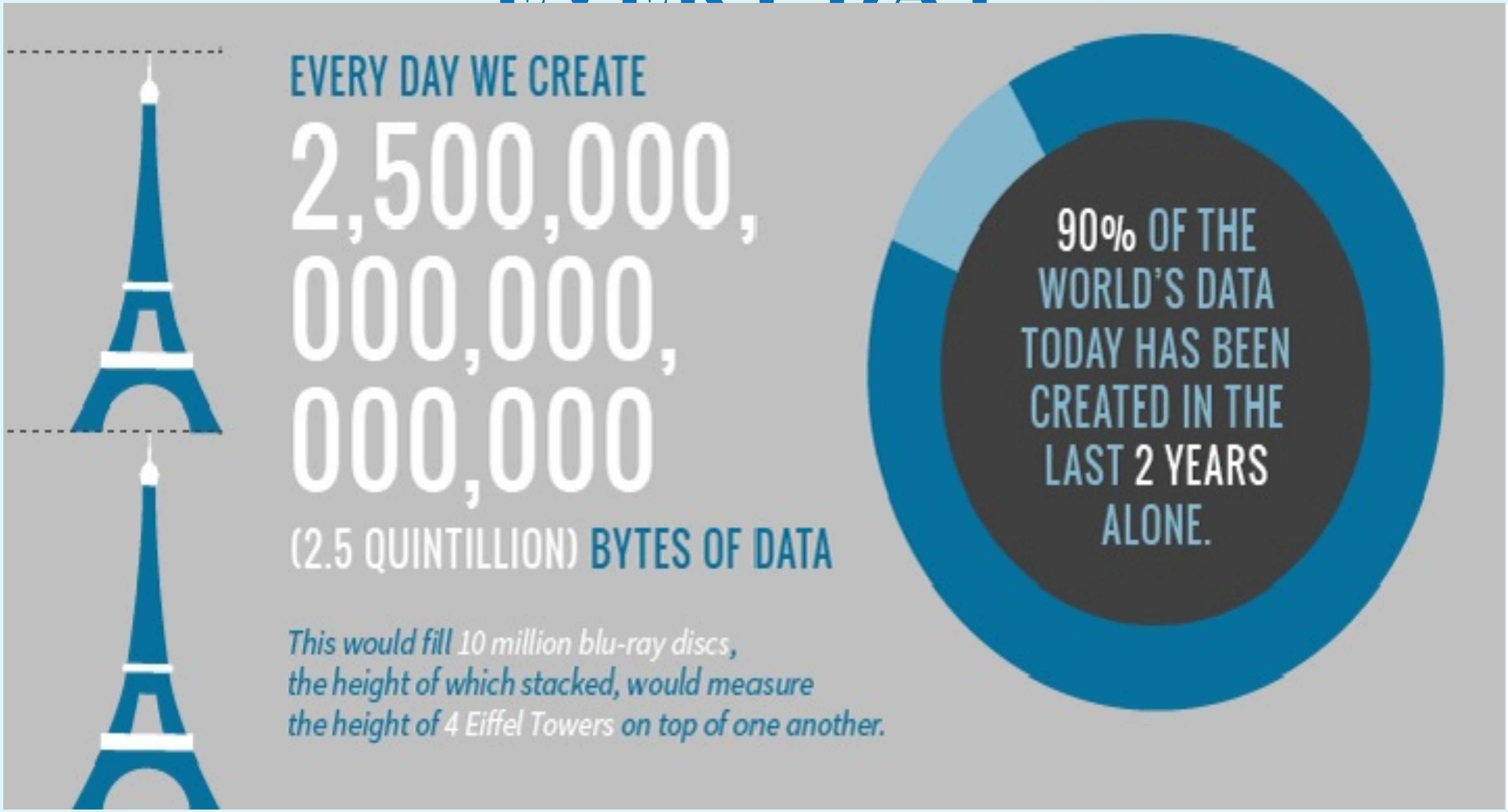
EVERYDAY BIG DATA

Big data describes the collection of complex and large data sets such that it's difficult to ***capture, process, store, search and analyze*** *using conventional data base systems. Its uses are shaping the world around us, offering more qualitative* ***insights*** *into our everyday lives.*

HOW MUCH DATA DO WE CREATE EVERY DAY?

- Data is everywhere !!
- Data never sleeps.
- We produce 2.5 quintillion bytes of data created each day at our current rate – rough estimate.
- 90% of the data in the world today has been created in the last two years alone.
- This pace is only accelerating with the growth of the Internet of Things (IoT).

HOW MUCH DATA DO WE CREATE EVERY DAY?



EVERY DAY WE CREATE

2,500,000,000,000,000,000,000

(2.5 QUINTILLION) BYTES OF DATA

This would fill 10 million blu-ray discs, the height of which stacked, would measure the height of 4 Eiffel Towers on top of one another.

90% OF THE WORLD'S DATA TODAY HAS BEEN CREATED IN THE LAST 2 YEARS ALONE.

Social Media

- Snapchat users share 527,760 photos
- More than 120 professionals join LinkedIn
- Users watch 4,146,600 YouTube videos
- 456,000 tweets are sent on Twitter
- Instagram users post 46,740 photos

These are numbers generated **every minute** of the day !!

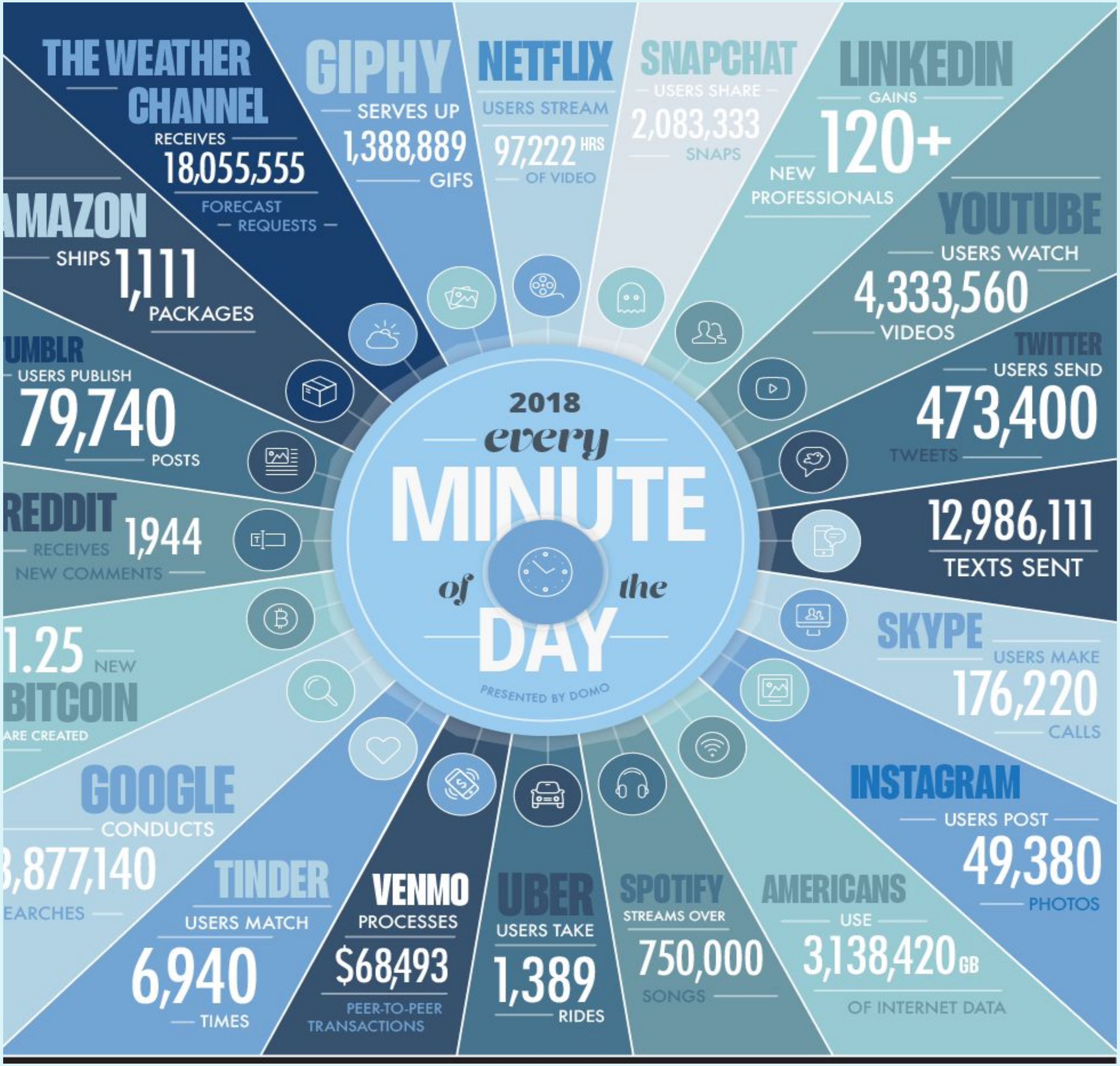
Facebook statistics

- 1.5 billion people are active on Facebook **daily.**
- There are five new Facebook profiles created every second!
- More than 300 million photos get uploaded per day.
- Every minute there are 510,000 comments posted and 293,000 statuses updated.

