# Introduction to Big Data

# What’s Big Data?

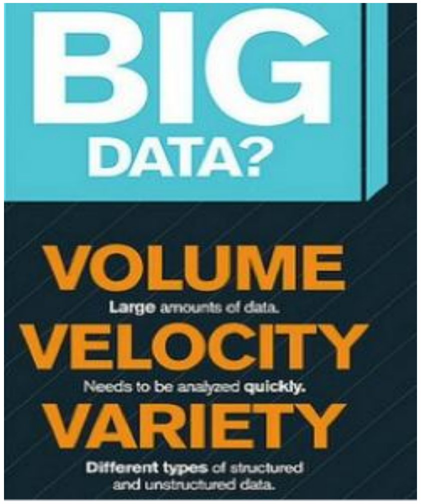
- **Big data** is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.
- The challenges include capture, curation, storage, search, sharing, transfer, analysis, and visualization.
- The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions.”

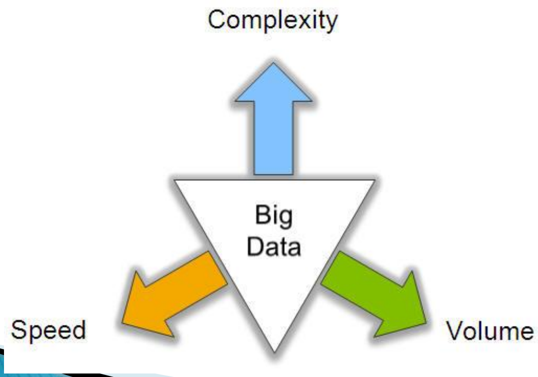
# Big Data Definition

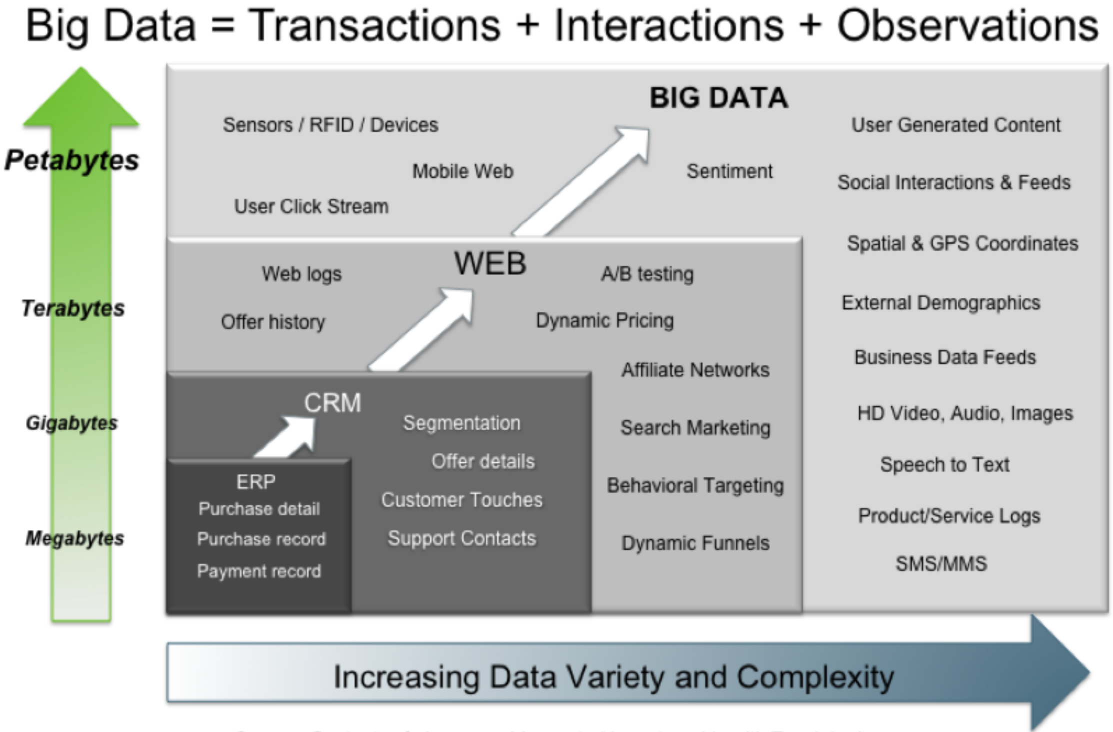
- No single standard definition…

"*Big Data*" is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it…

# Big Data: 5V's


BIG
DATA?
VOLUME
Large amounts of data.
VELOCITY
Needs to be analyzed quickly.
VARIETY
Different types of structured and unstructured data.

