Introduction to Big Data

BIG DATA FUNDAMENTALS

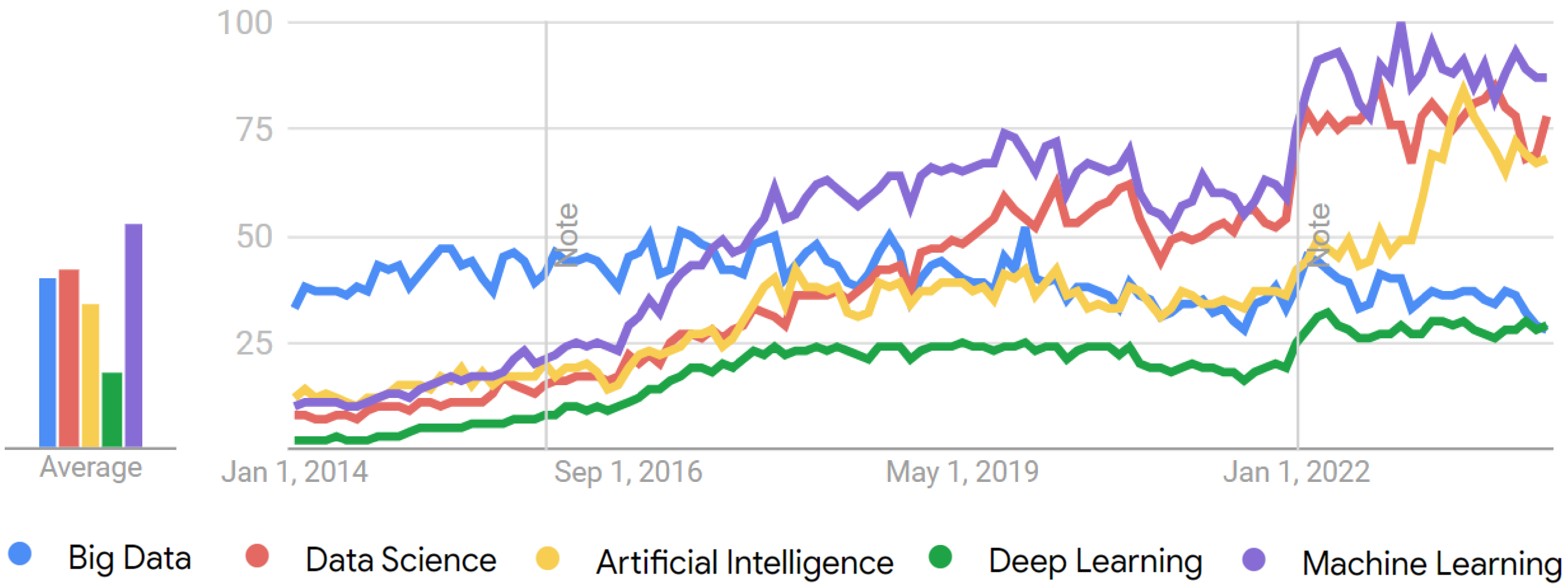
Part I

Le Ngoc Thanh – Nguyen Ngoc Thao

{lnthanh, nnthao}@fit.hcmus.edu.vn

Big data: The trending term

- Big data is among trending search terms in recent years.



Source: Google Trends, updated 01/2024

Outline

- What is Big data?
  - The definitions of Big data
  - The V’s characteristics of Big data
  - Common Issues in Big data
- Big data case studies
  - The applications of Big data
  - Big data projects in practice
- Motivations and opportunities

What is Big data?

It is not big. It is just bigger…

Big data: A definition

- A variety of definitions for Big data are available worldwide.

Big data is a term used to refer to the *study and applications of data sets that are so big and complex* that traditional data-processing application software are inadequate to deal with them. – Wikipedia.

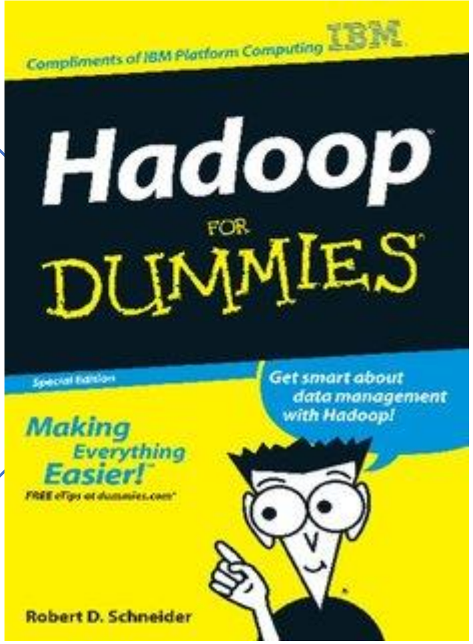
Big data is *high-volume, high-velocity and/or high-variety information assets* that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation. – Garner, 2001.

Big data refers to *the dynamic, large and disparate volumes of data being created by people, tools and machines*; it requires new, innovative and scalable technology to collect, host and analytically process the vast amount of data gathered in order to derive real-time business insights that relate to consumers, risk, profit, performance, productivity management and enhanced shareholder value. – Ernst & Young, 2014.

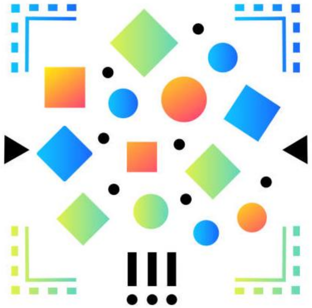
Big data: A definition

Big data is a term that describes at least three separate, but interrelated, trends.

- ✓ Capturing and managing lots of information
- ✓ Working with many new types of data
- ✓ Exploiting these masses of information and new data types with new styles of applications

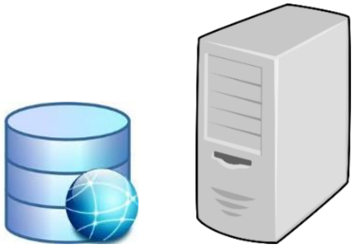


Compliments of IBM Platform Computing
IBM
Hadoop
FOR
DUMMIES
Special Edition
Get smart about data management with Hadoop!
Making Everything Easier!
FREE eTips at dummies.com
Robert D. Schneider



Small data vs. Big data

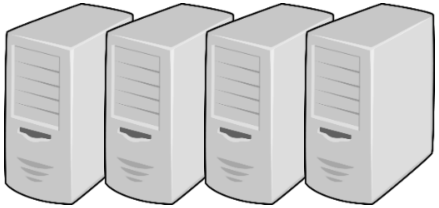
- “Big data” is just the “small data” that grows bigger.
- The new scale of data may require novel approaches for techniques and frameworks.
- It is now able solve new problems or existing problems in a better way.



small data, small computer

bigger data

bigger computer?



or more small computers?

Technologies in Big data

- Not a single technology but a combination of old and new technologies that helps companies gain actionable insight

There is an urgent need for parallel processing in distributed environment with high scalability.

- Capability to manage a huge volume of disparate data, at the right speed, and within the right time frame → allow for real-time analysis and reaction

The 5V’s characteristics

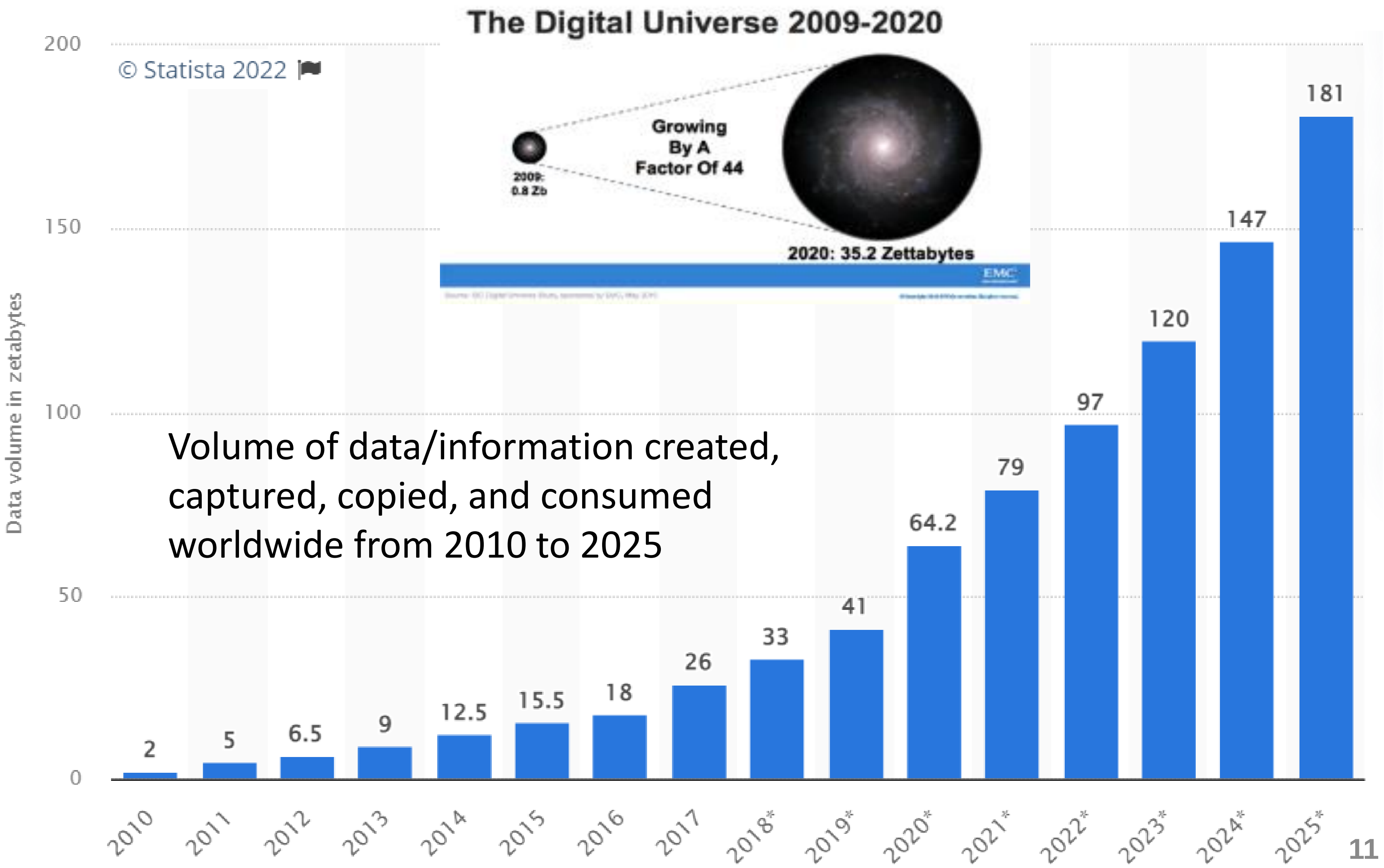
- The characteristics of Big data are characterized by the 5V’s.

