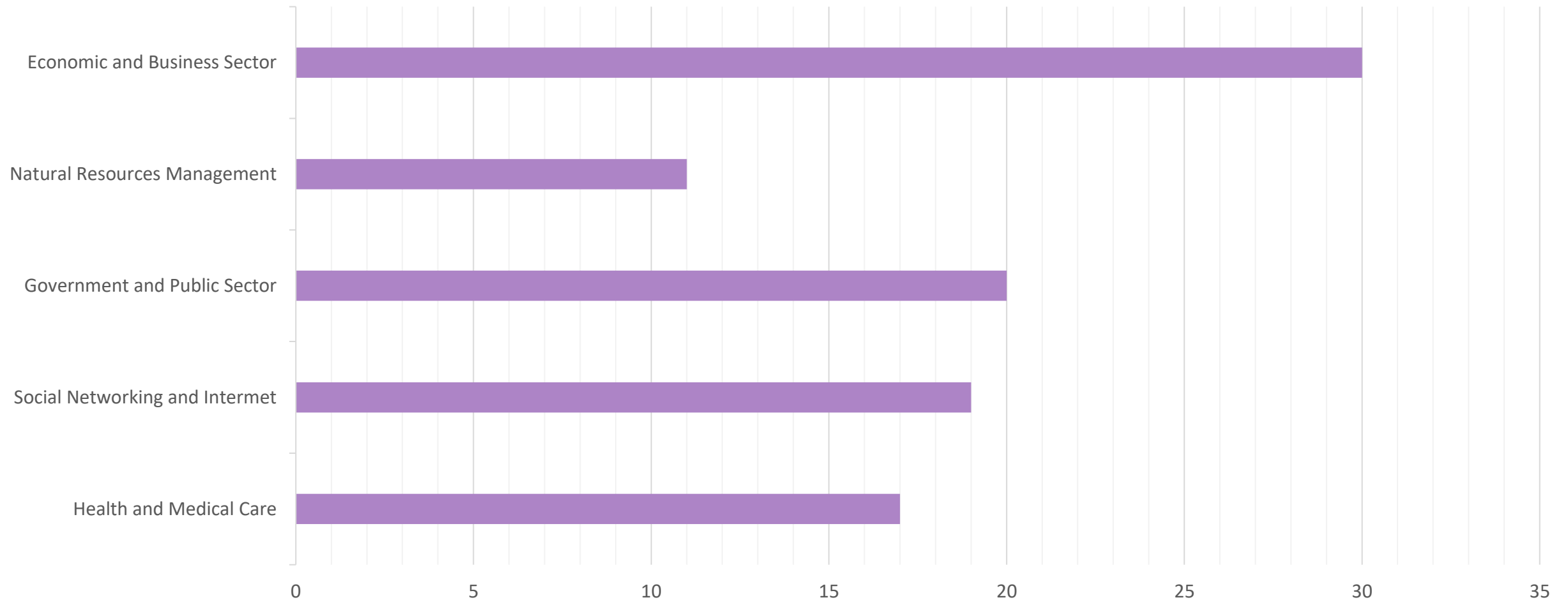


IBDAAI

INSTITUTE FOR BIG DATA ANALYTICS
AND ARTIFICIAL INTELLIGENCE



DOMAIN AREA





IBDAAI

INSTITUTE FOR BIG DATA ANALYTICS
AND ARTIFICIAL INTELLIGENCE



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



IBDAAI

INSTITUTE FOR BIG DATA ANALYTICS
AND ARTIFICIAL INTELLIGENCE



DATA PROCESSING & STORAGE

Batch Processing

- Hadoop / MapReduce
- Apache Pig
- Flume

Stream Processing

- Storm
- S4
- Kafka
- Flink
- Spark

Hybrid Processing

Google

facebook

YAHOO!



Alibaba.com™



Spark



IBDAAI

INSTITUTE FOR BIG DATA ANALYTICS
AND ARTIFICIAL INTELLIGENCE



BIG DATA ANALYTICS TECHNIQUES





IBDAAI

INSTITUTE FOR BIG DATA ANALYTICS
AND ARTIFICIAL INTELLIGENCE



MACHINE LEARNING

1st Generation



WEKA, SAS, R, Python

- Deep Analytics- large collection of serial ML Algo
- Vertical Scaling scalability
- 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, Hadoop, Apache Giraph, Pregel, Storm

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



IBDAAI

INSTITUTE FOR BIG DATA ANALYTICS
AND ARTIFICIAL INTELLIGENCE



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.