Volume of communication

(text to emails sent out every minute)

- We send 16 million text messages.
- There are 990,000 Tinder swipes.
- 156 million emails are sent; worldwide it is expected that there will be 9 billion email users by 2019.
- 15,000 GIFs are sent via Facebook messenger.
- Every minute there are 103,447,520 spam emails sent.
- There are 154,200 calls on Skype.


2018
every
MINUTE
of the
DAY
PRESENTED BY DOMO
THE WEATHER CHANNEL
RECEIVES 18,055,555
FORECAST REQUESTS
GIPHY
SERVES UP 1,388,889
GIFS
NETFLIX
USERS STREAM
97,222 HRS
OF VIDEO
SNAPCHAT
USERS SHARE
2,083,333
SNAPS
LINKEDIN
GAINS
120+
NEW PROFESSIONALS
YOUTUBE
USERS WATCH
4,333,560
VIDEOS
TWITTER
USERS SEND
473,400
TWEETS
12,986,111
TEXTS SENT
SKYPE
USERS MAKE
176,220
CALLS
INSTAGRAM
USERS POST
49,380
PHOTOS
AMERICANS
USE
3,138,420GB
OF INTERNET DATA
SPOTIFY
STREAMS OVER
750,000
SONGS
UBER
USERS TAKE
1,389
RIDES
VENMO
PROCESSES
$68,493
PEER-TO-PEER TRANSACTIONS
TINDER
USERS MATCH
6,940
TIMES
GOOGLE
CONDUCTS
,877,140
EARCHES
1.25 NEW
BITCOIN
ARE CREATED
REDDIT
RECEIVES 1,944
NEW COMMENTS
UMBLR
USERS PUBLISH
79,740
POSTS
MAZON
SHIPS 1,111
PACKAGES

FUTURE PREDICTIONS

By the year 2020 (not as far away as it sounds):

- 1.7 megabytes of new information will be created every second, per person.
- Around 1/3 of all data will be processed through the cloud.

- Today, less than 0.5% of available data is actually being analyzed.

SOURCES OF DATA

Data is obtained primarily from two sources :

- Internal Sources :

  Internal data is the information that the business already has on hand, has control of and currently owns, including details contained within the company’s own computer systems and cloud environments.

- External Sources:

  External data is information that is not currently owned by the company, and can include unstructured, public data as well as information gathered by other organizations.

TYPES OF DATA

- **Structured data**
- **Unstructured data**
- **Semi-structured data**

STRUCTURED DATA

“ Structured data is data that has been organized into a formatted repository, typically a database, so that its elements can be made addressable for more effective processing and analysis”.

“ Structured data is data that has a defined repeating pattern. This makes it easier for any program to sort, read and process the data”

Structured Data....

- Is organized data in a predefined format.
- Is stored in a tabular form.
- Is the data that resides in fixed fields within a record/file.
- Is formatted data that has entities & their attributes mapped.
- Is used to query and report against predetermined data types.

Sources : Relational Db’s, Legacy Db’s, MultiDimensional Db’s, Flat files in the form of records.

An 'Employee' Table In A Database Is An Example Of Structured Data

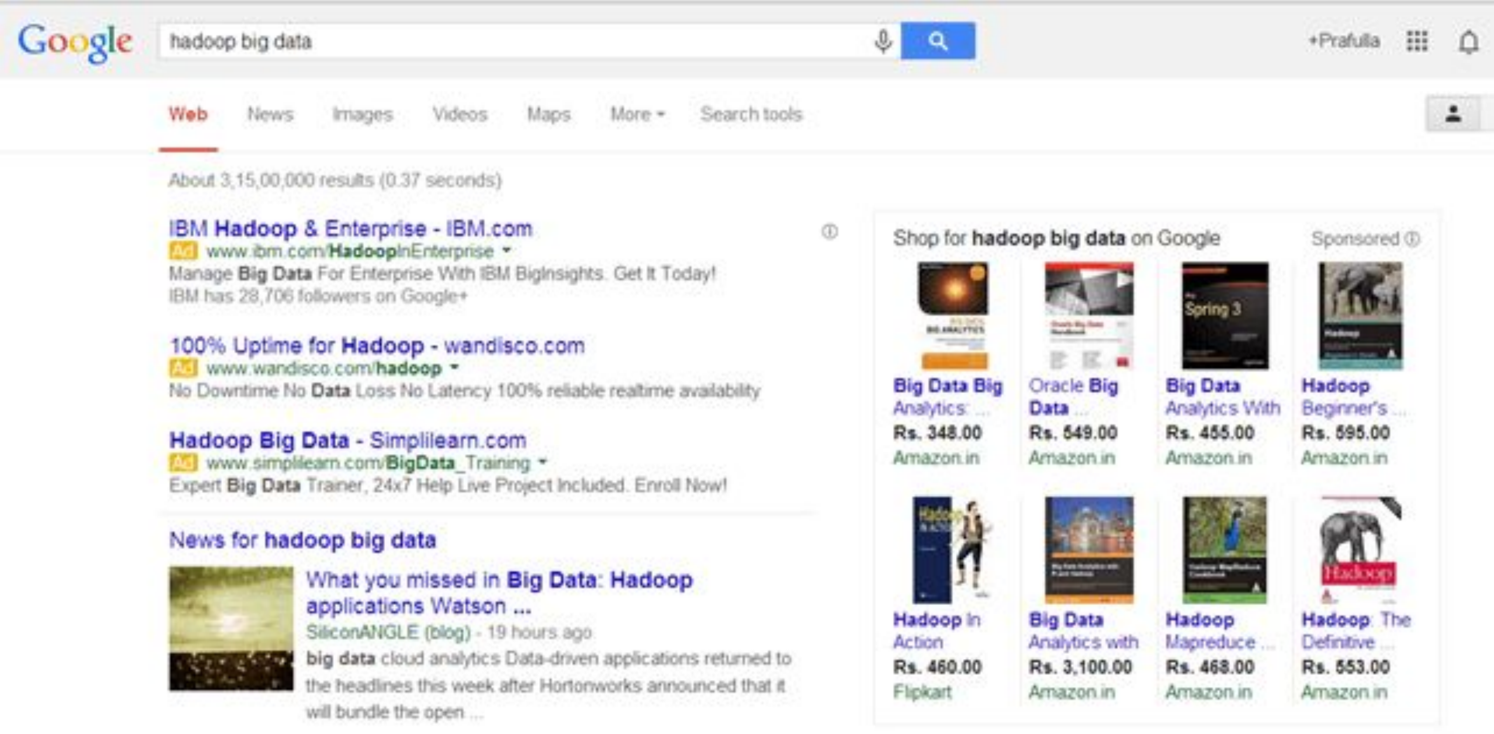
Employee_ID	Employee_Name	Gender	Department	Salary_In_lacs
2365	Rajesh Kulkarni	Male	Finance	650000
7465	Pratibha Joshi	Female	Admin	650000
7500	Shushil Roy	Male	Admin	500000

UNSTRUCTURED DATA

- Might/Might not have a repeating pattern.
- Typically consists of meta data.
- Comprises inconsistent data – data obtained from files, social media websites, satellites etc.
- Consists of data in different formats – emails, text, video, images etc

Sources : Text both internal and external to an organization, Social media, Mobile data.

EXAMPLE OF UN-STRUCTURED DATA



About 3,15,00,000 results (0.37 seconds)

IBM **Hadoop** & Enterprise - IBM.com
Ad www.ibm.com/**Hadoop**InEnterprise ▾
Manage **Big Data** For Enterprise With IBM BigInsights. Get It Today!
IBM has 28,706 followers on Google+

100% Uptime for **Hadoop** - wandisco.com
Ad www.wandisco.com/**hadoop** ▾
No Downtime No **Data** Loss No Latency 100% reliable realtime availability

Hadoop Big Data - Simplilearn.com
Ad www.simplilearn.com/**BigData**_Training ▾
Expert **Big Data** Trainer, 24x7 Help Live Project Included. Enroll Now!

News for **hadoop big data**

What you missed in **Big Data**: **Hadoop** applications Watson ...
SiliconANGLE (blog) - 19 hours ago
big data cloud analytics Data-driven applications returned to the headlines this week after Hortonworks announced that it will bundle the open ...

Shop for **hadoop big data** on Google

Big Data Big Analytics: ...
Rs. 348.00
Amazon.in

Oracle **Big Data** ...
Rs. 549.00
Amazon.in

Big Data Analytics With
Rs. 455.00
Amazon.in

Hadoop Beginner's ...
Rs. 595.00
Amazon.in

Hadoop In Action
Rs. 460.00
Flipkart

Big Data Analytics with
Rs. 3,100.00
Amazon.in

Hadoop Mapreduce ...
Rs. 468.00
Amazon.in

Hadoop: The Definitive ...
Rs. 553.00
Amazon.in

Output returned by 'Google Search
]

SEMI-STRUCTURED DATA

- Self-describing structure.
- Is a form of structured data that contains tags or mark-up elements.
- Do not follow the proper structure of data models as in relational db’s.
  - Data is stored in rows and columns inconsistently.