Image credit: Dreamtime

The 5V's characteristics

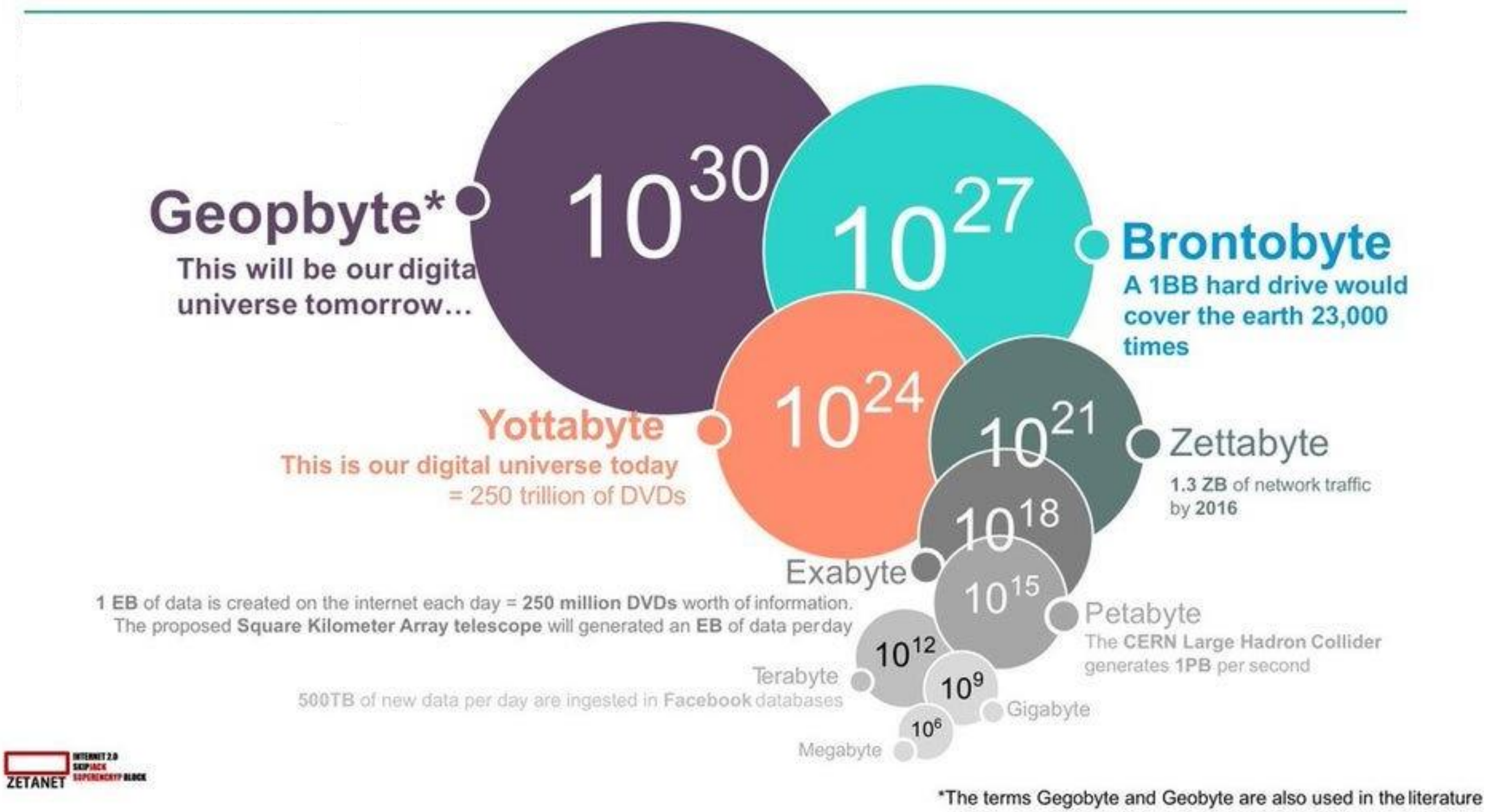
- Description: The amount of data generated is vast compared to traditional data sources.
- Attributes: Exabyte, zettabyte, yottabytes, etc.
- Drivers: Increase in data sources, higher resolution sensors, scalable infrastructure.

The explosive growth of data



Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2025

New data units for the Big data era



Geopbyte*
This will be our digita
universe tomorrow…

Brontobyte
A 1BB hard drive would cover the earth 23,000 times

Yottabyte
This is our digital universe today
= 250 trillion of DVDs

Zettabyte
1.3 ZB of network traffic by **2016**

Exabyte
1 EB of data is created on the internet each day = **250 million DVDs** worth of information.
The proposed **Square Kilometer Array telescope** will generated an **EB** of data perday

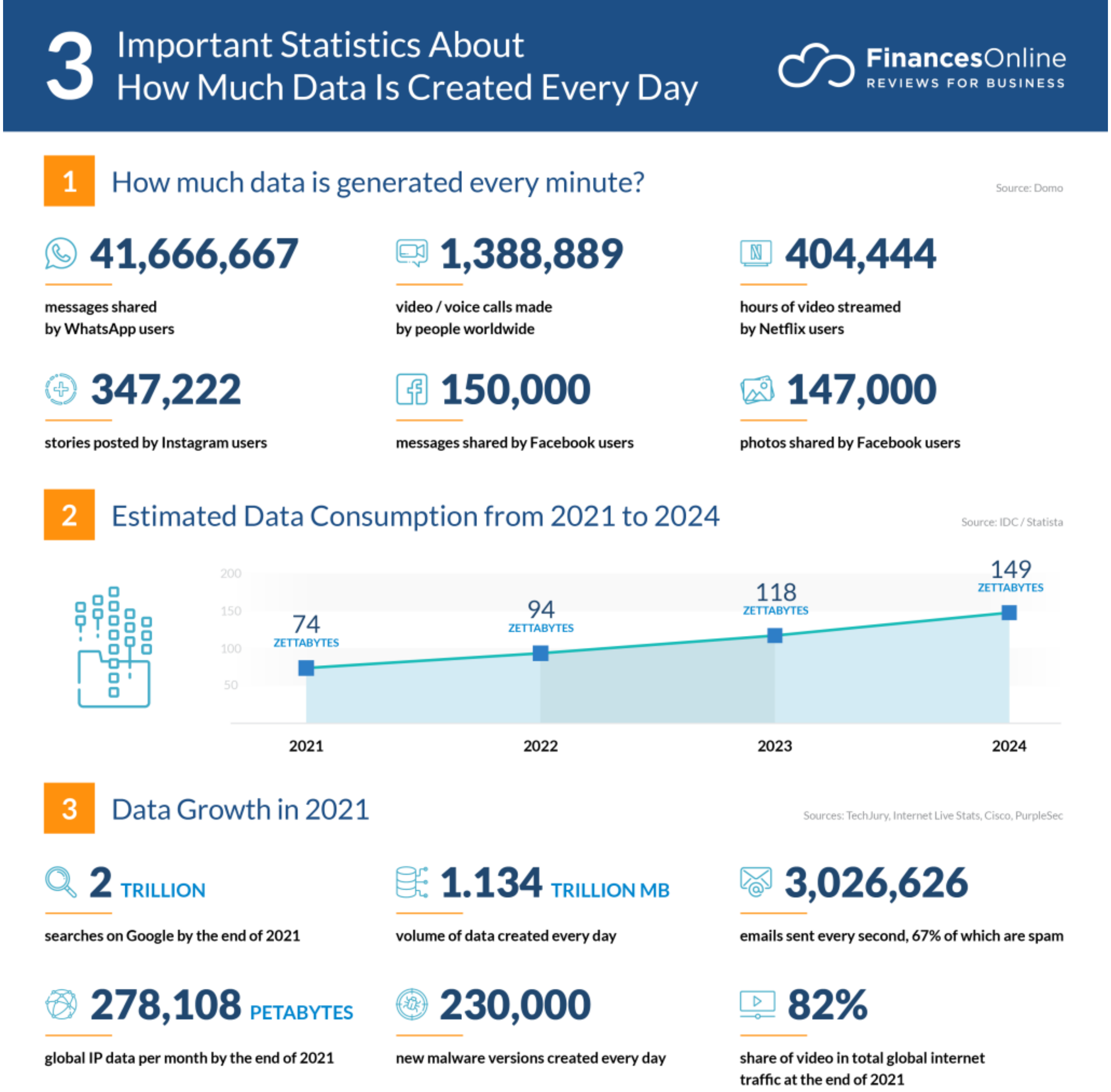
Petabyte
The **CERN Large Hadron Collider** generates **1PB** per second

Terabyte
500TB of new data per day are ingested in **Facebook** databases

Gigabyte

Megabyte

*The terms Gegobyte and Geobyte are also used in the literature



3 Important Statistics About How Much Data Is Created Every Day

FinancesOnline
REVIEWS FOR BUSINESS

1 How much data is generated every minute?

Source: Domo

41,666,667
messages shared by WhatsApp users

1,388,889
video / voice calls made by people worldwide

404,444
hours of video streamed by Netflix users

347,222
stories posted by Instagram users

150,000
messages shared by Facebook users

147,000
photos shared by Facebook users

2 Estimated Data Consumption from 2021 to 2024

Source: IDC / Statista

3 Data Growth in 2021

Sources: TechJury, Internet Live Stats, Cisco, PurpleSec

2 TRILLION
searches on Google by the end of 2021

1.134 TRILLION MB
volume of data created every day

3,026,626
emails sent every second, 67% of which are spam

278,108 PETABYTES
global IP data per month by the end of 2021

230,000
new malware versions created every day

82%
share of video in total global internet traffic at the end of 2021

Why does data become big now?

- Key enablers of appearance and growth of data are

Increase of storage capacity

Increase of processing power

Availability of data

The 5V's characteristics

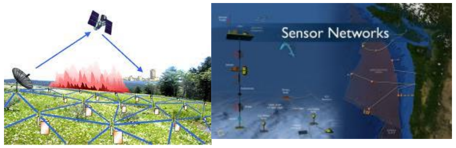
- Description: Data is being generated extremely fast, a process that never stops; and the speed at which data is transformed into insight
- Attributes: Batch; near/real-time; streams
- Drivers: Improved connectivity; competitive advantage; precomputed information

Real-time and/or fast data

Social media and networks
(all of us are generating data)

Scientific instruments
(collecting all sorts of data)

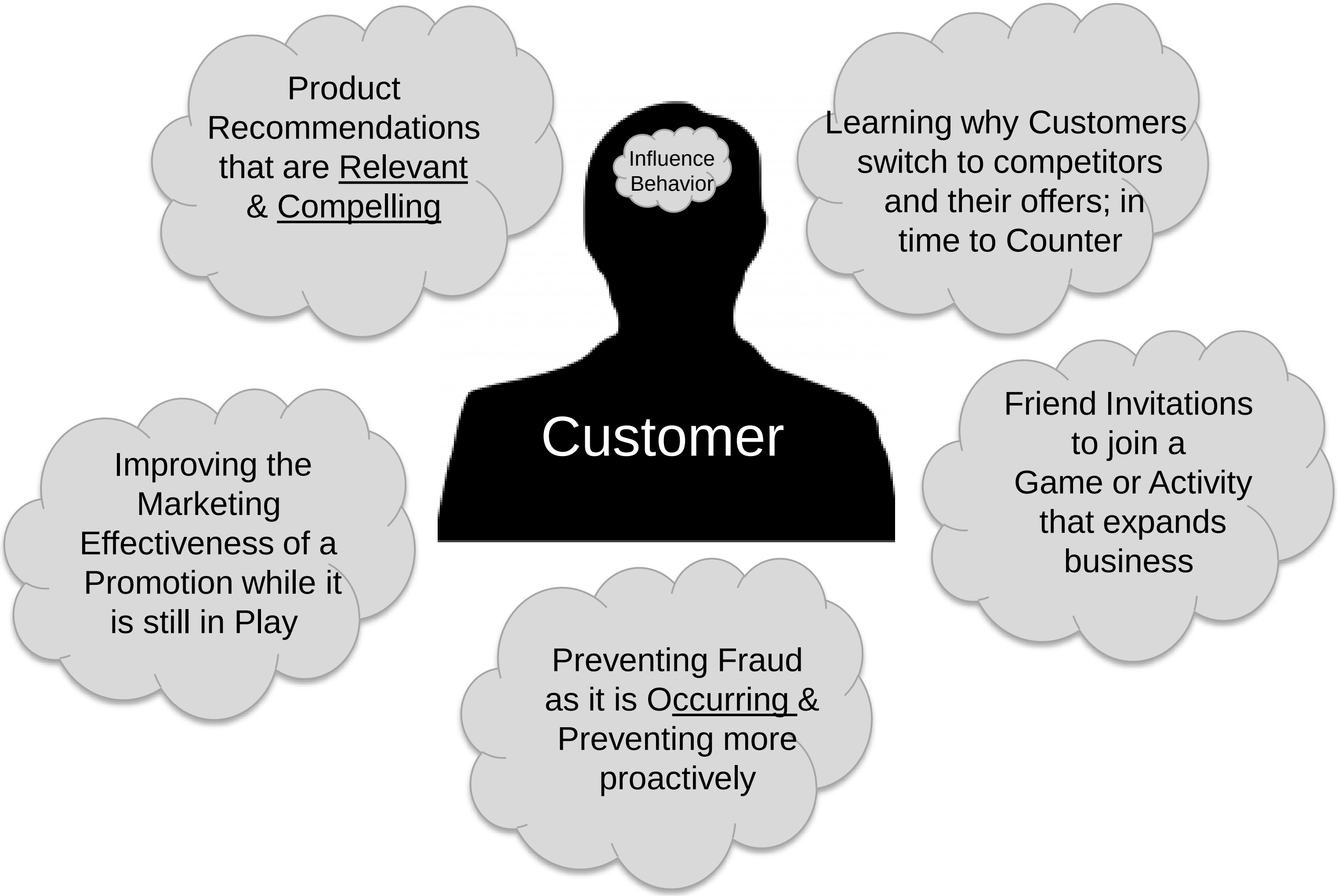
Mobile devices
(tracking all objects all the time)



Sensor technology
and networks
(measuring all kinds of data)

- Innovations and their progresses are no longer hindered by the ability to collect data but by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion.