IBDAAI

INSTITUTE FOR BIG DATA ANALYTICS
AND ARTIFICIAL INTELLIGENCE



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

Web
Mining

Natural
Language
Processing

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

Text Analytics

Mining useful information from text sources.



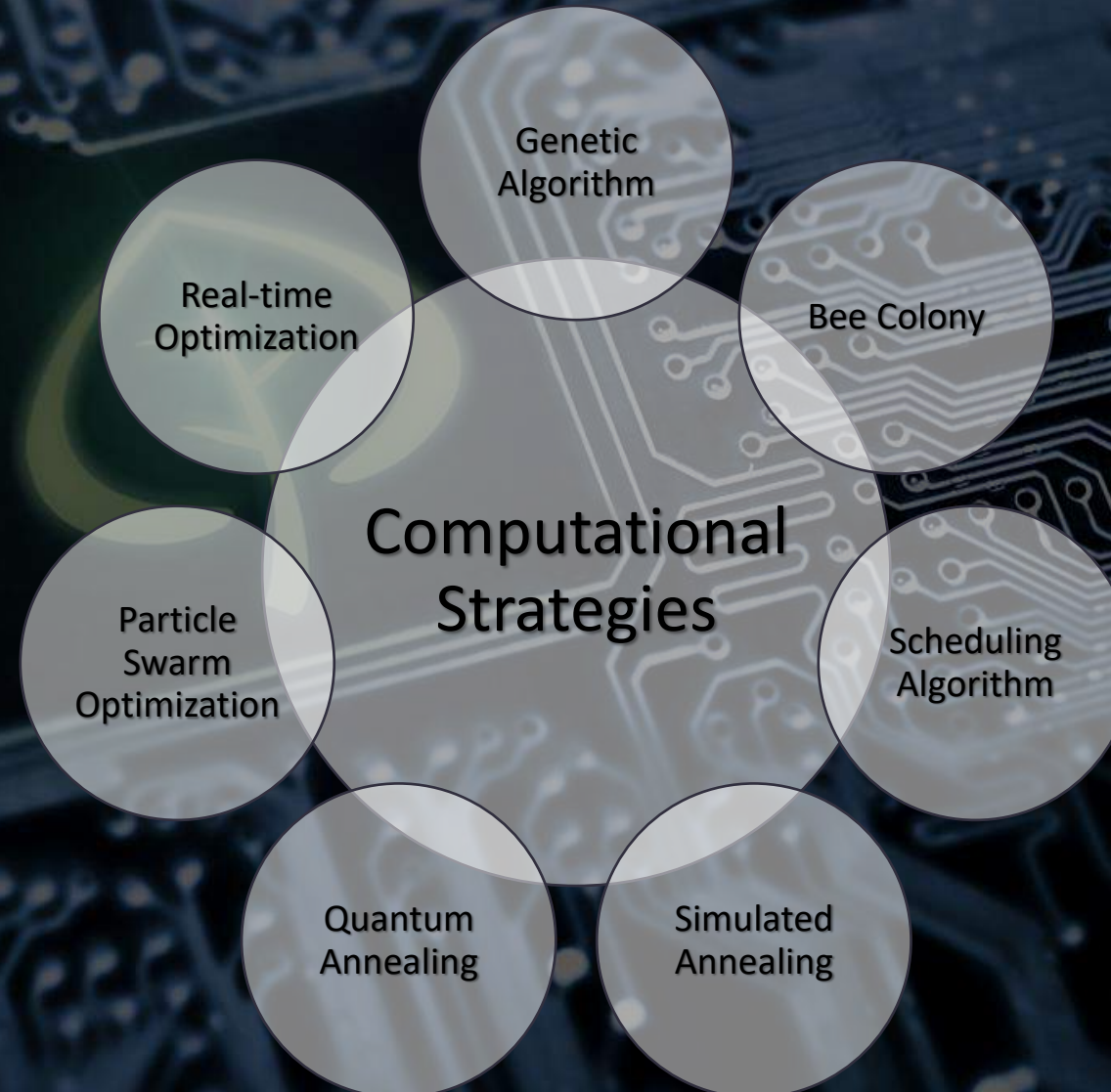
IBDAAI

INSTITUTE FOR BIG DATA ANALYTICS
AND ARTIFICIAL INTELLIGENCE



OPTIMIZATION TECHNIQUES

Solve Quantitative Problems –
Multidisciplinary Fields





IBDAAI

INSTITUTE FOR BIG DATA ANALYTICS