Sources : Web data(cookies), JSON data etc.

EXAMPLE OF SEMI-STRUCTURED DATA

- Personal data stored in a XML file-

```
<rec><name>Prashant Rao</name><sex>Male</sex><age>35</age></rec>
<rec><name>Seema R.</name><sex>Female</sex><age>41</age></rec>
<rec><name>Satish Mane</name><sex>Male</sex><age>29</age></rec>
<rec><name>Subrato Roy</name><sex>Male</sex><age>26</age></rec>
<rec><name>Jeremiah J.</name><sex>Male</sex><age>35</age></rec>
```

CHARACERISTICS OF BIG DATA

#1: Volume

#2: Velocity

#3: Variety

#4:Veracity

#5: Variability

#6: Validity

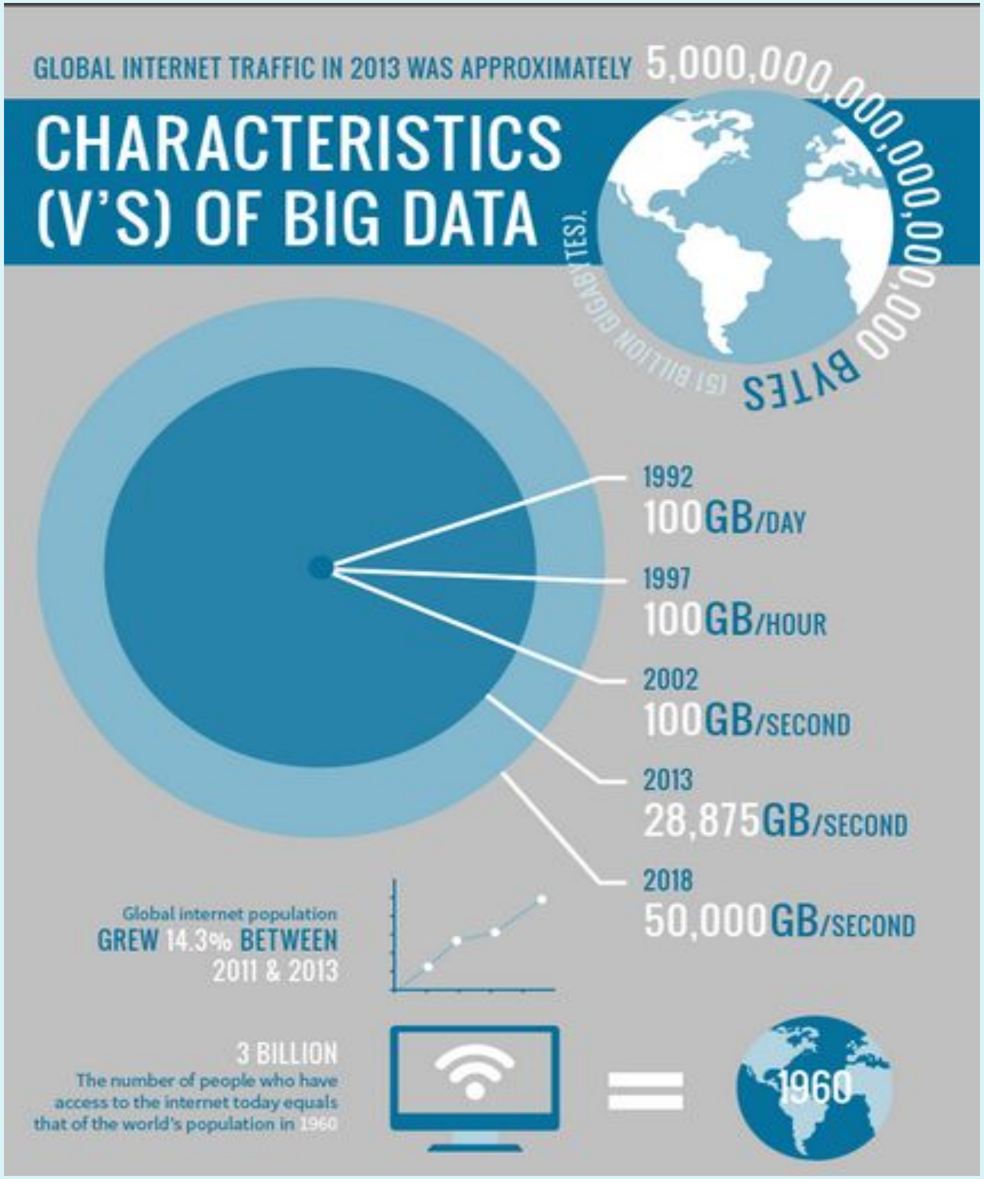
#7: Vulnerability

#8: Volatility

#9: Visualization

#10: Value

VOLUME



GLOBAL INTERNET TRAFFIC IN 2013 WAS APPROXIMATELY 5,000,000,000,000,000,000 BYTES (5 BILLION GIGABYTES).

CHARACTERISTICS (V'S) OF BIG DATA

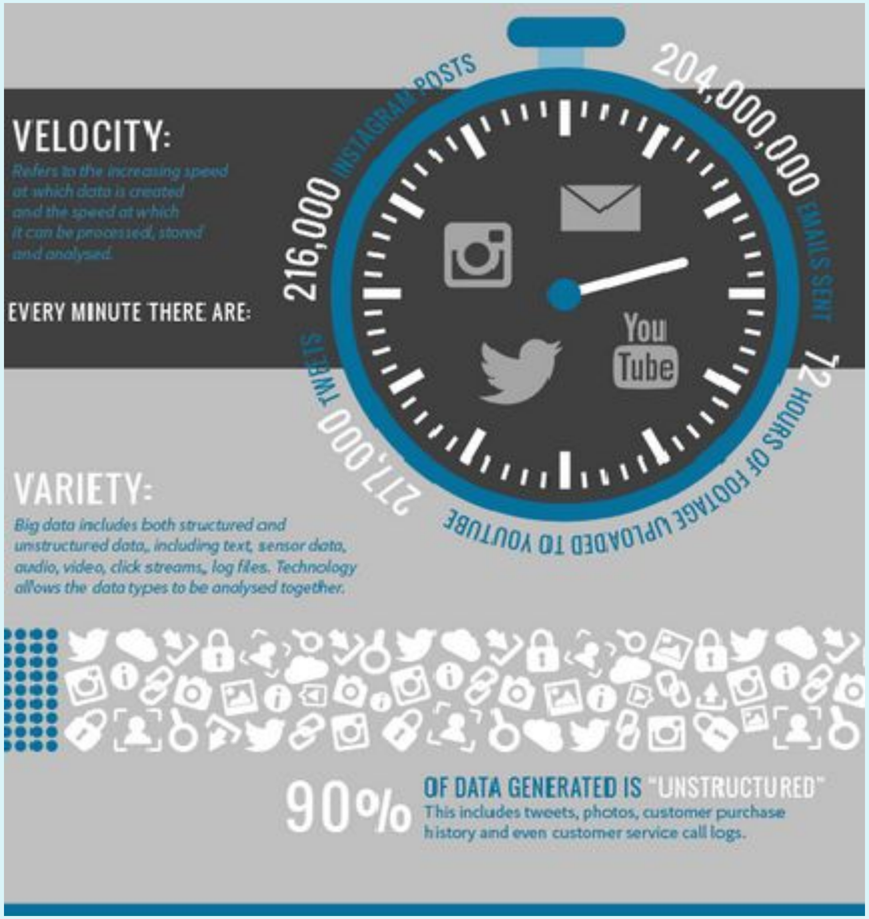
Global internet population GREW 14.3% BETWEEN 2011 & 2013

3 BILLION

The number of people who have access to the internet today equals that of the world's population in 1960

- In total, 2.7 Zettabytes of data exists in our digital universe. (A zettabyte is equal to 1,024 exabytes)
- 149, 513 emails are sent every minute.
- 3.3 million (undoubtedly extremely insightful) Facebook posts are created every minute.
- 3.8 Google million searches are performed each minute.
- Each minute, 65,972 Instagram photos are uploaded.
- 448,800 Tweets are constructed. How often? Every minute.
- 500 hours of YouTube videos are uploaded every minute.

VELOCITY



VELOCITY:

Refers to the increasing speed at which data is created and the speed at which it can be processed, stored and analysed.

EVERY MINUTE THERE ARE:

VARIETY:

Big data includes both structured and unstructured data,, including text, sensor data, audio, video, click streams,, log files. Technology allows the data types to be analysed together.