Real-time analytics/decision requirement



Customer

Influence Behavior

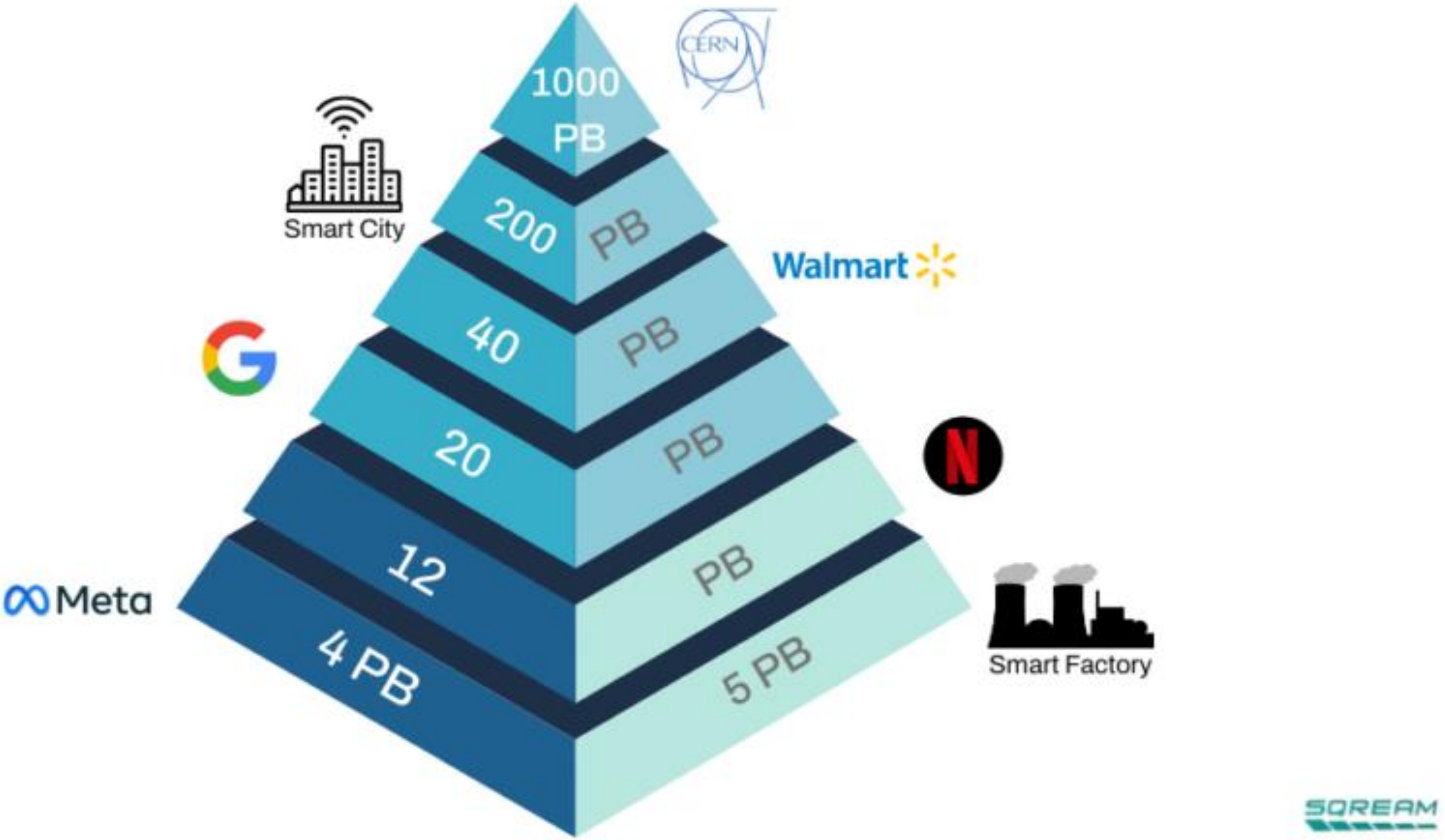
Product Recommendations that are Relevant & Compelling

Learning why Customers switch to competitors and their offers; in time to Counter

Friend Invitations to join a Game or Activity that expands business

Preventing Fraud as it is Occurring & Preventing more proactively

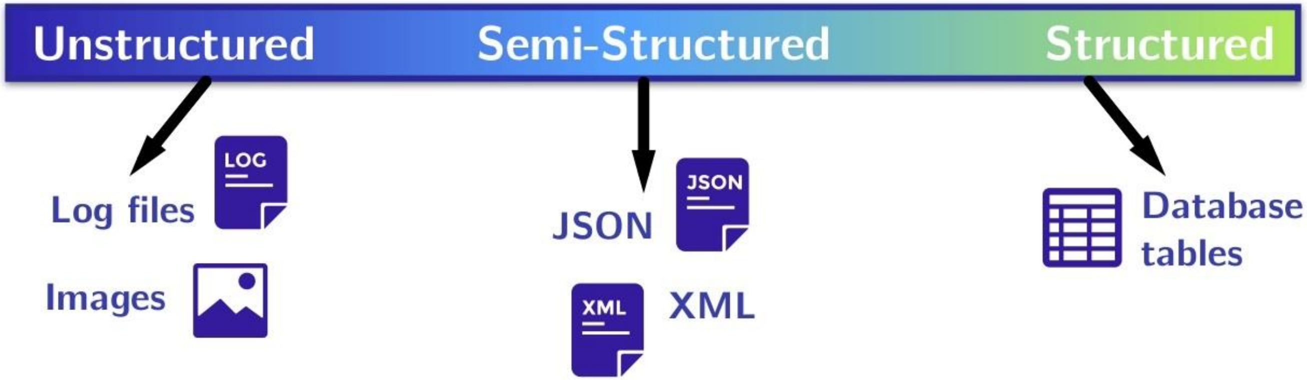
Improving the Marketing Effectiveness of a Promotion while it is still in Play



Data processed per day (Updated in January 2022)

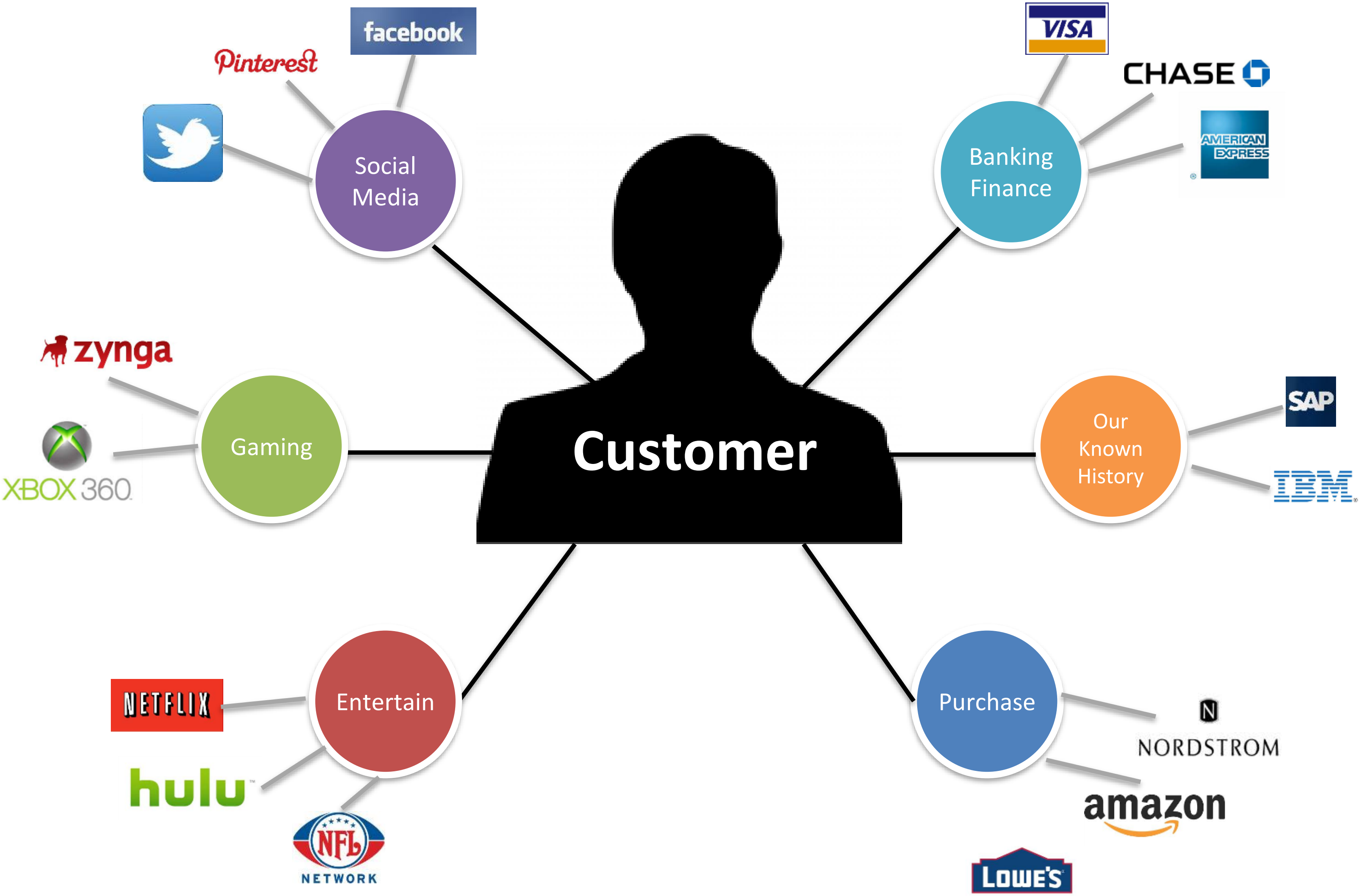
Image credit: sqream

The 5V's characteristics



- <span>Description:</span> Data comes from different sources, machines, people, and processes both from outside and inside the organizations
- Attributes: Degree of structure; complexity
- Drivers: Mobile; social media; video; genomics; IoT

A single view to the customer



The 5V's characteristic

- Description: Quality and origin of data
- Attributes: Consistency; completeness; integrity; ambiguity
- Drivers: Cost; need of traceability and justification

- The ability and need to turn data into value
- Value is not only profit but also medical or social benefits, or personal satisfaction (customer, employee, etc.).

Common issues related to the 5V’s

- As the data volume increases, the value of different data records will decrease in proportion to age, type, richness, and quantity among other factors.
- It is hard to handle complex data by existing traditional analytic systems.
  - Big data with relational databases, statistics/visualization packages
  - Massively parallel software running on tens, hundreds, or even thousands of computing units.
  - Data analytics with data that is constantly in motion.

Issues of personnel

- There is a considerable gap between Business leaders and IT professionals.