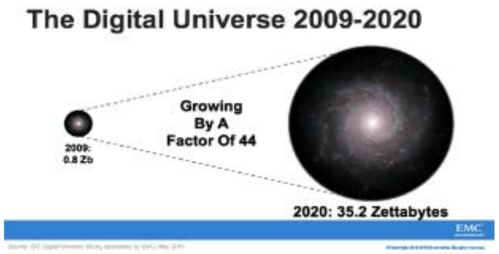


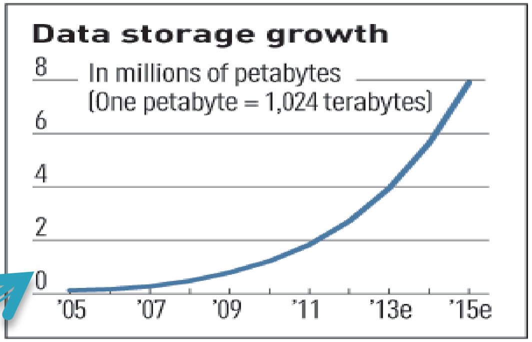


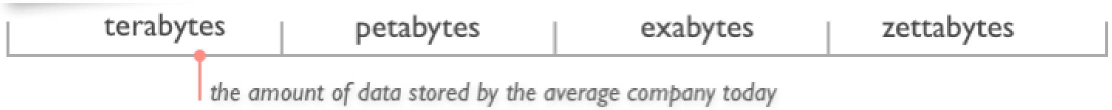
***Source**: Contents of above graphic created in partnership with Teradata, Inc.*

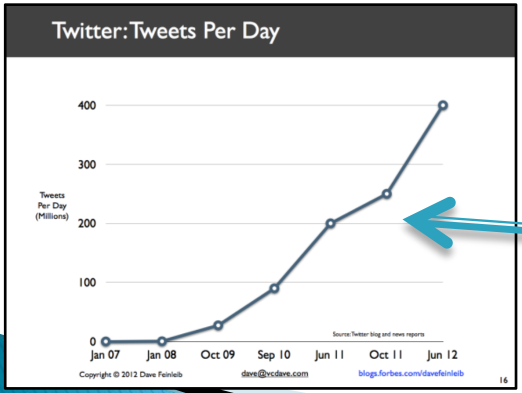
# Volume (Scale)

- **Data Volume**
  - 44x increase from 2009 2020
  - From 0.8 zettabytes to 35zb
- Data volume is increasing exponentially


The Digital Universe 2009-2020


Data storage growth




Twitter: Tweets Per Day

Source: Twitter blog and news reports

Copyright © 2012 Dave Feinleib dave@vcdave.com blogs.forbes.com/davefeinleib 16

*Exponential increase in collected/generated data*

*? TBs* of
data every day

***12+ TBs***
**of tweet data**
**every day**

***25+ TBs*** **of**
**log data**
**every day**

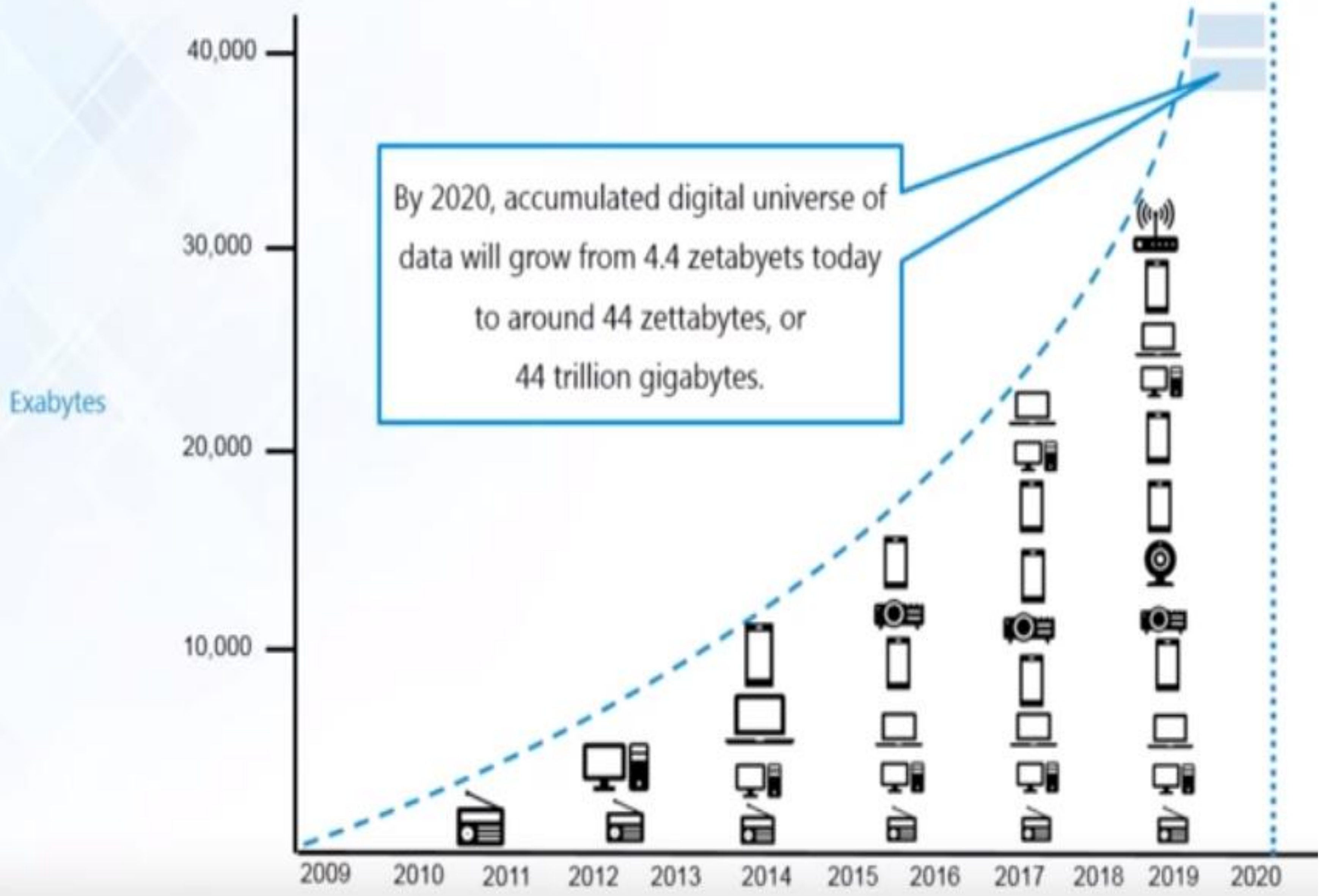
*30 billion* RFID
tags today
(1.3B in 2005)