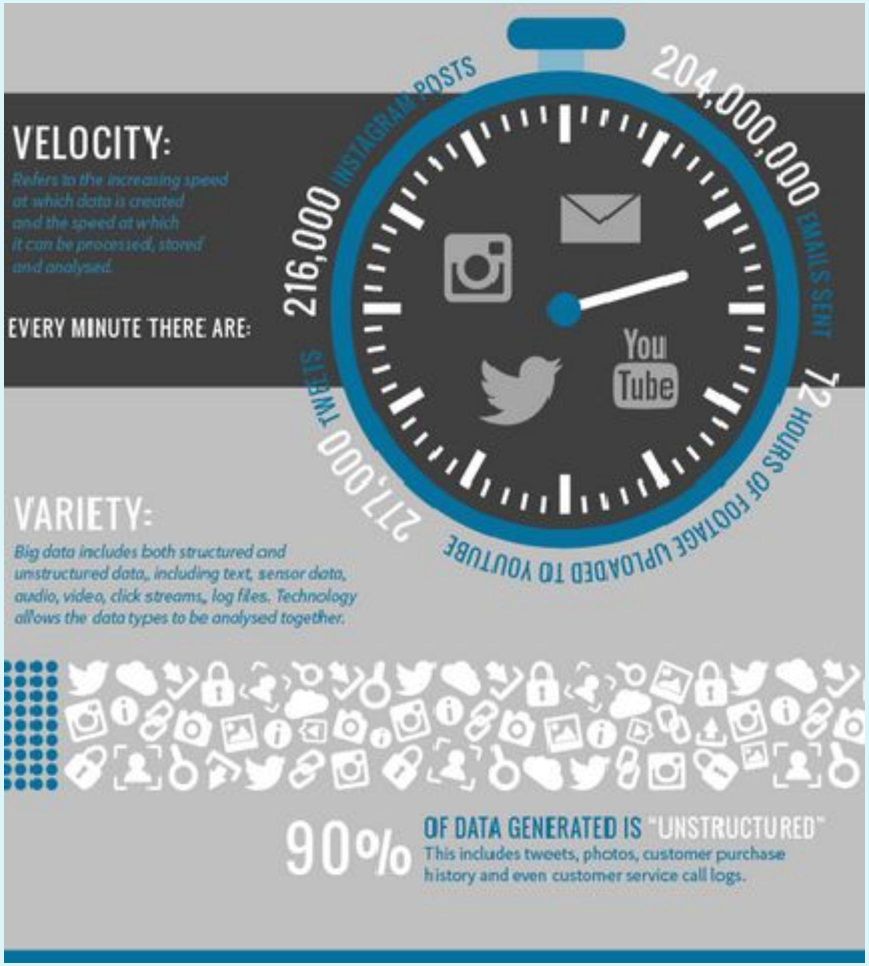
90% OF DATA GENERATED IS "UNSTRUCTURED"

This includes tweets, photos, customer purchase history and even customer service call logs.

Velocity refers to the speed at which new data is being generated, produced, created, or refreshed.

- Facebook claims 600 terabytes of incoming data per day.
- Google alone processes on average more than "40,000 search queries every second," which roughly translates to more than 3.5 billion searches per day

VARIETY.



VELOCITY:

Refers to the increasing speed at which data is created and the speed at which it can be processed, stored and analysed.

EVERY MINUTE THERE ARE:

VARIETY:

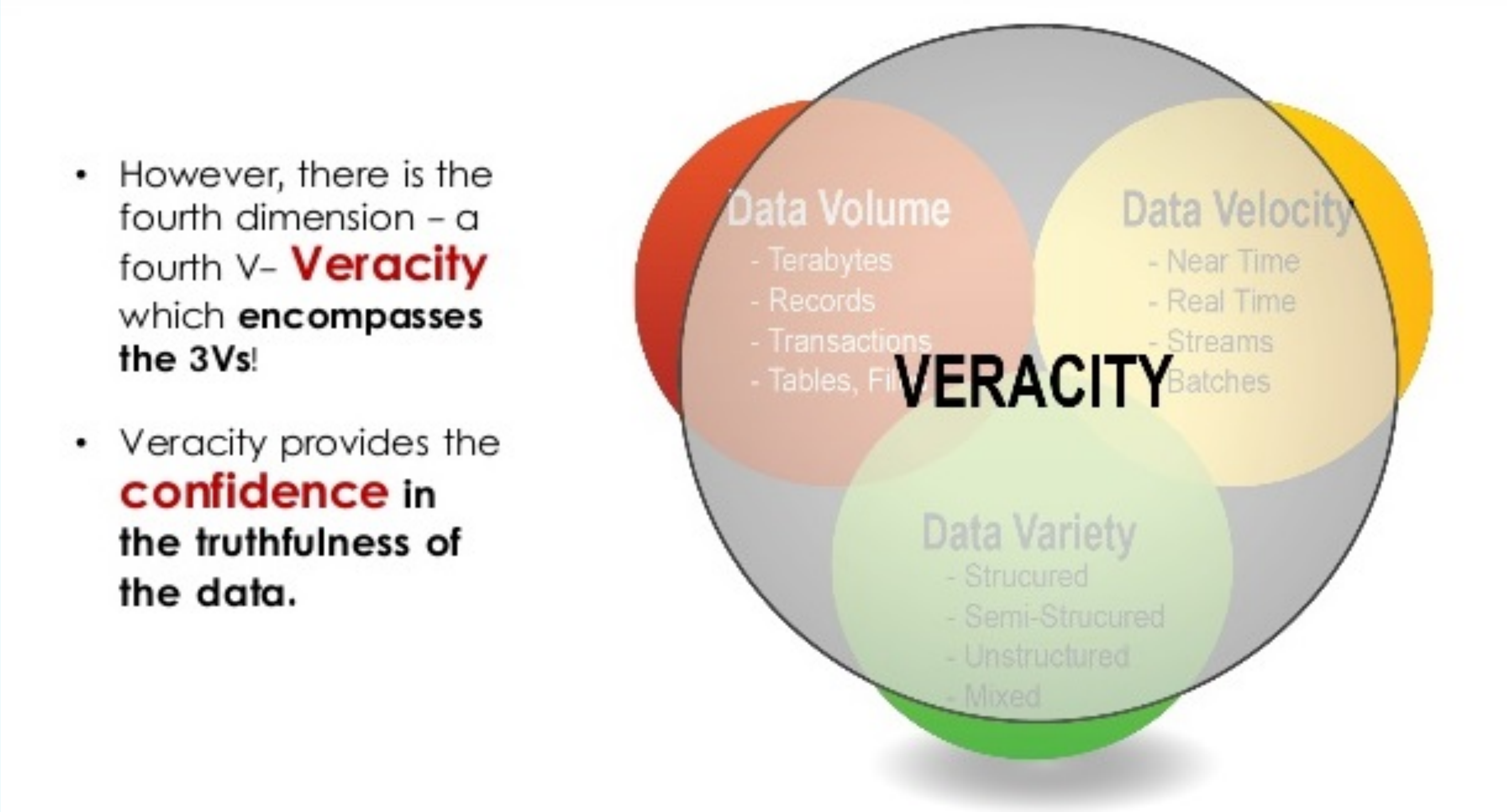
Big data includes both structured and unstructured data,, including text, sensor data, audio, video, click streams,, log files. Technology allows the data types to be analysed together.

90% OF DATA GENERATED IS "UNSTRUCTURED"

This includes tweets, photos, customer purchase history and even customer service call logs.

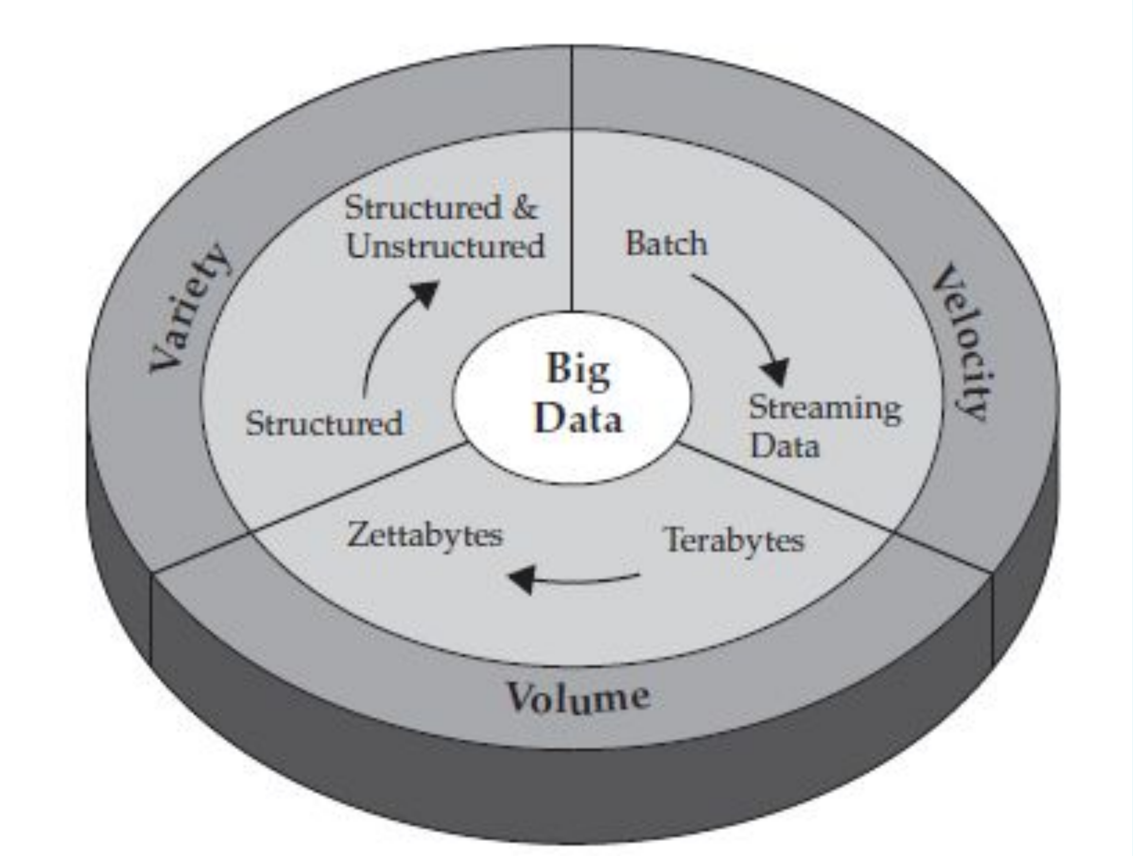
- Variety refers to heterogeneous sources and the nature of data- both structured and unstructured.
- Most big data seems to be unstructured, but besides audio, image, video files, social media updates, and other text formats there are also log files, click data, machine and sensor data, etc.