- Business leaders concern about adding value to their business and getting more and more profit.
- Meanwhile, IT leaders focus on the technicalities of the storage and processing only.

Issues of storage and transport

- Assume that we have an exabyte (10^{18}) of data.
- Modern hard disks can store several TBs (10^{12})

  → thousands of disks are required for an exabyte.
- A single computer system would be unable to directly attach the requisite number of disks.
- Accesses to that data also overwhelms the networks.
  - E.g., a 1GB/second network operating at 80% efficiency and a sustainable bandwidth of 100 MB/second would take approximately 14,465 days to transfer an exabyte of data.

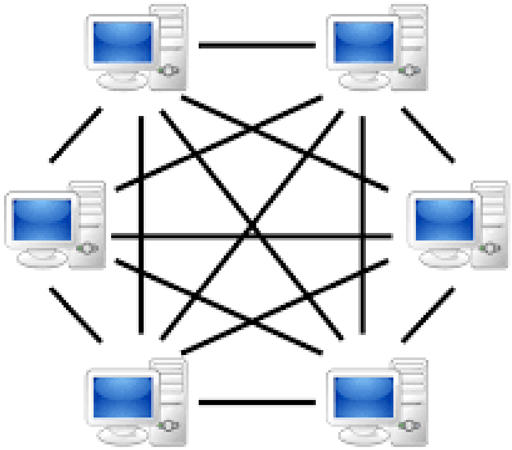
Issues of data management

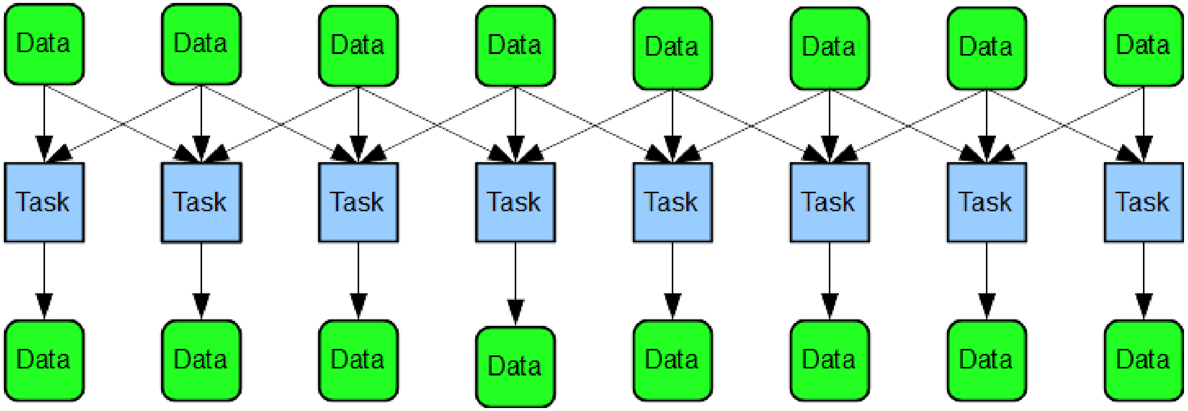
- Possibly the most difficult problem
- Issues of access, utilization, updating, governance, and reference (in publication) are major stumbling blocks.

- Data sources are varied by size, format, and by method of collection.
  - What, when, where, who, why and how it was collected.
- It is impractical to validate every data item in a huge source.

Issues of processing power

- Extensive parallel processing and new analytics algorithms are required.





Assume that an exabyte of data need to be processed and it is chunked into blocks of 8 words → 1 exabytes = 1K petabytes.

Assuming a processor expends 100 instructions on one block at 5 gigahertz → 1K petabytes would require a processing time of 635 years.

Big data case studies

The more data, the better decisions, and then the better outcomes…

Big Data use case categories

Big Data Exploration

Find, visualize, understand all big data to improve decision making

Enhanced 360º View of the Customer

Extend existing customer views (MDM, CRM, etc) by incorporating additional internal and external information sources

Security/Intelligence Extension

Lower risk, detect fraud and monitor cyber security in real-time

Operations Analysis

Analyze a variety of machine data for improved business results

Data Warehouse Augmentation

Integrate big data and data warehouse capabilities to increase operational efficiency

Potential Use Cases for Big Data Analytics

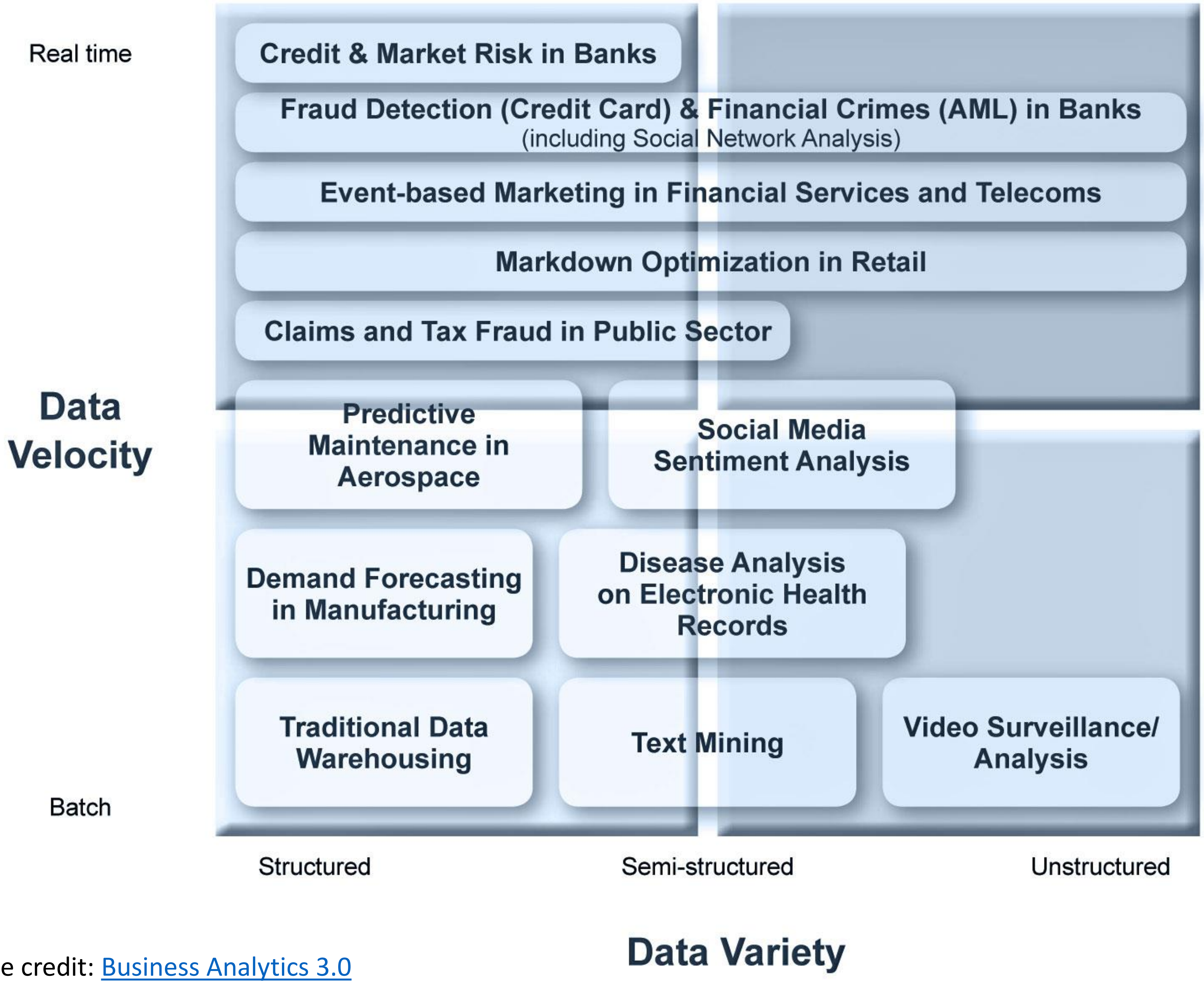
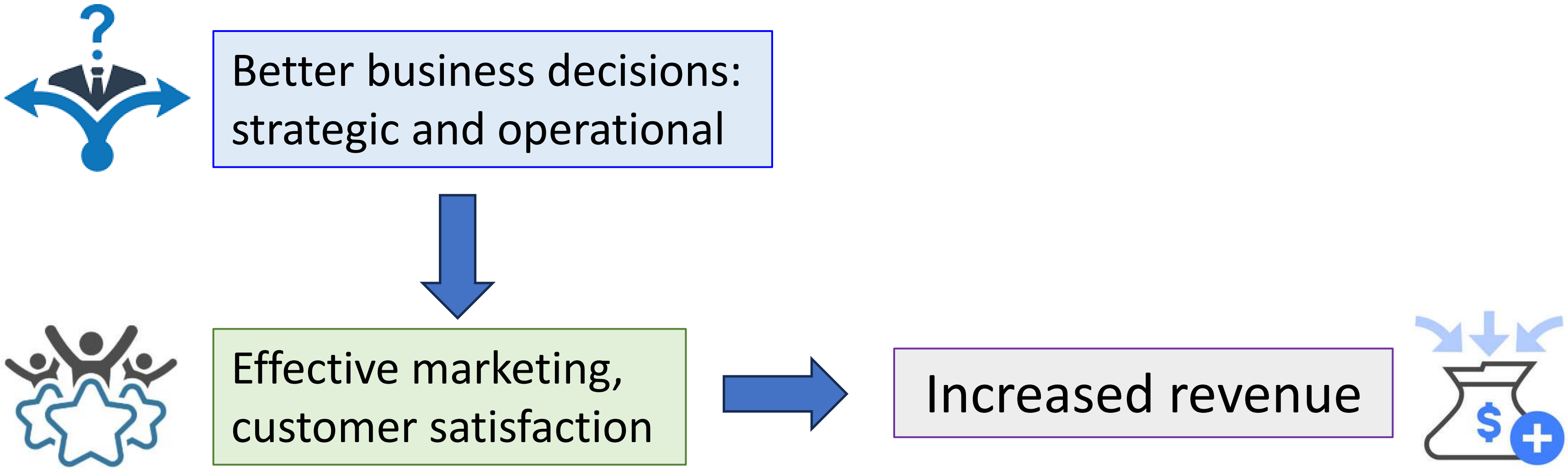

Real time
Credit & Market Risk in Banks
Fraud Detection (Credit Card) & Financial Crimes (AML) in Banks
(including Social Network Analysis)
Event-based Marketing in Financial Services and Telecoms
Markdown Optimization in Retail
Claims and Tax Fraud in Public Sector
Data Velocity
Predictive Maintenance in Aerospace
Social Media Sentiment Analysis
Demand Forecasting in Manufacturing
Disease Analysis on Electronic Health Records
Traditional Data Warehousing
Text Mining
Video Surveillance/ Analysis
Batch
Structured
Semi-structured
Unstructured
Data Variety

Image credit: Business Analytics 3.0

Big data analytics

- This analytics exploits a **large** amount of data for interesting data relationships to gain competitive advantage.

- Appropriate information: hidden patterns or correlations
- Competitive advantage:



Better business decisions: strategic and operational

Effective marketing, customer satisfaction

Increased revenue

Big data analytics

- Big data is more real-time in nature than traditional data warehouse applications.
  - Conventional architectures are ill-suited for big data apps (e.g., Exadata, Teradata)
- There are many challenges in handling Big data, mainly about the technology bottleneck and the lack of experts.

