Fundamental of Big Data Analytics

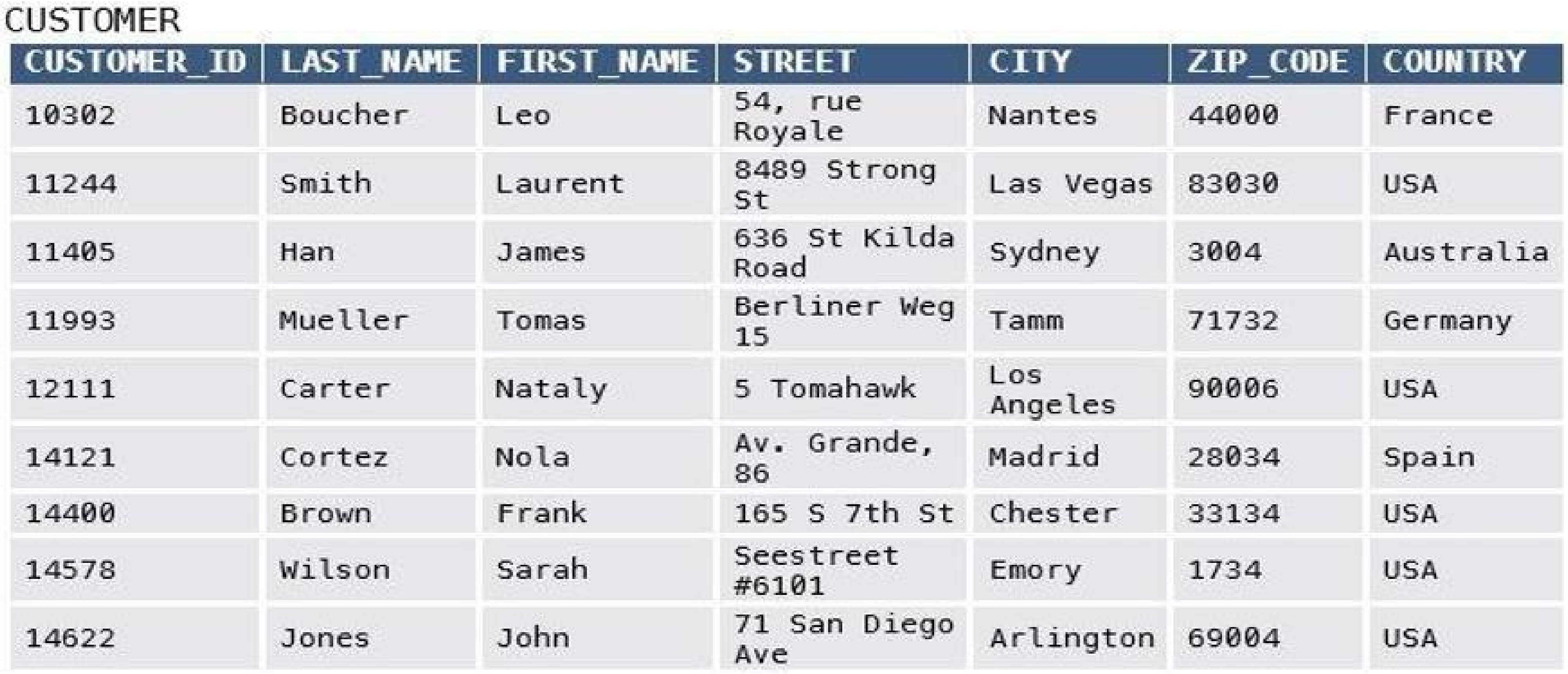
Lec 01

Data: Data is collection of raw fact and figures.

Types of data

Structure data	Un-structure data	Semi-structure data
quantitative data that consists of numbers and values.	qualitative data that consists of audio, video, sensors, descriptions, and more	that lies midway between structured and unstructured data
used in machine learning and drives machine learning algorithms.	used in natural language processing and text mining	
stored in tabular formats like excel sheets or SQL databases.	Stored as audio files, videos files, or NoSQL databases , pdf, word	
pre-defined data model.	does not have a pre-defined data model.	
requires less storage space and is difficult to scale.	requires more storage space and is difficult to scale.	