AND ARTIFICIAL INTELLIGENCE



DATA VISUALIZATION

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

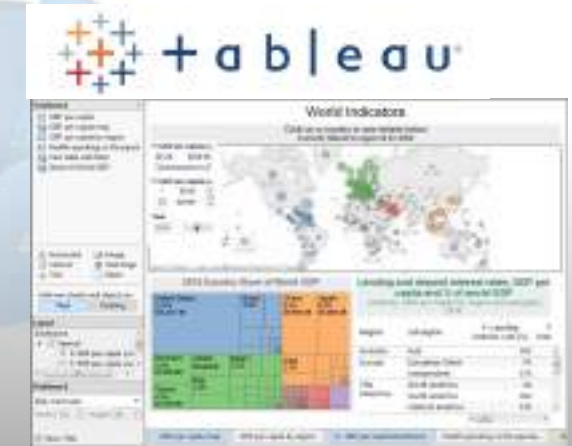
Narrative + Data = Explain

Data + Visual = Enlighten

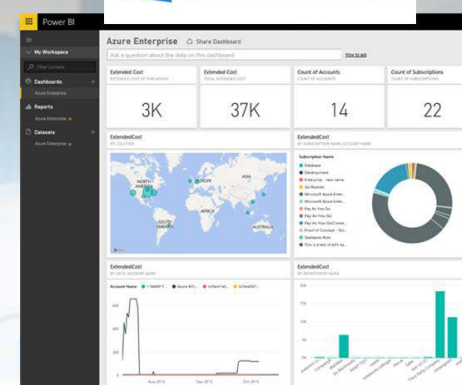
Visual + Narrative = Engage



ORACLE BI



Microsoft Azure





IBDAAI

INSTITUTE FOR BIG DATA ANALYTICS
AND ARTIFICIAL INTELLIGENCE



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

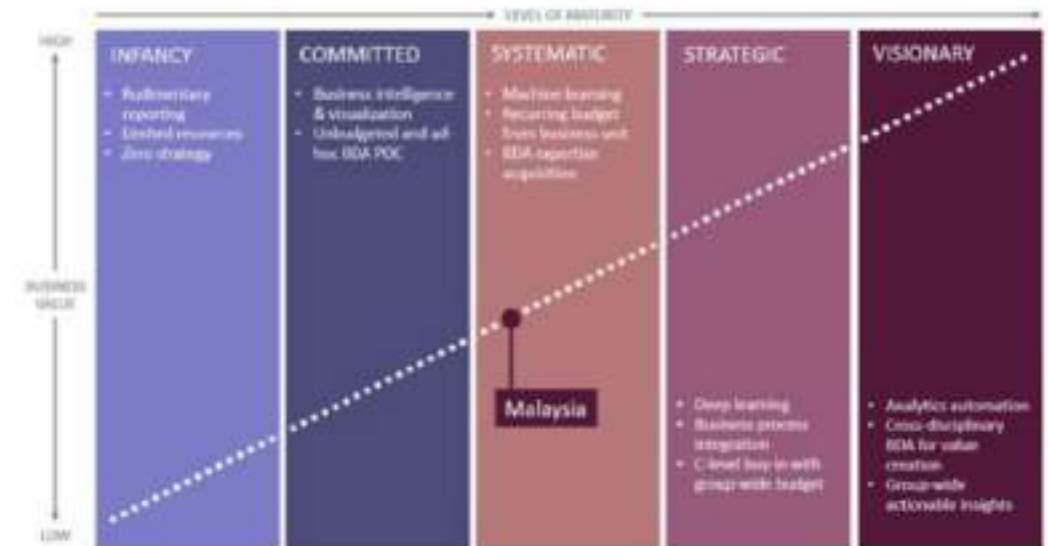
MALAYSIA AI BLUEPRINT 2020 ANNUAL REPORT