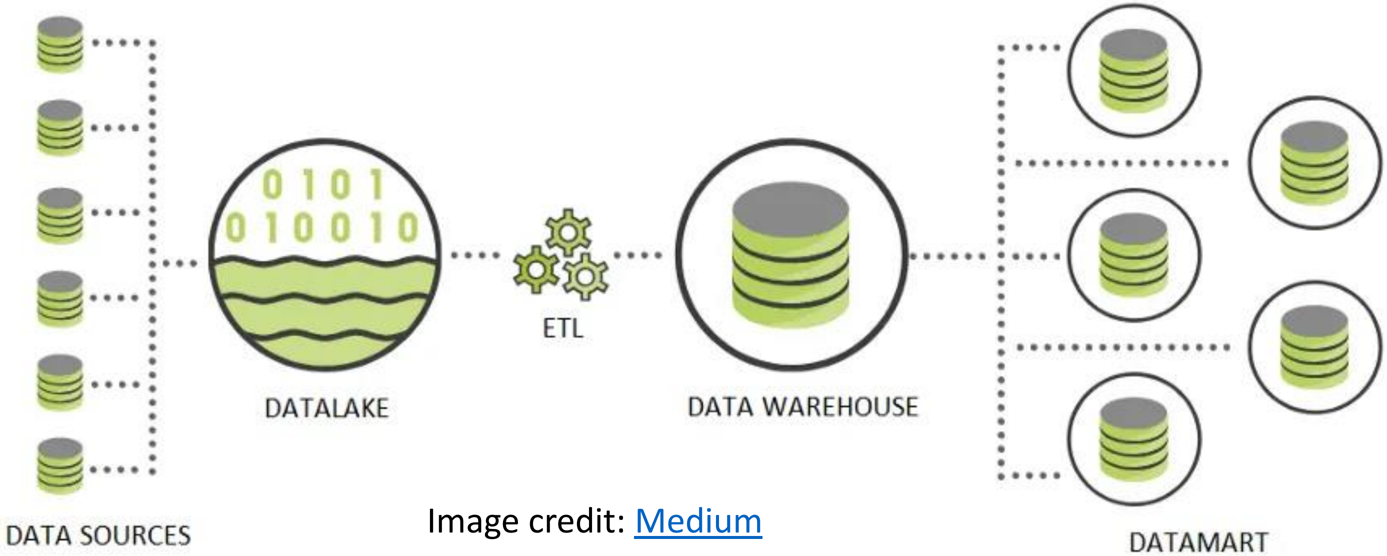
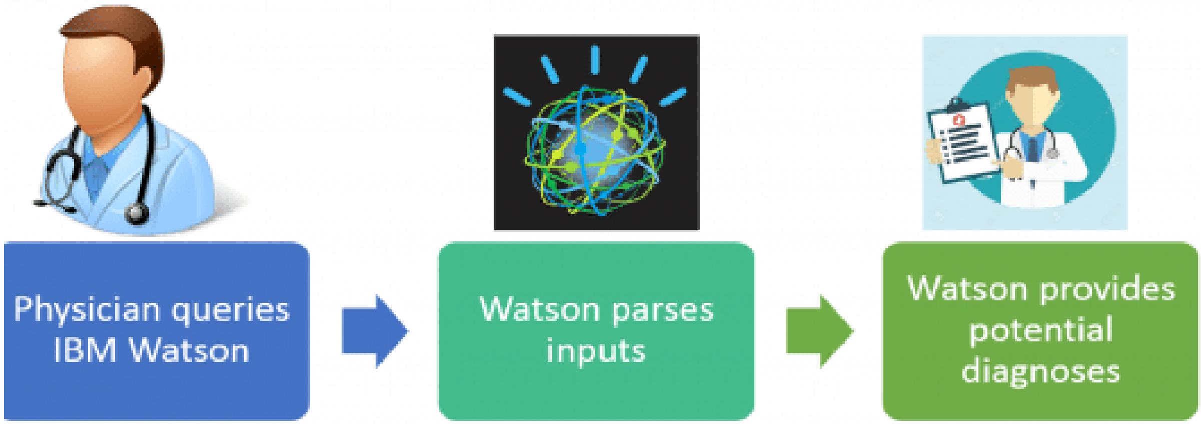


Image credit: Medium

Big data in Healthcare

- 80% of medical data is unstructured and clinically relevant, residing in multiple places.
    - Individual EMRs, labs and imaging systems, physician notes, medical correspondence, etc.
- Big data may help increase access to healthcare.
    - Build sustainable healthcare systems, collaborate to improve care and outcomes.

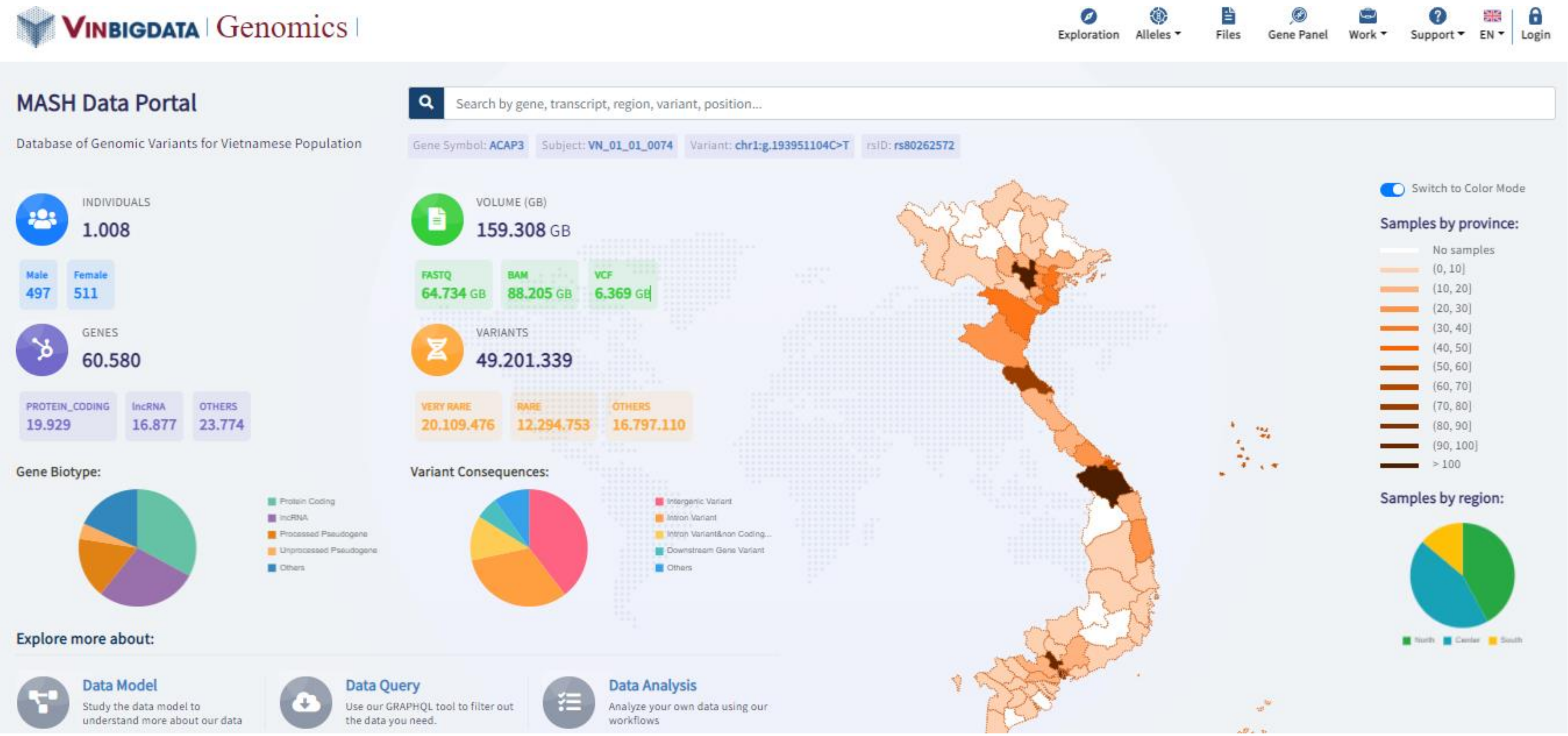


Physician queries IBM Watson

Watson parses inputs

Watson provides potential diagnoses

Big data in Healthcare



MASH Data Portal

Database of Genomic Variants for Vietnamese Population

Search by gene, transcript, region, variant, position...

Gene Symbol: ACAP3 | Subject: VN_01_01_0074 | Variant: chr1:g.193951104C>T | rsID: rs80262572

INDIVIDUALS
1.008

Male 497 | Female 511

GENES
60.580

PROTEIN_CODING 19.929 | lncRNA 16.877 | OTHERS 23.774

VOLUME (GB)
159.308 GB

FASTQ 64.734 GB | BAM 88.205 GB | VCF 6.369 GB

VARIANTS
49.201.339

VERY RARE 20.109.476 | RARE 12.294.753 | OTHERS 16.797.110

Gene Biotype:

Variant Consequences:

Samples by province:

Samples by region:

Explore more about:

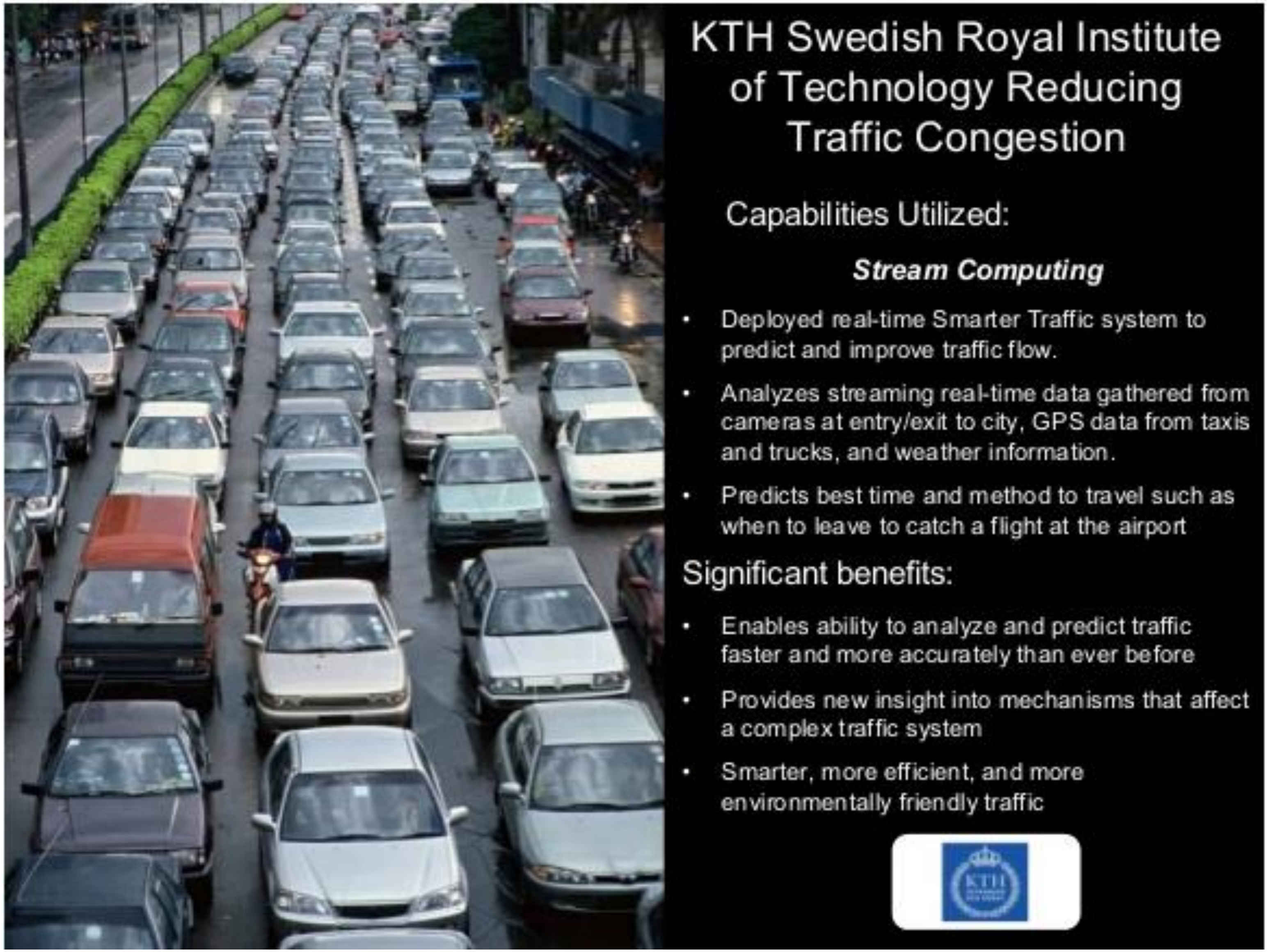
Data Model
Study the data model to understand more about our data

Data Query
Use our GRAPHQL tool to filter out the data you need.