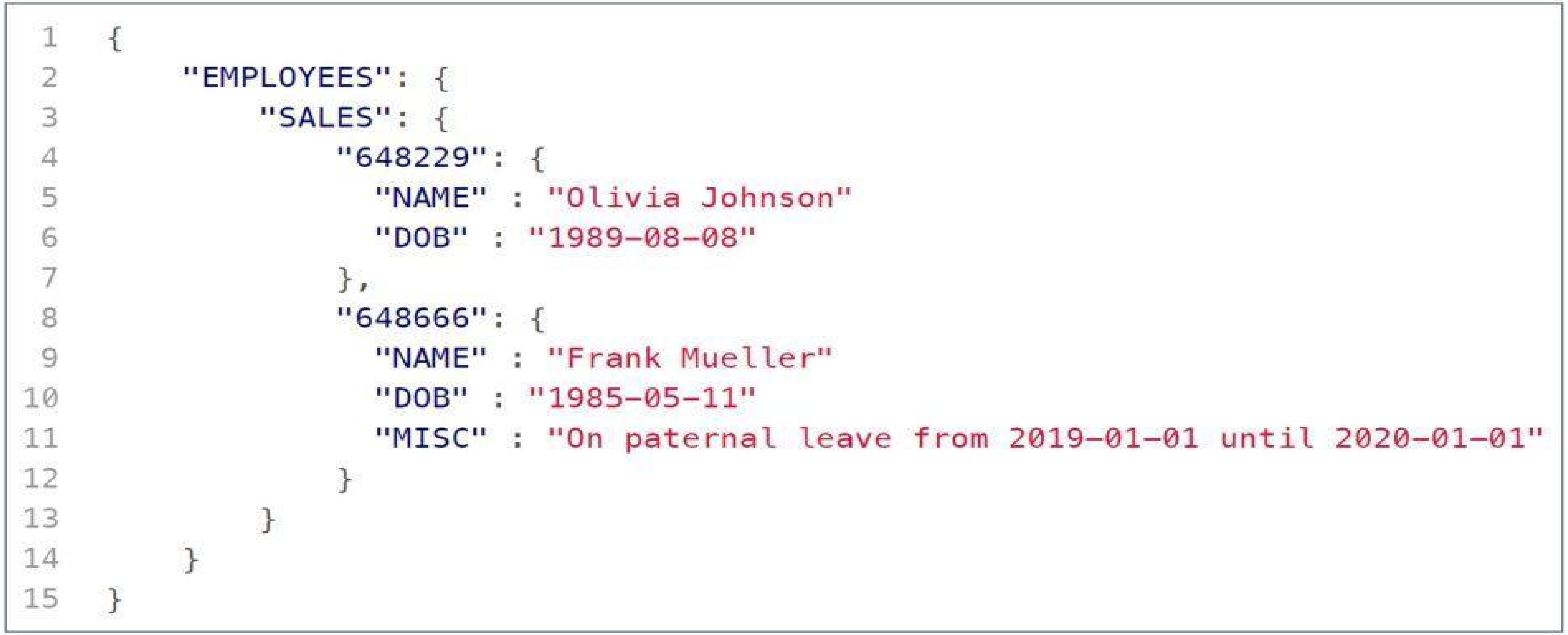
Structure Data



CUSTOMER

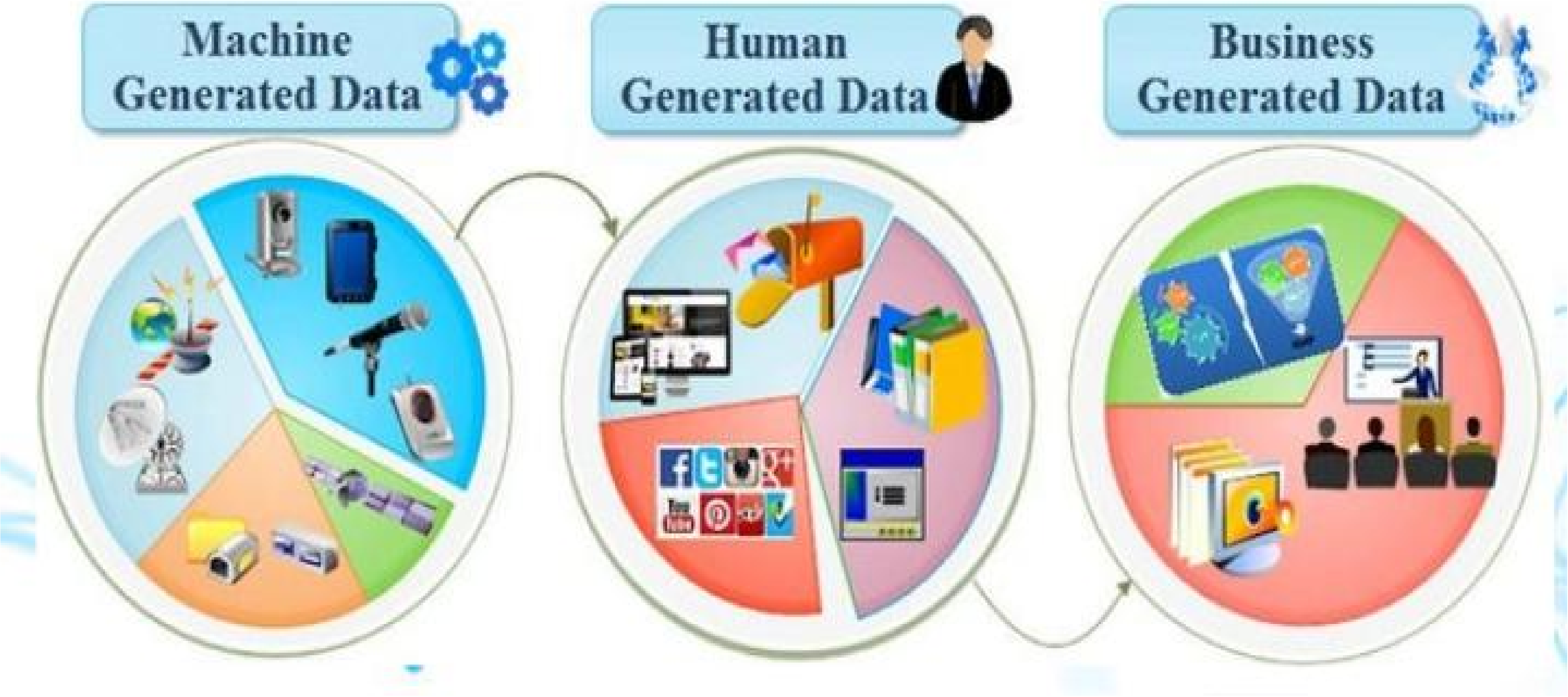
CUSTOMER_ID	LAST_NAME	FIRST_NAME	STREET	CITY	ZIP_CODE	COUNTRY
10302	Boucher	Leo	54, rue Royale	Nantes	44000	France
11244	Smith	Laurent	8489 Strong St	Las Vegas	83030	USA
11405	Han	James	636 St Kilda Road	Sydney	3004	Australia
11993	Mueller	Tomas	Berliner Weg 15	Tamm	71732	Germany
12111	Carter	Nataly	5 Tomahawk	Los Angeles	90006	USA
14121	Cortez	Nola	Av. Grande, 86	Madrid	28034	Spain
14400	Brown	Frank	165 S 7th St	Chester	33134	USA
14578	Wilson	Sarah	Seestreet #6101	Emory	1734	USA
14622	Jones	John	71 San Diego Ave	Arlington	69004	USA

Simi-Structure Data



```
{
    "EMPLOYEES": {
        "SALES": {
            "648229": {
              "NAME" : "Olivia Johnson"
              "DOB" : "1989-08-08"
            },
            "648666": {
              "NAME" : "Frank Mueller"
              "DOB" : "1985-05-11"
              "MISC" : "On paternal leave from 2019-01-01 until 2020-01-01"
            }
        }
    }
}
```

Sources of Data



What is Big Data?