A Digital Solution Towards Data Driven Agriculture in Malaysia

The Center of Applied Data Science (CAIDS) Feb 25, 2020 • 8 min read



Source : medium.com, as of Feb 2020



Malaysia's BDA maturity (overall)

Source : bigittechnology.com, as of 2020

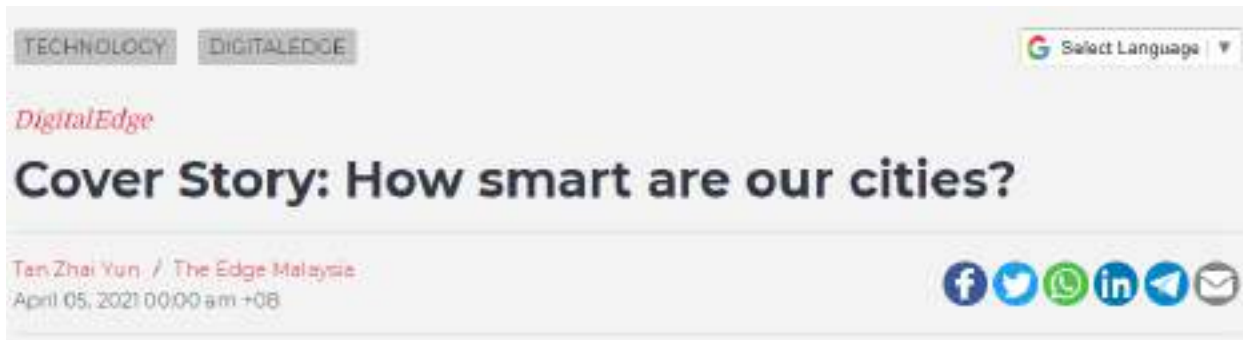


IBDAAI

INSTITUTE FOR BIG DATA ANALYTICS
AND ARTIFICIAL INTELLIGENCE



BIG DATA ANALYTICS : MALAYSIA VIEW



"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

© 24th Jul 2021 | MOH COVID-19 DATA, GITHUB



<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

INSTITUTE FOR BIG DATA ANALYTICS
AND ARTIFICIAL INTELLIGENCE



IBDAAI INITIATIVES



“ ”

Social Listening



Health Analytics



Oil and Gas



Data Drifting
(Streaming Data)



Leveraging Big Data Analytics Big Data : Future Direction

183.102

154.178

2455



IBDAAI

INSTITUTE FOR BIG DATA ANALYTICS
AND ARTIFICIAL INTELLIGENCE



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

Gartner Top 10 Data and Analytics Trends, 2021



gartner.com/SmarterWithGartner

Source: Gartner
© 2021 Gartner, Inc. All rights reserved. CTMKT1164473

Gartner



IBDAAI

INSTITUTE FOR BIG DATA ANALYTICS
AND ARTIFICIAL INTELLIGENCE



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

Hype Cycle for Emerging Technologies, 2021



gartner.com

Source: Gartner
© 2021 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner and Hype Cycle are registered trademarks of Gartner, Inc. and its affiliates in the U.S. 1000000

Gartner.



IBDAAI

INSTITUTE FOR BIG DATA ANALYTICS
AND ARTIFICIAL INTELLIGENCE



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



The poster for SCDS2021 features a circular logo on the top left with a hand holding a blue diamond, surrounded by the text 'SCDS 2021' and 'SCDS IN DATA SCIENCE'. To the right, the title 'SCDS2021' is written in large, bold, blue letters. Below the title, it says 'THE 6TH INTERNATIONAL CONFERENCE ON SOFT COMPUTING IN DATA SCIENCE' and '2-3 NOV 2021 VIRTUAL CONFERENCE'. A blue banner across the middle contains the text 'SCIENCE IN ANALYTICS: HARNESSING DATA AND SIMPLIFYING SOLUTIONS' and the Springer logo. Below this, a paragraph describes the conference's organization and goals. To the right of the paragraph is a 'TOPICS' section with a list of subjects. At the bottom, a 'KEYNOTE SPEAKERS' section displays seven circular portraits of the speakers with their names and affiliations below them.