Data Analysis
Analyze your own data using our workflows

VinBigData's Genomics project.
Homepage: https://genome.vinbigdata.org/ (Updated 2021)

KTH: Reducing traffic congestion



KTH Swedish Royal Institute of Technology Reducing Traffic Congestion

Capabilities Utilized:

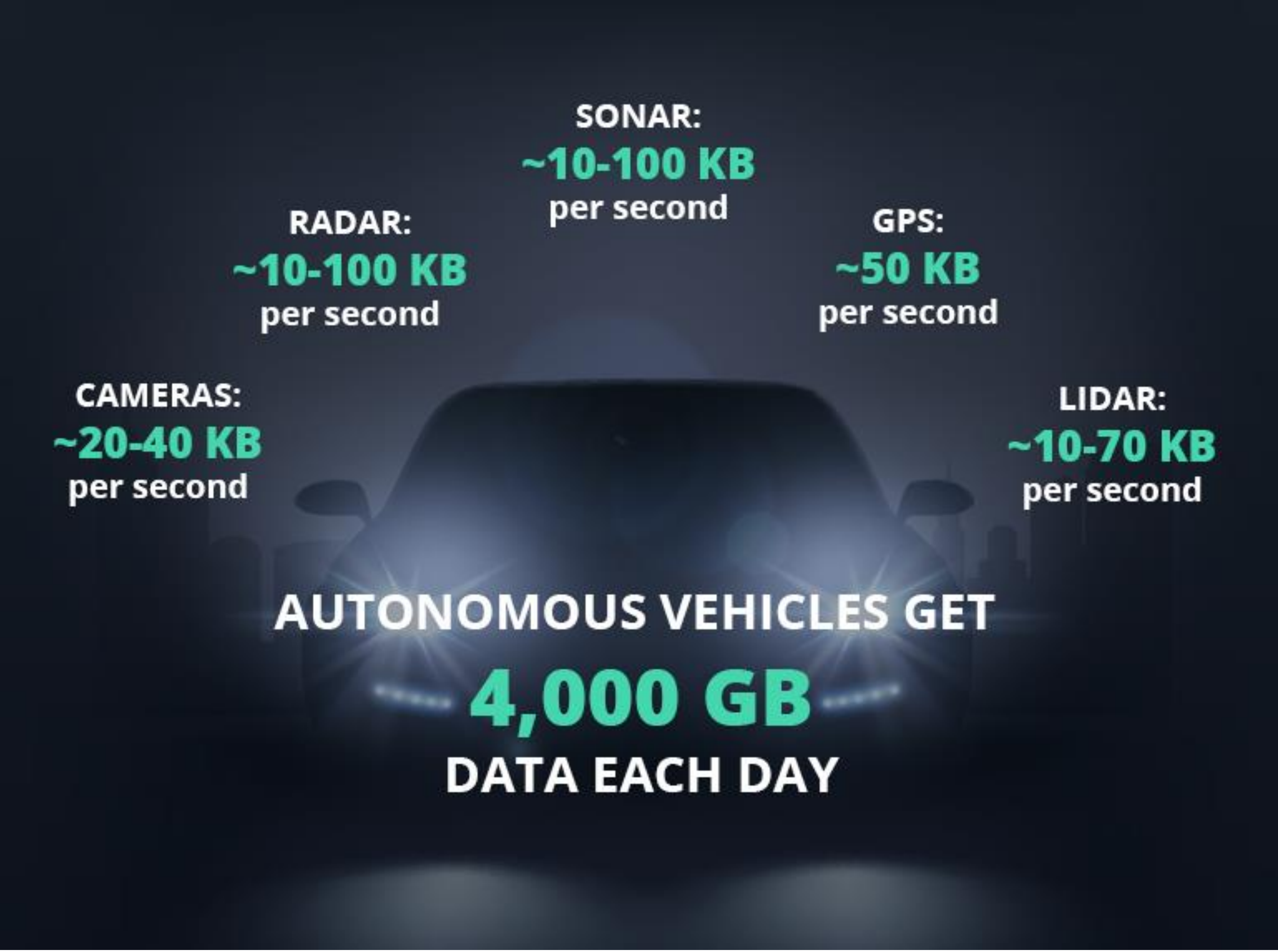
Stream Computing

- Deployed real-time Smarter Traffic system to predict and improve traffic flow.
- Analyzes streaming real-time data gathered from cameras at entry/exit to city, GPS data from taxis and trucks, and weather information.
- Predicts best time and method to travel such as when to leave to catch a flight at the airport

Significant benefits:

- Enables ability to analyze and predict traffic faster and more accurately than ever before
- Provides new insight into mechanisms that affect a complex traffic system
- Smarter, more efficient, and more environmentally friendly traffic

Source: SlideShare (updated 11/2012)



SONAR:
~10-100 KB
per second

RADAR:
~10-100 KB
per second

GPS:
~50 KB
per second

CAMERAS:
~20-40 KB
per second

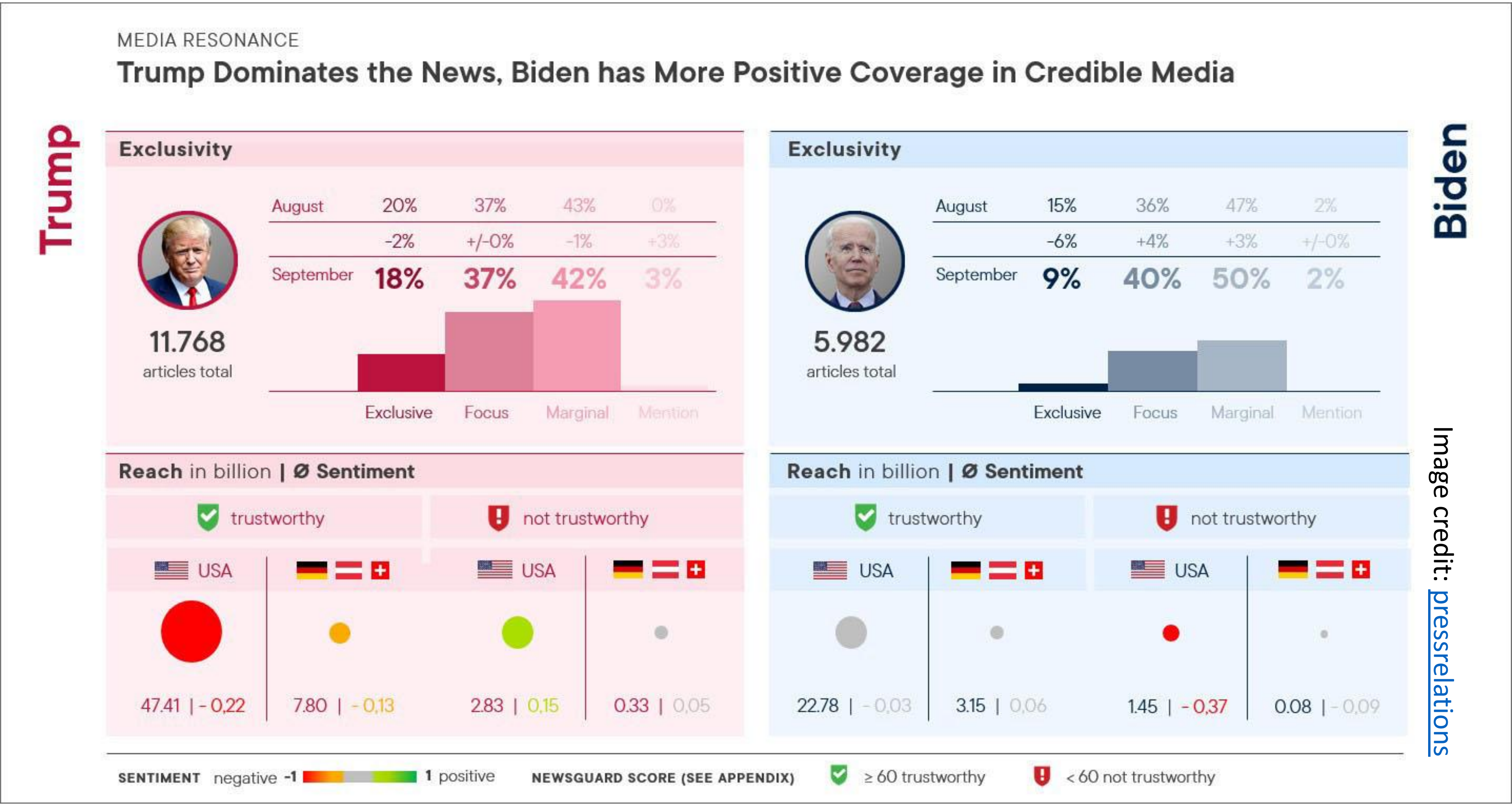
LIDAR:
~10-70 KB
per second

AUTONOMOUS VEHICLES GET

4,000 GB

DATA EACH DAY

Big data in social network analysis



MEDIA RESONANCE

Trump Dominates the News, Biden has More Positive Coverage in Credible Media

Trump

Exclusivity

11.768 articles total

	Exclusive	Focus	Marginal	Mention
August	20%	37%	43%	0%
	-2%	+/-0%	-1%	+3%
September	18%	37%	42%	3%

Reach in billion | Ø Sentiment

✓ trustworthy		! not trustworthy	
🇺🇸 USA	🇩🇪 🇦🇹 🇨🇭	🇺🇸 USA	🇩🇪 🇦🇹 🇨🇭
47.41 | - 0,22	7.80 | - 0,13	2.83 | 0,15	0.33 | 0,05

Biden

Exclusivity

5.982 articles total

	Exclusive	Focus	Marginal	Mention
August	15%	36%	47%	2%
	-6%	+4%	+3%	+/-0%
September	9%	40%	50%	2%