Big Data

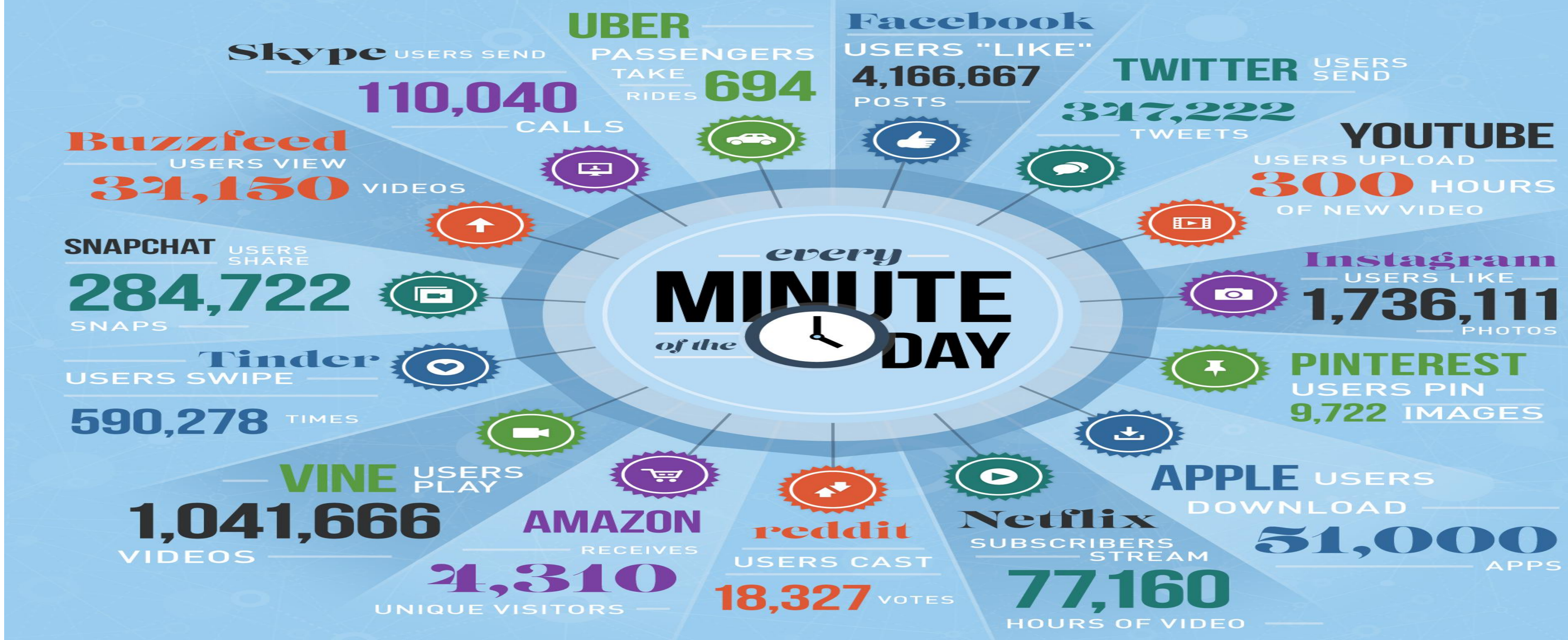
- Massive amount of data which **cannot** be stored, processed and analyzed using **traditional tools** is know as big data.
- Data that contains greater **variety**, arriving in increasing volumes and with more velocity.
- It is in the combination of structured, unstructured and semi-structured data.
- Hadoop, Spark is the solution of storing Big Data. It stored data in **distributed system** and process data using parallel processing methods.