SCDS2021
THE 6TH INTERNATIONAL CONFERENCE
ON SOFT COMPUTING
IN DATA SCIENCE
2-3 NOV 2021
VIRTUAL CONFERENCE

SCIENCE IN ANALYTICS:
HARNESSING DATA AND SIMPLIFYING SOLUTIONS

SCDS2021 is organized by IBDAAI (Institute for Big Data Analytics and Artificial Intelligence), UTM, with the collaborations of Faculty of Computer Science and Mathematics, UTM, Department of Statistics, Faculty of Science and Data Analytics, Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia, and Research Nexus, UTM from 2-3 November 2021. SCDS2021 aims to provide a platform for researchers and practitioners to share their research work and to create rigorous international research collaborations. All accepted and registered papers of SCDS2021 will be published in the renowned series of "Communications in Computer and Information Science series" by Springer Verlag. Selected papers will have the opportunity for extended papers to be published in Scopus indexed journals.

TOPICS

- Big Data Analytics
- Artificial Intelligence
- Big Data Technologies
- Deep Learning
- Data Mining
- Machine Learning
- Bayesian models and methods
- Genetic algorithms
- Data Representation and Visualization
- Fuzzy Logic
- Optimization Algorithms
- Graph/Network/Social Mining
- Image Processing
- Computational Intelligence
- Cloud Computing
- Computational Statistics

...and any other related topics.

KEYNOTE SPEAKERS

| | | | | | | |
|---|--|---|---|--|---|--|
|  YANG DATTO SRI DR. MOHD UZIR MAHIDIN Chief Statistician, Malaysia |  PROF. DR. CHUAN-HSIN HUANG National Yang Ming Chiao Tung University, Taiwan |  PROF. DR. RER. POL. NERI KUSWANTO Institut Teknologi Sepuluh Nopember, Indonesia |  ASSOC. PROF. DR. AIDEN DOHERTY Robert Higgs Institute, Bedford University, UK |  ASSOC. PROF. DR. SIMON FONG University of Macau, China |  PROF. DR. RICHARD C. MILLHAM Durban University of Technology, Africa |  ASSIST. PROF. DR. UTHMAN HAIDOU University of North Texas, Health Science Center |
|---|--|---|---|--|---|--|



IBDAAI

INSTITUTE FOR BIG DATA ANALYTICS AND ARTIFICIAL INTELLIGENCE



Artificial Intelligence Review
ISSN 2689-5915 (Print) 2689-5923 (Online)

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

Azlinah Mohamed¹, Marwan Hameed Najafabadi², * Tay Bee Wuk³
Ezzat El-Mahalawi Elmaghrabi⁴, Rydhya Mubaid⁵

¹ Springer Nature, 19, 2019

Abstract

Big data has become a significant research area due to the high volume of data generated from various sources like social media, internet of things, and multimedia applications. Big data has played critical roles in many decision making and forecasting domains such as recommendation systems, business analysis, healthcare, retail display advertising, education, transportation, brand detection and tourism marketing. The rapid development of various big data tools such as Hadoop, Storm, Spark, Flink, Kafka and Pig is research and industrial communities has allowed the huge number of data to be distributed, processed and processed. Big data applications use big data analytics techniques to efficiently analyze large amounts of data. However, choosing the suitable big data tool based on both real-world data processing and analytics techniques for development a big data system are difficult due to the challenges in processing and applying big data. Practitioners and researchers who are developing big data systems have inadequate information about the current technology and requirement concerning the big data platform. Hence, the strengths and weaknesses of big data technologies and effective solutions for the Big Data challenges are needed to be discussed. Hence, due to that, this paper presents a review of the literature that analyze the use of big data tools and big data analytics techniques in areas like health and medical care, social networking and internet, government and public sector, natural resource management, economic and business sector. The goals of this paper are to (1) explain the trend of big data related research and current status of big data technologies; (2) identify trends in the use of research of big data tools based on health and natural resource processing and big data analytics techniques; (3) assist and provide new researchers and practitioners to plan new research activities in this domain appropriately. The findings of this study will provide insights and knowledge on the existing data science trends and their applications domains, the strengths and weaknesses of big data tools, big data analytics techniques and their use, and new research opportunities in future development of big data systems.