Reach in billion | Ø Sentiment

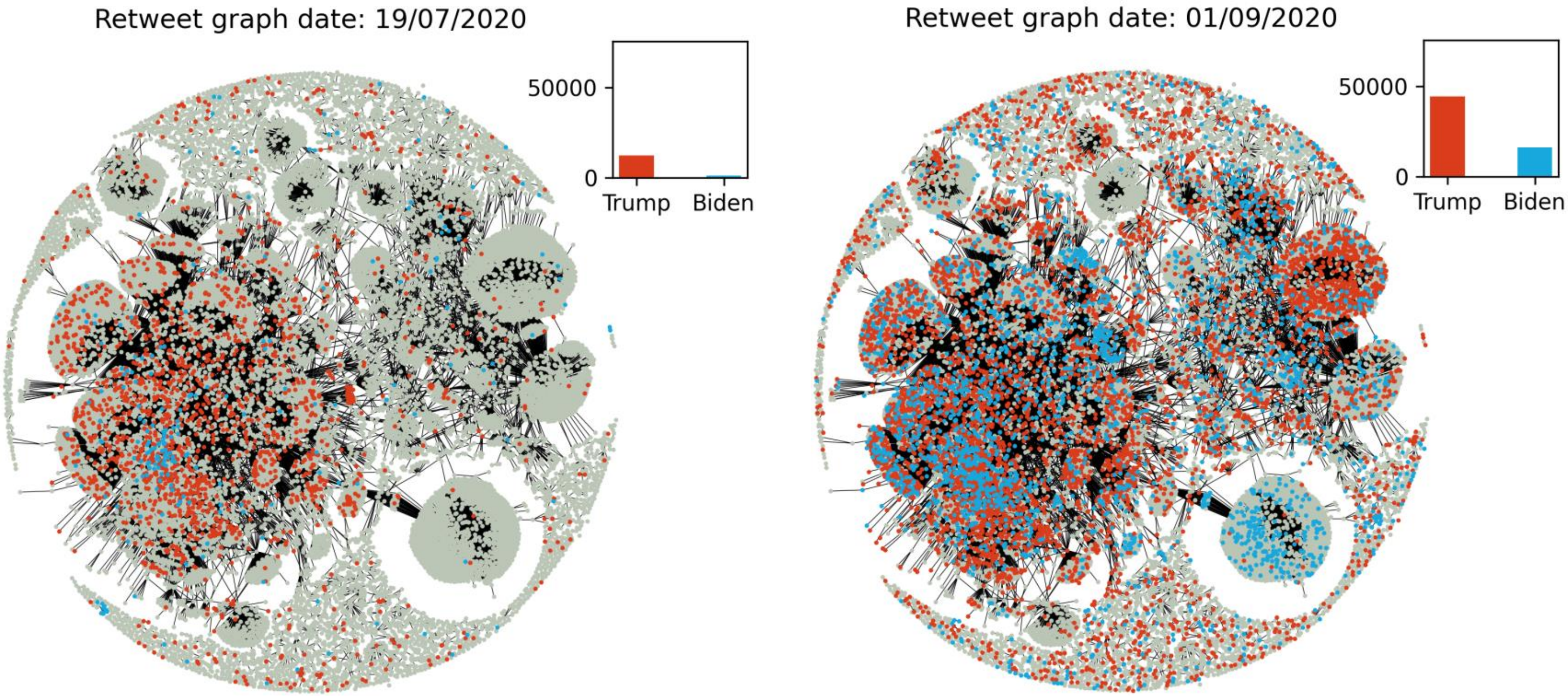
✓ trustworthy		! not trustworthy	
🇺🇸 USA	🇩🇪 🇦🇹 🇨🇭	🇺🇸 USA	🇩🇪 🇦🇹 🇨🇭
22.78 | - 0,03	3.15 | 0,06	1.45 | - 0,37	0.08 | - 0,09

SENTIMENT negative -1 1 positive NEWSGUARD SCORE (SEE APPENDIX) ✓ ≥ 60 trustworthy ! < 60 not trustworthy

Image credit: pressrelations

The trend has even strengthened compared to the previous month. The more trustworthy the media outlet, the more positive is its coverage of Biden. DACH media report neutrally to positively about Biden and predominantly negatively about Trump.

Big data in social network analysis



Retweet graph date: 19/07/2020

Retweet graph date: 01/09/2020

Shevtsov, Alexander, et al. "Analysis of Twitter and YouTube during US elections 2020." *arXiv e-prints* (2020): arXiv-2010.

Motivations and opportunities

A new horizon that changes our lives…

New insight into data

- Why deal with more data? ***New insights.***
- The insights are for people throughout the enterprise, not top-level executives only.
    - People involve may include the CEO, marketing staffs, data analyst and programmers.
- They can be transformed into actionable intelligence

  → better business, better service to customers

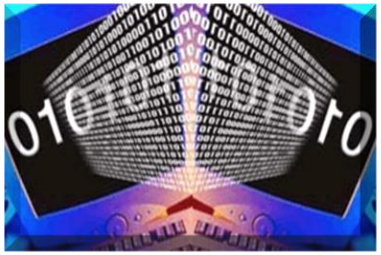
  → more profit gained

Big data analytics: Applications

Smarter Healthcare

Multi-channel sales

Finance



Log Analysis

Homeland Security

Traffic Control

Telecom

Search Quality

Manufacturing

Trading Analytics

Fraud and Risk

Retail: Churn, NBO

Big data analytics: Applications

amazon

310 million active users globally

12 million items across all its categories and services

Over 1 billion monthly active users

There were a total of 8.6 billion videos uploaded in TikTok in 2021. The number in 2023 approximates 14.4 billion videos.

TikTok

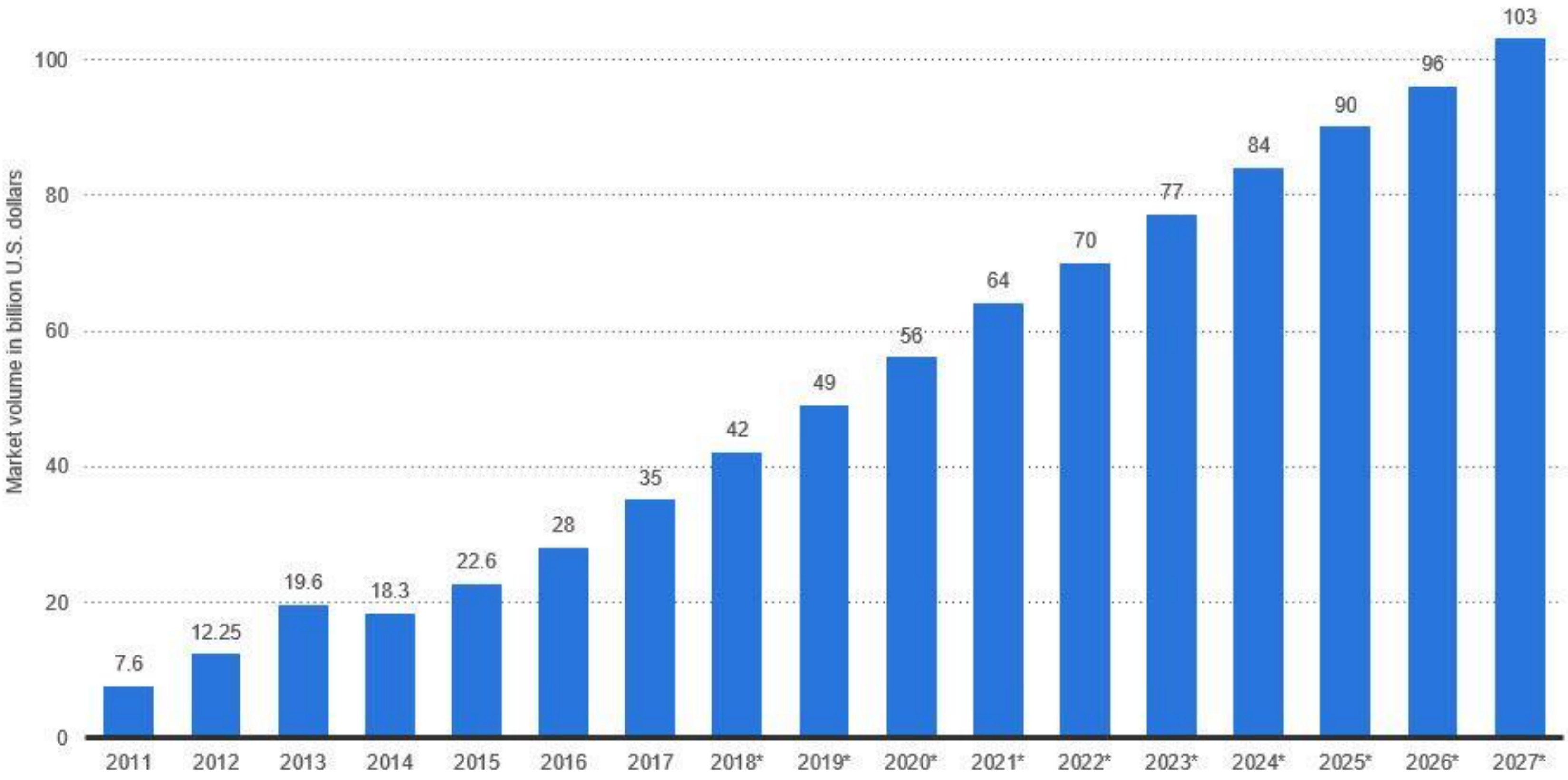
STEAM®

20,919,374 players online, 5,163,503 players in-game

105,281 tracked Steam games

Forecast Revenue Big Data Market Worldwide 2011-2027

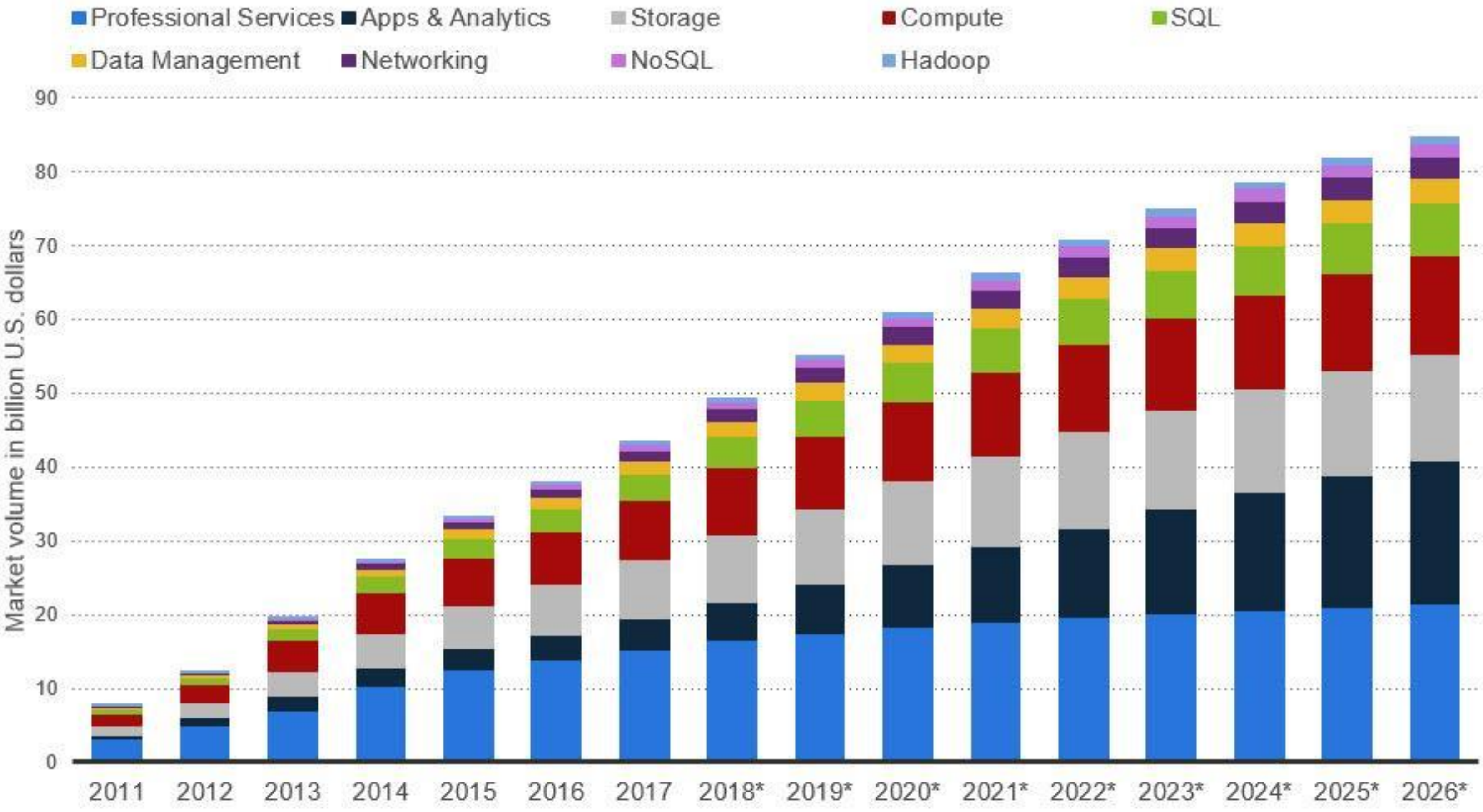
Big Data Market Size Revenue Forecast Worldwide From 2011 To 2027 (in billion U.S. dollars)