DOMO

DATA NEVER SLEEPS 3.0

How much data is generated **every minute?**

Data is being created all the time without us even noticing it. Much of what we do every day now happens in the digital realm, leaving an ever-increasing digital trail that can be measured and analyzed. Just how much data do our tweets, likes and photo uploads really generate? For the third time, Domo has the answer—and the numbers are staggering.



every MINUTE of the DAY

Skype USERS SEND 110,040 CALLS

UBER PASSENGERS TAKE 694 RIDES

Facebook USERS "LIKE" 4,166,667 POSTS

TWITTER USERS SEND 347,222 TWEETS

YOUTUBE USERS UPLOAD 300 HOURS OF NEW VIDEO

Instagram USERS LIKE 1,736,111 PHOTOS

PINTEREST USERS PIN 9,722 IMAGES

APPLE USERS DOWNLOAD 51,000 APPS

Netflix SUBSCRIBERS STREAM 77,160 HOURS OF VIDEO

reddit USERS CAST 18,327 VOTES

AMAZON RECEIVES 4,310 UNIQUE VISITORS

VINE USERS PLAY 1,041,666 VIDEOS

Tinder USERS SWIPE 590,278 TIMES

SNAPCHAT USERS SHARE 284,722 SNAPS

Buzzfeed USERS VIEW 34,150 VIDEOS

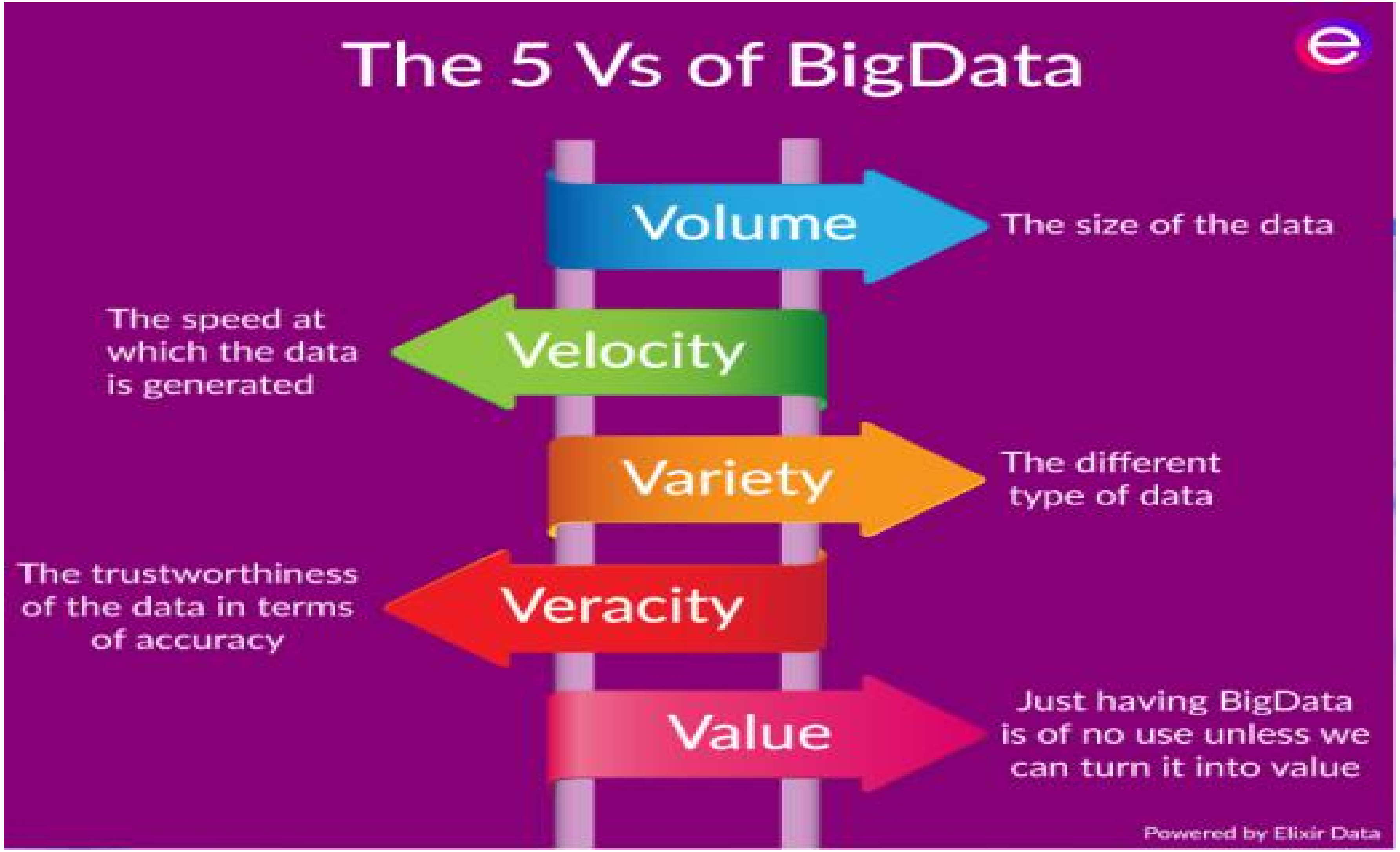
THE GLOBAL INTERNET POPULATION GREW 18.5% FROM 2013–2015 AND NOW REPRESENTS **3.2 BILLION PEOPLE.**