Keywords: Parallel and distributed computing; Big data tools; Big data analytics techniques; Domain area

¹ Springer Nature Singapore
marwan.hameed@univnet.edu.jo

Extended author information available on the last page of the article.

Published online: 11 February 2019



Professorial Lecture **UiTM**

Higher Education Institution and BIG DATA ANALYTICS

OBSERVE & DEDUCE

AZLINAH MOHAMED

ISSN 2222-0208 | Institute for Analytics of Agri-food Systems and Food Safety | January 17-18-19 | 100 pages | 1000

Tag recommendation model using feature learning via word embedding

Marwan Hameed Najafabadi¹
Department of Systems Engineering
and Computer Science
University of Jordan, Amman
Jordan, Amman, Jordan
marwan.hameed@univnet.edu.jo

Marwan Hameed Najafabadi²
Department of Systems Engineering
and Computer Science
University of Jordan, Amman
Jordan, Amman, Jordan
marwan.hameed@univnet.edu.jo

Azlinah Mohamed³
Faculty of Computer and Mathematical
Science
Universiti Teknologi MARA
Johor Bahru, Malaysia
azlinahmohamed@uitm.edu.my

Abstract—Tag recommendation models serve as extracting metadata for target objects like images, videos and Web pages. However, these models could not solve problem due to absence of initial tags. To improve tag quality in tag recommendation systems, some of previous works exploit the statistical properties such as co-occurrence patterns or topic frequency to predict the candidate tags as a target object. However, tag recommendation models fail to be effective when initial tags are absent or low quality tags are available for objects. Recently, authors studying the word embeddings achieve success in many natural language processing tasks. Therefore, this paper aims to introduce a novel tag recommendation algorithm that can address the relation between words in a text generated with target object using word embeddings. In fact, we further generalize relations between words in a text or sentence with focus on human learning methods. This paper model is used to explain human relation and learn the dependencies, which are useful for tag recommendation. Our method shows improvement in previous research methods with gain of up to 18 percent in precision when using data from Amazon dataset.

Keywords—Tag recommendation model; Tag recommendation; Word embeddings; Feature learning

1. INTRODUCTION

Social tags have gained in popularity in systems without exception with Web 2.0 applications for example photo sharing like Flickr, social bookmarking sites like del.icio.us, and Amazon.com. In such applications, users can tag items and improve their objects by e.g. images, text or videos with their tags as well as their friends to find useful and sharing. In particular, tags or words are key words featured by users in the objects for supporting various recommendation systems classification and information retrieval [1-4].

Most tag recommendation methods require to take into account the semantic social tag sets as the target object for tag suggestions [5-10]. These methods also advantage of co-occurrence patterns to predict tags that co-occur with the initial set of tags to rank the candidate tag recommendations. However, recommendation methods were designed for recommendation the content and not their metadata and

In this paper, we solve the cold start problem by studying human dependencies from multiple sets of word embeddings. Our method uses the relation and similarity between words in text to description of objects via random walk. Our tag recommendation system exploits graph models with some information features related to related grammatical relations between words to address the absence of the target object. Therefore, combining the evidence between words via random walk models, such as dependencies and similarity between objects is important. Related techniques applied to solve tag recommendation methods in most research papers. When related candidate tags in our method can be extracted, they are taken advantage of data related social network embeddings [6] to analyze the dependencies between words in a text. It is reasonable to identify good candidate tags for target object in object by assessing the dependency-based features for which words it is a source.

Hence, to rank tags in a cold start scenario without depending on word frequency metrics, we focus from the topic related network embeddings [6]. Our model without user comments have high degree of similarity in response to topic belonging to different communities. Our method uses the dependency-based to provide effective information or features for learning representation vector of words in related semantic context as tag and description related to target object. Subsequently, we employ the best gradient descent to optimize the learning representation vector of words.

We evaluate our method on real dataset, namely Flickr dataset dataset and compare results with previous research methods. The experimental results show the efficiency of our method with cold start scenario in contrast to other tag recommendation models.

The remainder of this paper is organized as follows. Section 2 outlines related research on tag recommendation systems. Section 3 explains details and structure of our method. We evaluate the experimental results in Section 4. Finally, we conclude the paper and present future works.

© 2019 IEEE

THANK YOU