statista

Big Data Market Worldwide Segment Revenue Forecast 2011-2026

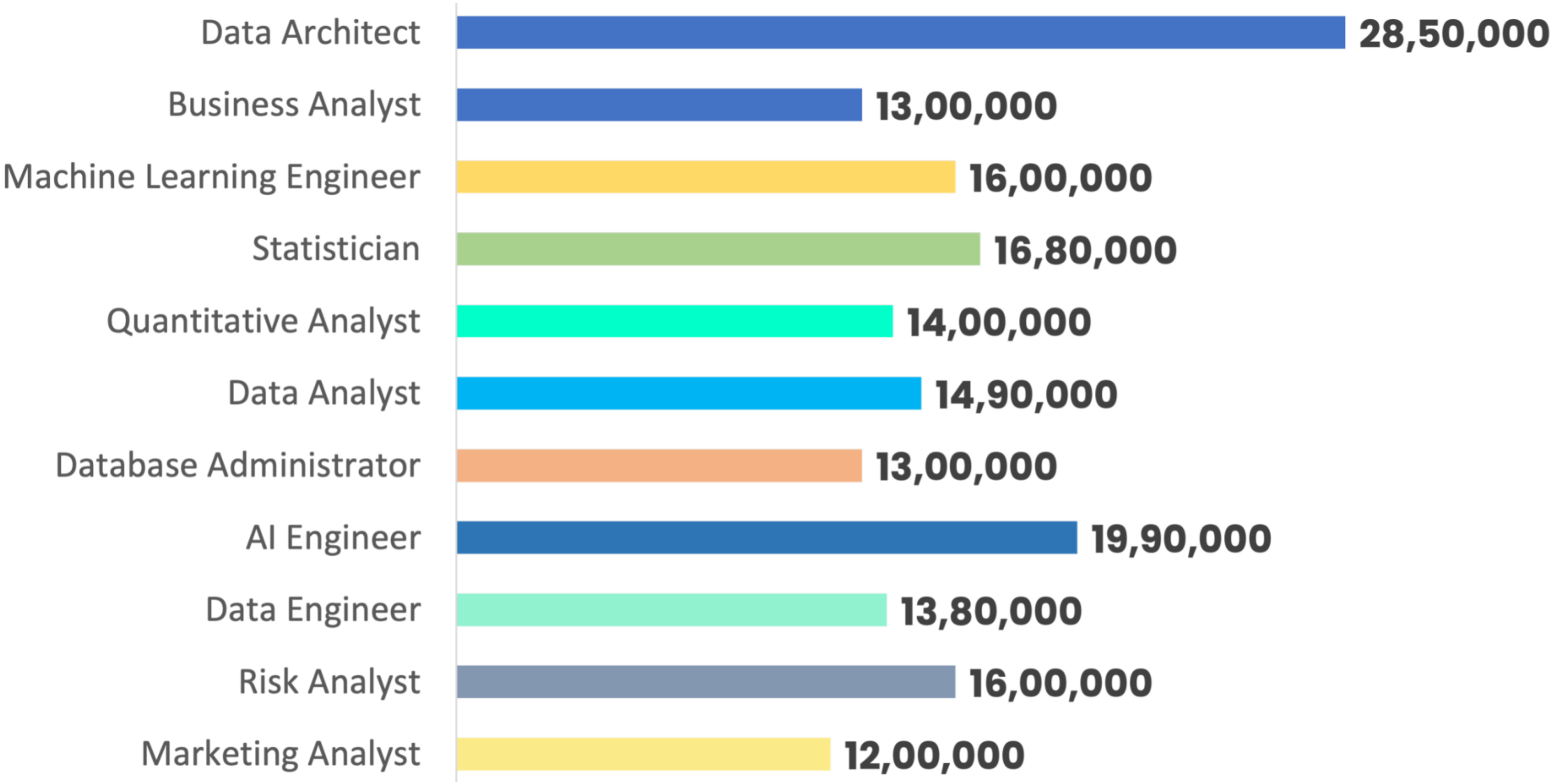
Big Data Market Forecast Worldwide from 2011 to 2026, by segment (in billion U.S. dollars)



statista

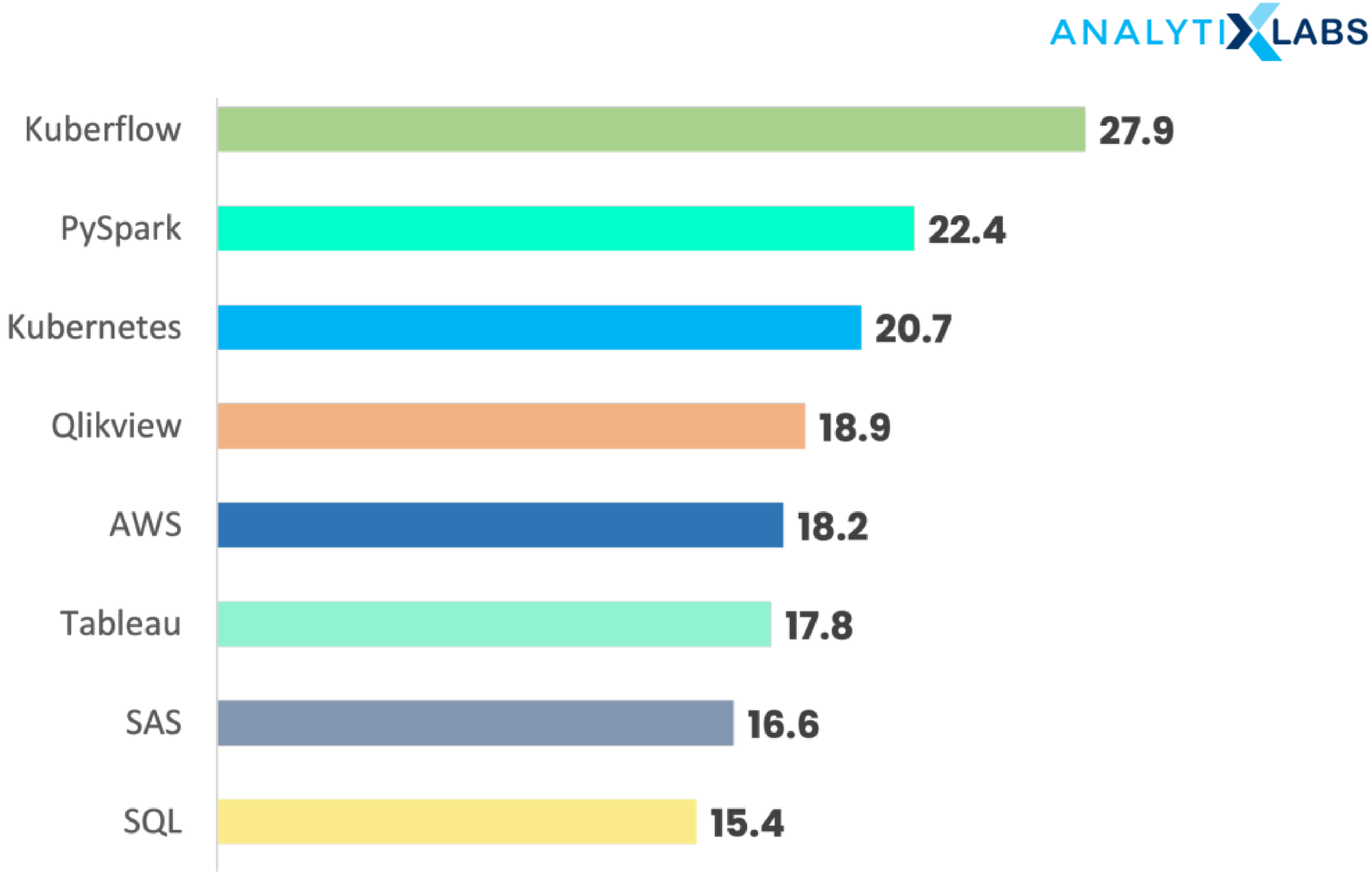
Salary by Job roles in India (July 2023)

ANALYTIXLABS



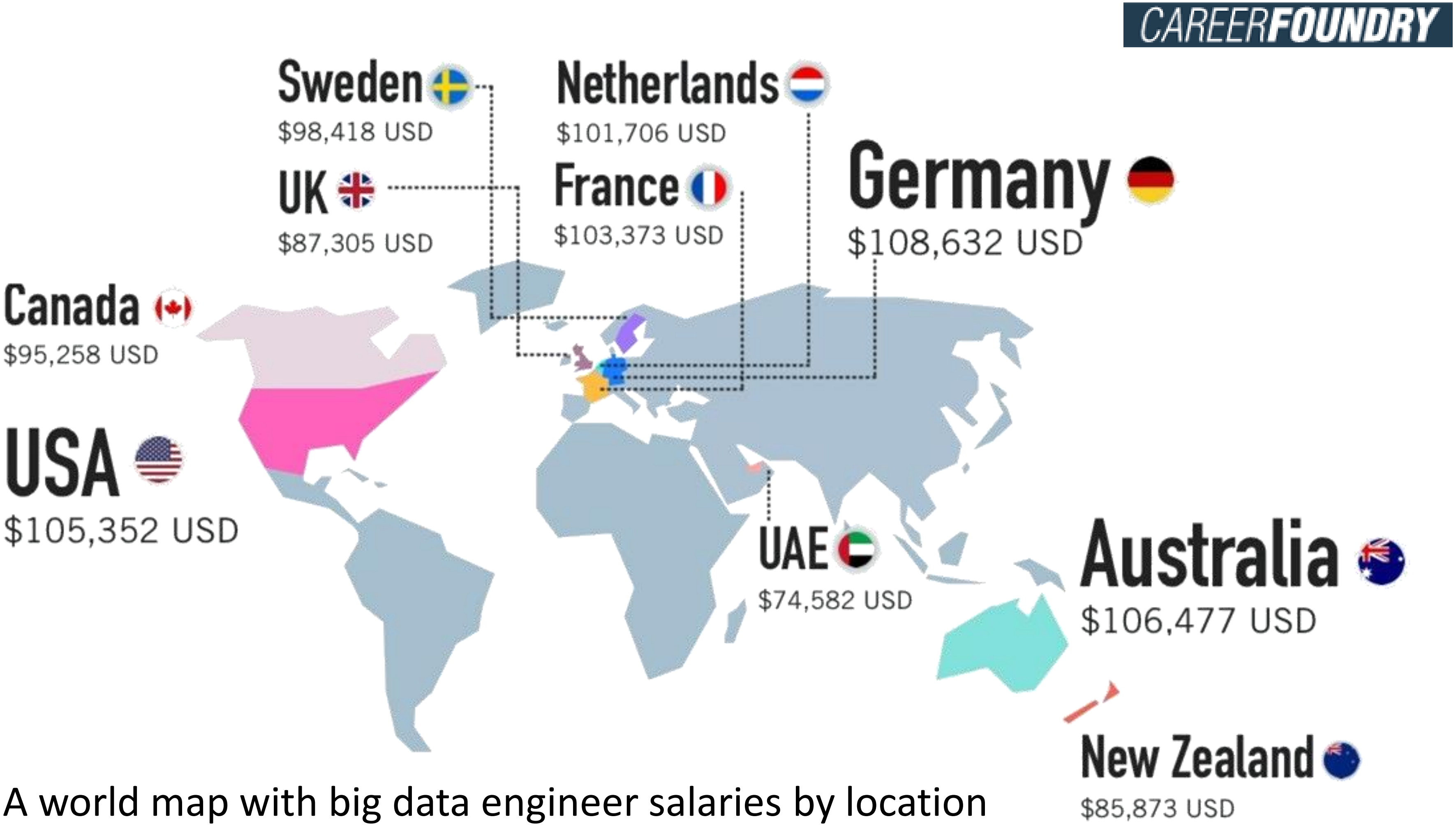
* The currency unit is INR (Indian Rupee)

Salary by tools in India (July 2023)



* The currency unit is Lakhs in INR (Indian Rupee)

Big data engineer salaries by location



A world map with big data engineer salaries by location (updated in January 2023)

Average big data engineer salary

CAREERFOUNDRY

(updated in January 2023)

Meta	$229K
Google	$205K
Microsoft	$183K
amazon	$167K
Apple	$170K

...the end.