With each click, share and like, the world's data pool is expanding faster than we can comprehend. Businesses today are paying attention to scores of data sources to make crucial decisions about the future. The team at Domo can help your business make sense of this endless stream of data by providing executives with all their critical information in one intuitive platform. Domo delivers the insights you need to transform the way you run your business. **Learn more at www.domo.com.**

DOMO

SOURCES:
FACEBOOK, TWITTER, YOUTUBE, INSTAGRAM, PINTEREST, APPLE, NETFLIX, REDDIT, AMAZON, TINDER, BUZZFEED, MARKETINGPROFS, STATISTICBRAIN.COM

How do we Classify any data as Big Data



The 5 Vs of BigData

Volume

The size of the data

Velocity

The speed at which the data is generated

Variety

The different type of data

Veracity

The trustworthiness of the data in terms of accuracy

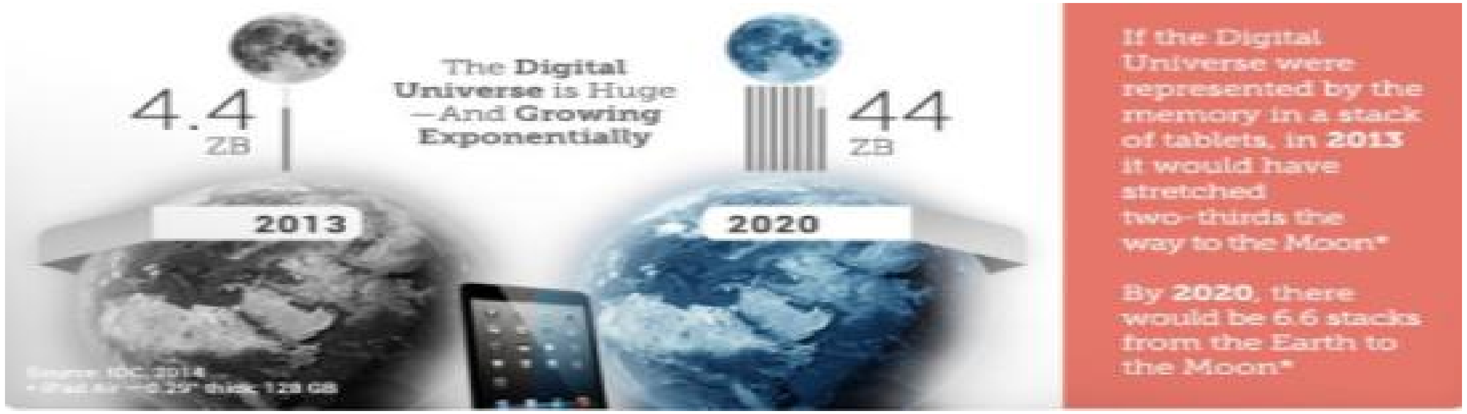
Value

Just having BigData is of no use unless we can turn it into value

Powered by Elixir Data

Volume

- Volume: size, scale, dimensionality,
- 204m emails/minute, if an email is 100KB, see the volume



If the Digital Universe were represented by the memory in a stack of tablets, in 2013 it would have stretched two-thirds the way to the Moon*