VERACITY



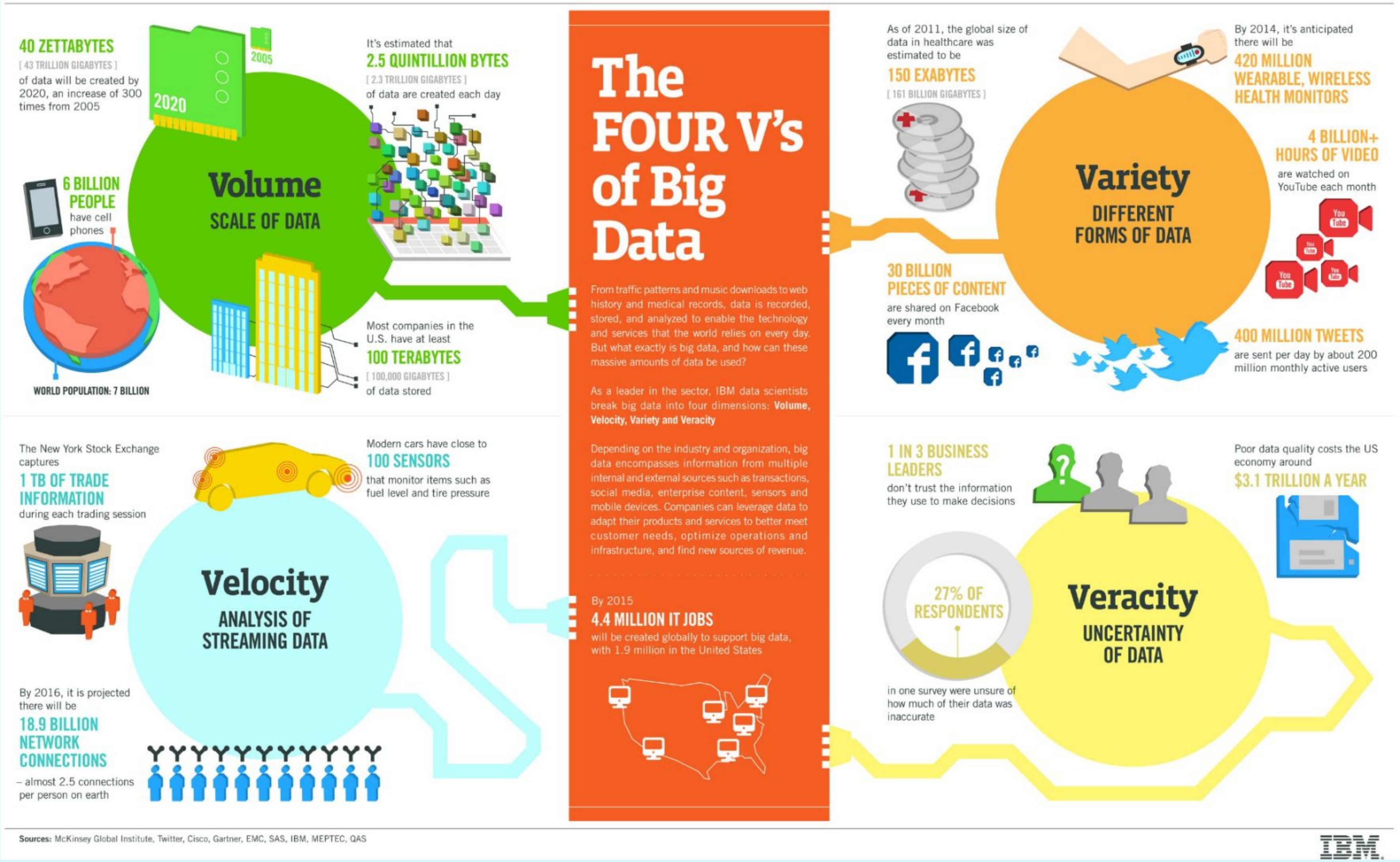
- However, there is the fourth dimension – a fourth V– **Veracity** which **encompasses the 3Vs**!
- Veracity provides the **confidence in the truthfulness of the data.**

CHARACERISTICS OF BIG DATA



REVIEW.....

- Companies have a wealth of big data under their own roofs.
  - Internal sources – internal data - reflect those data that are under the control of the business.
  - External sources - external data -any data generated outside the wall of the business.
- Structured data - organized data in a predefined format
- Semi-structured data – self describing nature
- Unstructured data - not been organized into a format –



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015

4.4 MILLION IT JOBS

will be created globally to support big data, with 1.9 million in the United States

Volume

SCALE OF DATA

40 ZETTABYTES

[43 TRILLION GIGABYTES]

of data will be created by 2020, an increase of 300 times from 2005

It's estimated that

2.5 QUINTILLION BYTES

[2.3 TRILLION GIGABYTES]

of data are created each day

6 BILLION PEOPLE have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least

100 TERABYTES

[100,000 GIGABYTES]

of data stored

Velocity

ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures

1 TB OF TRADE INFORMATION

during each trading session

Modern cars have close to

100 SENSORS

that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

– almost 2.5 connections per person on earth

Variety

DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES

[161 BILLION GIGABYTES]

By 2014, it's anticipated there will be

420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO

are watched on YouTube each month

30 BILLION PIECES OF CONTENT

are shared on Facebook every month

400 MILLION TWEETS

are sent per day by about 200 million monthly active users

Veracity

UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS

don't trust the information they use to make decisions

Poor data quality costs the US economy around

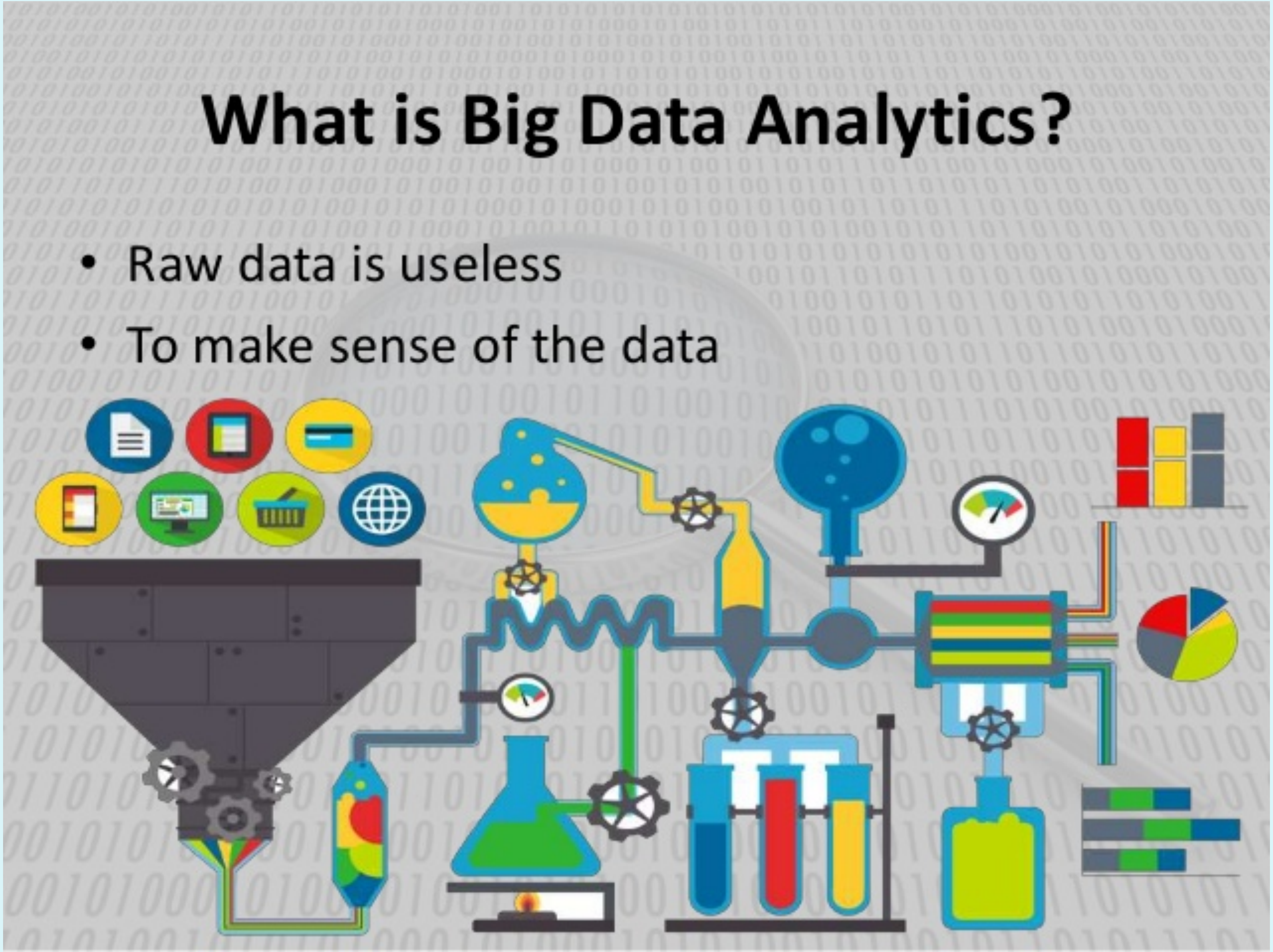
$3.1 TRILLION A YEAR

27% OF RESPONDENTS

in one survey were unsure of how much of their data was inaccurate

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM.