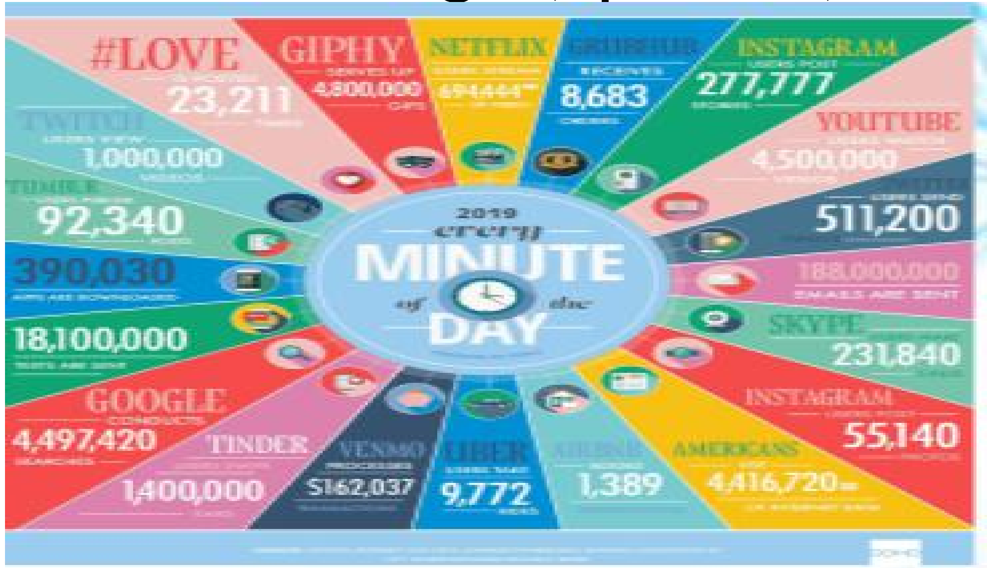
By 2020, there would be 6.6 stacks from the Earth to the Moon*

Source: IDC, 2014
* iPad Air – 0.29" thick: 128 GB

- **Challenges:** Acquisition, Storage, Retrieval, Processing Time
- Large dimensional data has more information, it is a blessing
- It is a also a **big curse**, dealing with large dimensions is a core topic in this course

Velocity

- Speed of data is very high
- Number of emails, twitter messages, photos, videos etc. per second



- Late decisions implies missed opportunities
- Real time processing vs Batch Processing

Variety

- Structural variety, different formats, models
- Medium variety, audio, text, video,
- DBMS, files, traffic logs, XML, code
- Online vs Offline,
- Real time vs Intermittent data (another way data varies)
- Challenges: requirement of analytics, Semantic, how to interpret