What is Big Data Analytics?

- Raw data is useless
- To make sense of the data

BIG DATA ANALYTICS (BDA)

- What ?

  “BDA is the process of examining large and varied data sets (big data) to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful information that can help organizations make more-informed business decisions.”

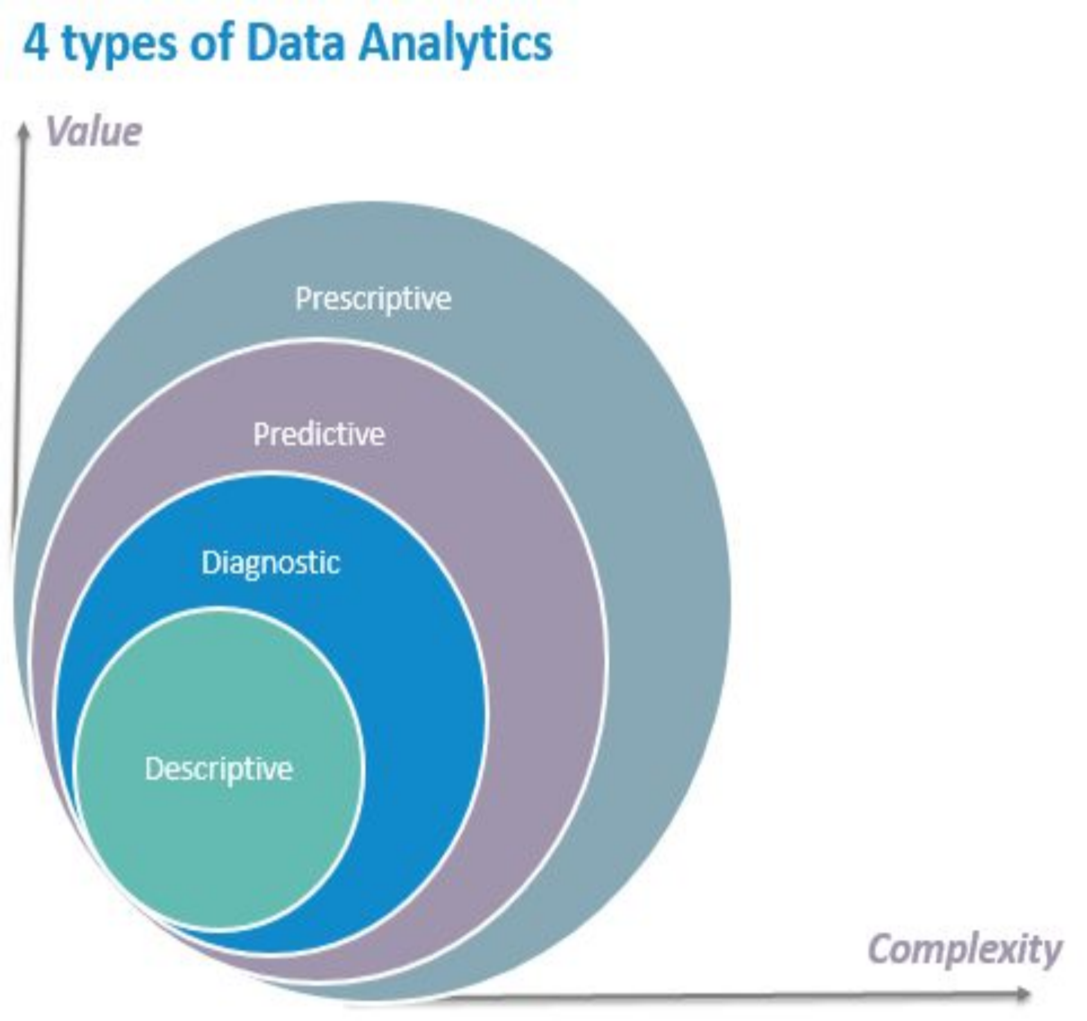
- Why ?

  Companies implement BDA because they want to make more informed business decisions.

IS DATA ANALYTICS ANYTHING NEW ?

- Even in the 1950’s - businesses were using basic analytics to uncover insights and trends.
  - essentially numbers in a spreadsheet that were manually examined
- Benefits that big data analytics brings to the table- speed and efficiency.
  - Few years ago -gathered information - run analytics -unearthed information - used for future decisions.
  - Today - business can identify insights for immediate decisions. The ability to work faster – and stay agile – gives organizations a competitive edge they didn’t have before.

TYPES OF BDA



Descriptive analytics: What happened?

Diagnostic analytics: Why did it happen?

Predictive analytics: What could happen in the future?

Prescriptive analytics: How should we respond to those potential future events?

DESCRIPTIVE ANALYTICS

- Serves as base for Advanced analytics
- It answers the question what happened in the business
- Analyses database to provide information on the trends of past or current business event that can help manager planners leaders to develop a road map for future action
- Performs in depth analysis of data to reveal details such as frequency of events, operation costs and underlying reason for failure
- Helps in identifying the root cause of the problem

DIAGNOSTIC ANALYTICS:

- takes descriptive data a step further and provides deeper analysis to answer the question: Why did this happen? Often, diagnostic analysis is referred to as root cause analysis.
- This includes using processes such as data discovery , data mining, and drill down and drill through.
- diagnostic analytics would explore the data and make correlations.

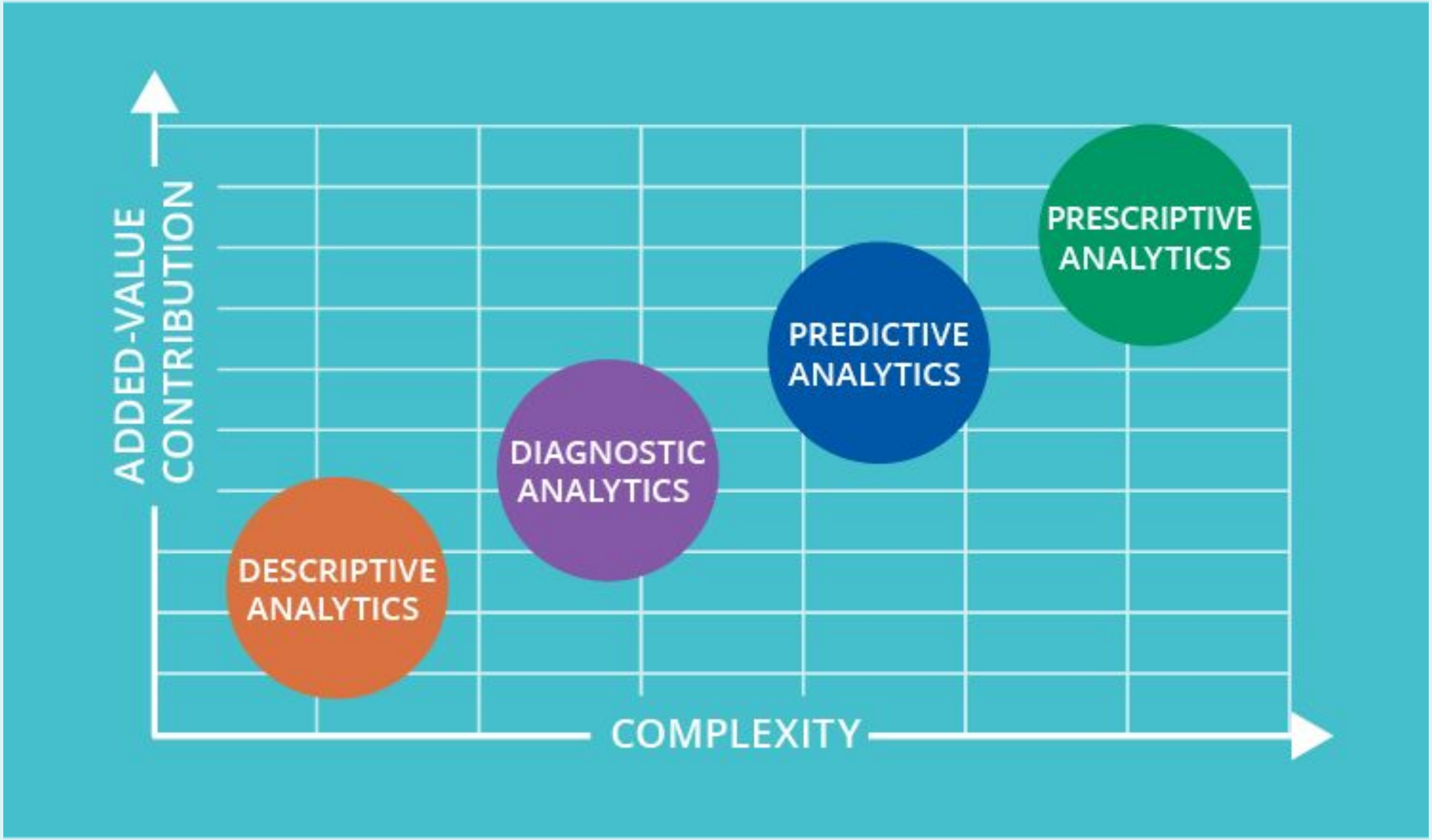
PREDICTIVE ANALYTICS

- Is about and understanding predicting the future and answers the question What could happen
- By using statically models and different forecast techniques
- It predicts the near future probabilities and trends and helps in analysis
- To analyze future predictive analytics use statistics,data mining technique and machine learning

PRESCRIPTIVE ANALYTICS

- takes predictive data to the next level.
- Now that you have an idea of what will likely happen in the future, what should you do?
- It suggests various courses of action and outlines what the potential implications would be for each.
- Answers what should we do on the basis of complex data obtained from descriptive and predictive analyses
- Determine the finest substitute to minimize or maximize some equitable finanace,marketing and many other areas
- For example if we have to find the best way of shipping goods from a factry to a destination to minimize costs we will use prescriptive analytics.

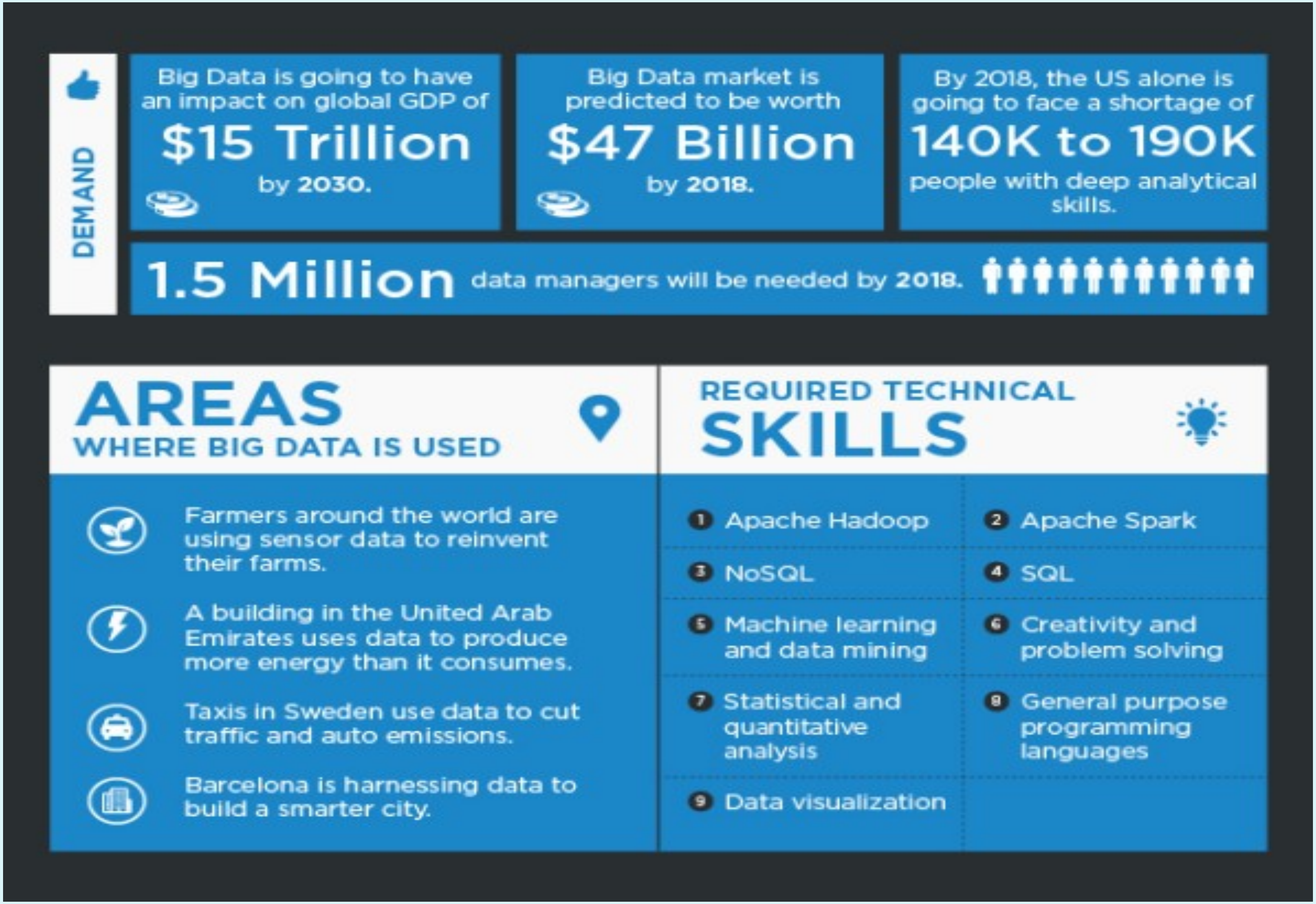
4 TYPES OF DATA ANALYTICS TO IMPROVE DECISION-MAKING



SUMMARY

- Both descriptive analytics and diagnostic analytics look to the past to explain what happened and why it happened.
- Predictive analytics and prescriptive analytics use historical data to forecast what will happen in the future and what actions you can take to affect those outcomes.
- Forward-thinking organizations use a variety of analytics together to make smart decisions that help your business

CAREERS IN BIG DATA



DEMAND

Big Data is going to have an impact on global GDP of **$15 Trillion** by 2030.

Big Data market is predicted to be worth **$47 Billion** by 2018.

By 2018, the US alone is going to face a shortage of **140K to 190K** people with deep analytical skills.

1.5 Million data managers will be needed by 2018.

AREAS

WHERE BIG DATA IS USED

Farmers around the world are using sensor data to reinvent their farms.

A building in the United Arab Emirates uses data to produce more energy than it consumes.

Taxis in Sweden use data to cut traffic and auto emissions.

Barcelona is harnessing data to build a smarter city.

REQUIRED TECHNICAL SKILLS

1. Apache Hadoop
2. Apache Spark
3. NoSQL
4. SQL
5. Machine learning and data mining
6. Creativity and problem solving
7. Statistical and quantitative analysis
8. General purpose programming languages
9. Data visualization