Veracity

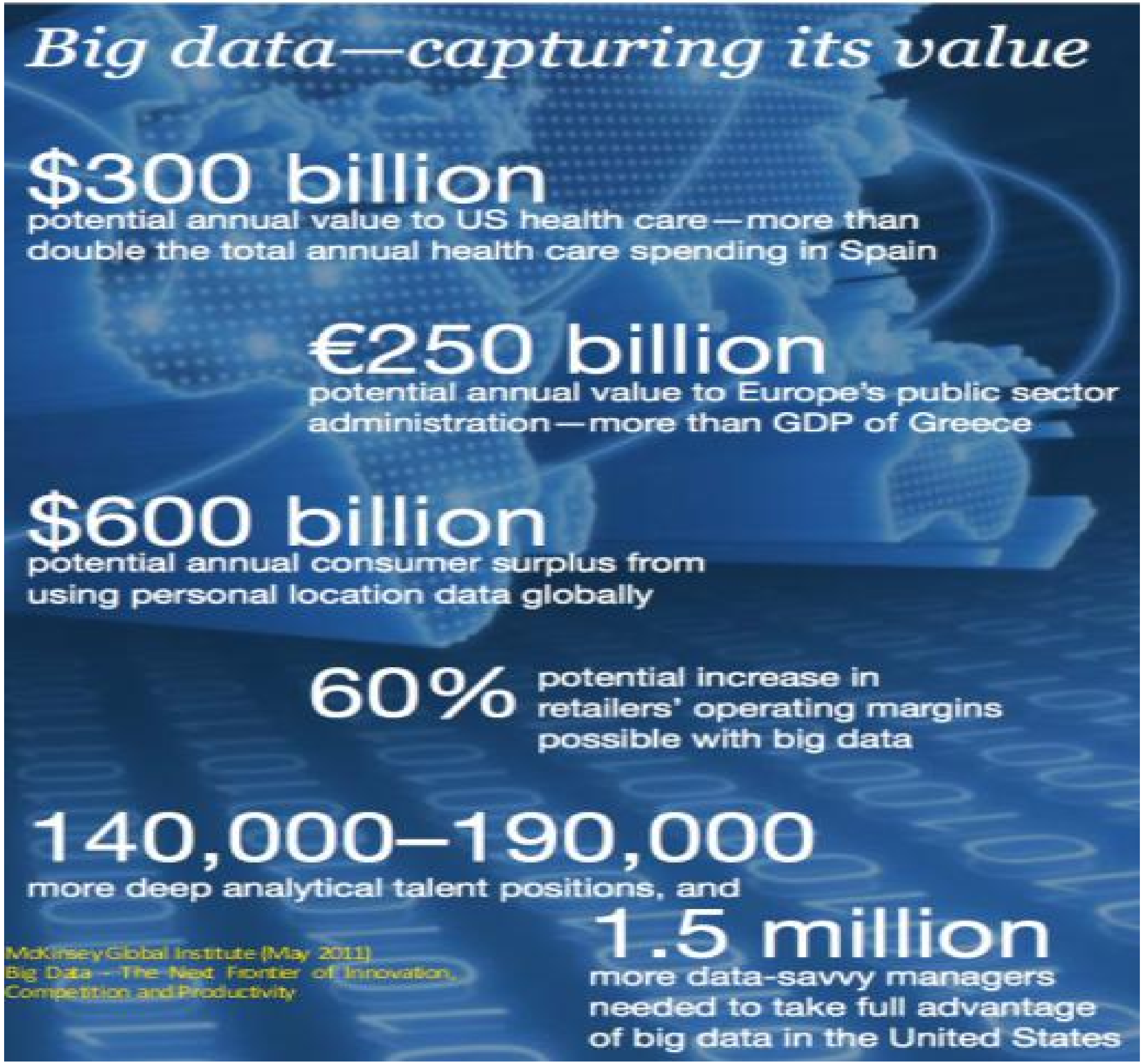
- Quality of data
- Data could have many issues (biases, anomalies, inconsistent measurements and units, incomplete and duplicate records)
- Volatility in data, updated/outdated, changing trends/ sentiments
- **Trustworthiness** and reliability of sources and generation/ processing
- Fake news, rumours, fake likes, fake followers

Value

Value refers to abality to turn your data **useful for business.**

The **Economist Intelligence Unit** report on surveying **476 executives**

- 60% feel that data is generating revenue within their organizations
- 83% say it is making existing services and products more profitable
- 63% executives based in Asia said they are routinely generating value from data
- In the US, the figure was 58% and in Europe, 56%



Big data—capturing its value

$300 billion
potential annual value to US health care—more than double the total annual health care spending in Spain

€250 billion
potential annual value to Europe's public sector administration—more than GDP of Greece

$600 billion
potential annual consumer surplus from using personal location data globally

60% potential increase in retailers' operating margins possible with big data

140,000–190,000
more deep analytical talent positions, and

1.5 million
more data-savvy managers needed to take full advantage of big data in the United States

McKinsey Global Institute (May 2011)
Big Data – The Next Frontier of Innovation, Competition and Productivity

BIG DATA ANALYTICS

Analysis vs. Analytics

Analysis: Making use of past data and deriving results from the data

Analytics: Using data to obtain future insights and details regarding trends etc..

Data Analytics

- The process of examining data in order to draw and communicate useful conclusions about the information it contains.

Tools that used in Big Data Analytics

- **Hadoop**
- **spark**
- **mongoDB**
- talend
- kalfa
- Storm
- cassandra