TOP BIG DATA PROFILES

- DATA SCIENTIST
- DATA ENGINEER
- BIG DATA ENGINEER
- BIG DATA DEVELOPER
- BIG DATA ADMINISTRATOR

SKILLS ESSENTIAL FOR A BIG DATA JOB

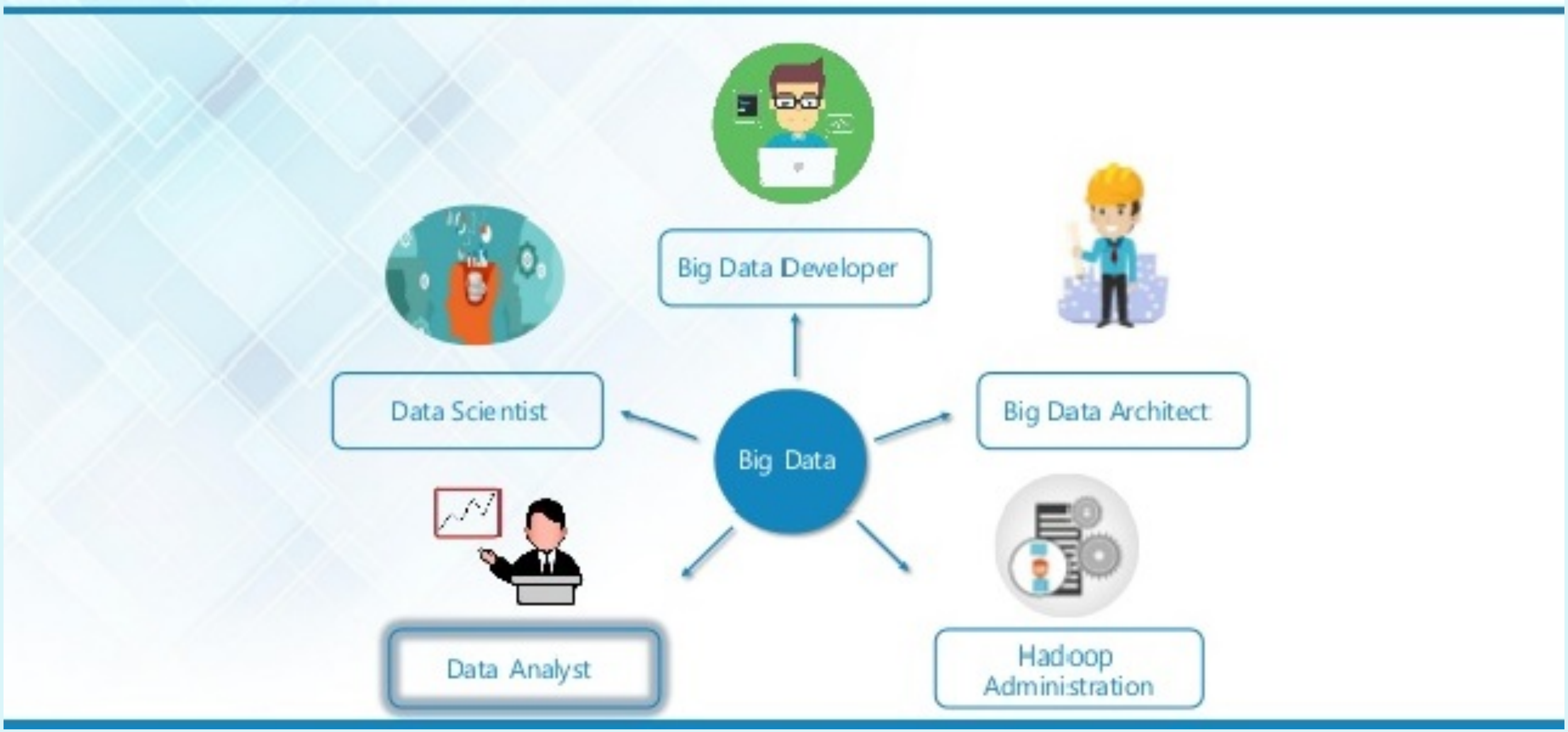
- Apache Hadoop.
- Apache Spark.
- NoSQL.
- Machine learning and Data Mining.
- Statistical and Quantitative Analysis.
- SQL.
- Data Visualization.
- General Purpose Programming language.

Big Data

=

Programming skills +Data Structure & Algorithms+ Analytical skills + Database Skills + Mathematics + Machine Learning + NLP +OS + Cryptography + Parallel Programming.

ROLE OF DIFFERENT JOB TITLES



1) BIG DATA ANALYST

A well trained professional who is able to :

- Collect data from different sources
- Organize it in a suitable format
- Analyze the data to generate desired results

2) BIG DATA SCIENTIST

are big data wranglers. They:

- take an enormous mass of messy data points (unstructured & structured)
- clean, massage and organize them (math, statistics and programming)
- to uncover hidden solutions to business

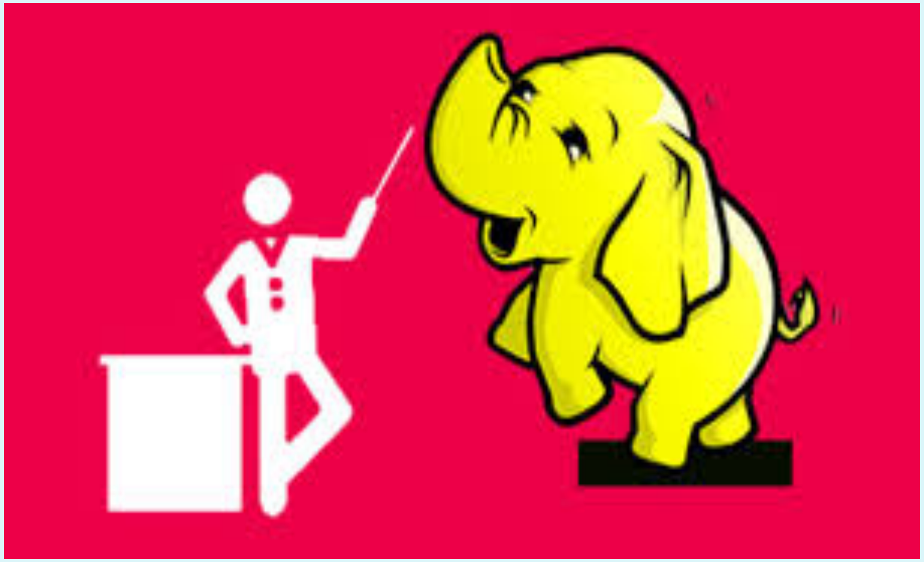
3) BIG DATA DEVELOPER

A programmer who can design, create, manage & administer :

- Large datasets
- Custom tools &
- Scripts

to achieve business goals

4) BIG DATA ADMINISTRATOR



An admin who is responsible for:

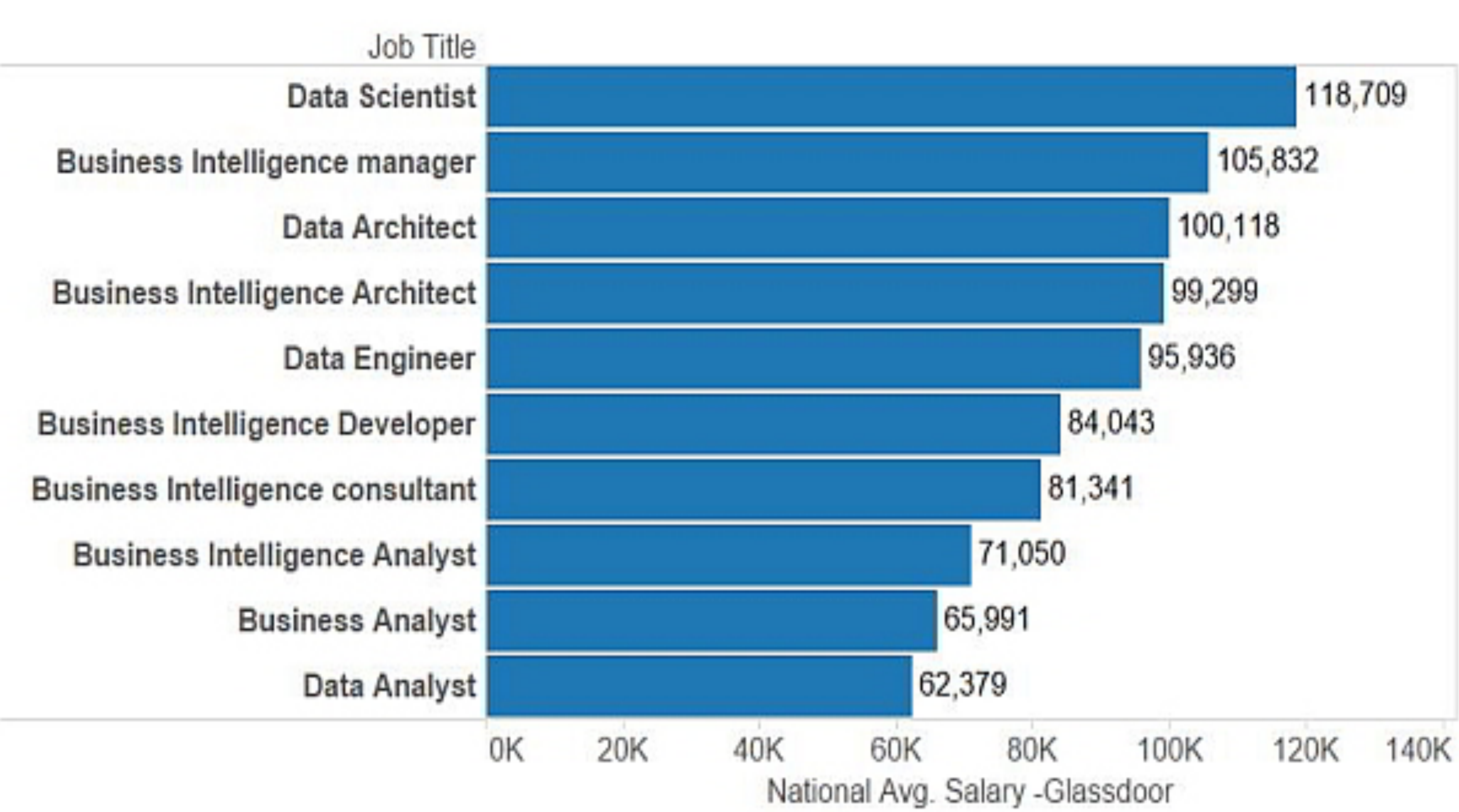
- System Upgrades
- Mgmt of the data warehouse
- Allocation of work load
- Storage

5) BIG DATA ENGINEER

A Professional who can:

- Design & develop applications using various frameworks and tools

BIG DATA....BIG BUCKS !!



FUTURE OF BIG DATA

- Allow the storage and use of transactional data in digital form
- Provide more specific information
- Refine analytics that can improve decision making
- Classify customers for providing customized products and services based on buying patterns

DATA SCIENCE VS. BIG DATA VS. DATA ANALYTICS