*4.6*
*billion*
camera
phones
world
wide

*100s of*
*millions*
*of GPS*
*enabled*
devices
sold
annually

*76 million* smart
meters in 2009…
200M by 2014

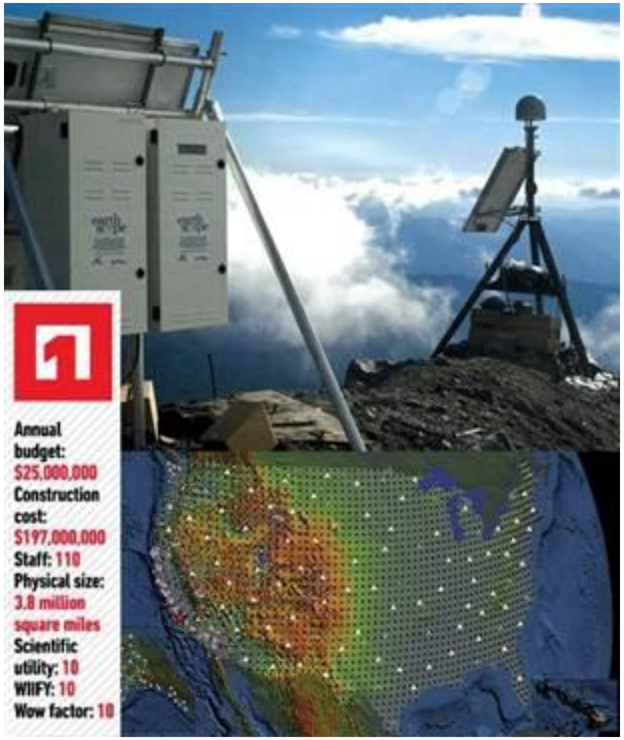
*2+*
*billion*
people
on the
Web by
end
2011



By 2020, accumulated digital universe of data will grow from 4.4 zetabyets today to around 44 zettabytes, or 44 trillion gigabytes.

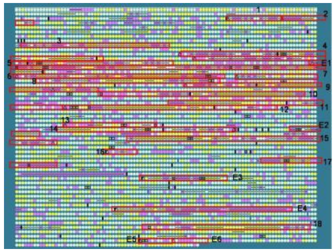
# The Earthscope

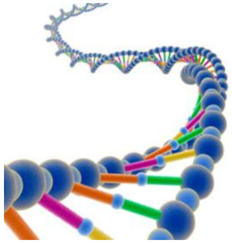
- The Earthscope is the world's largest science project. Designed to track North America's geological evolution, this observatory records data over 3.8 million square miles, amassing 67 terabytes of data. It analyzes seismic slips in the San Andreas fault, sure, but also the plume of magma underneath Yellowstone and much, much more. (http://www.msnbc.msn.com/id/44363598/ns/technology_and_science-future_of_technology/#.TmetOdQ--ul)

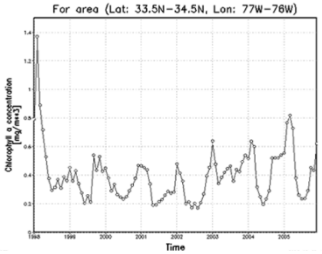

1
Annual budget: $25,000,000
Construction cost: $197,000,000
Staff: 110
Physical size: 3.8 million square miles
Scientific utility: 10
WIIFY: 10
Wow factor: 10

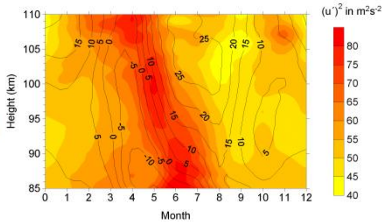
# Variety (Complexity)

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Web data (applied to data sourced from the World Wide Web and the Internet as a whole)
- Graph Data
  - Social Network, Semantic Web (RDF), …
- Streaming Data
  - You can only scan the data once
- A single application can be generating/collecting many types of data
- Big Public Data (online, weather, finance, etc)

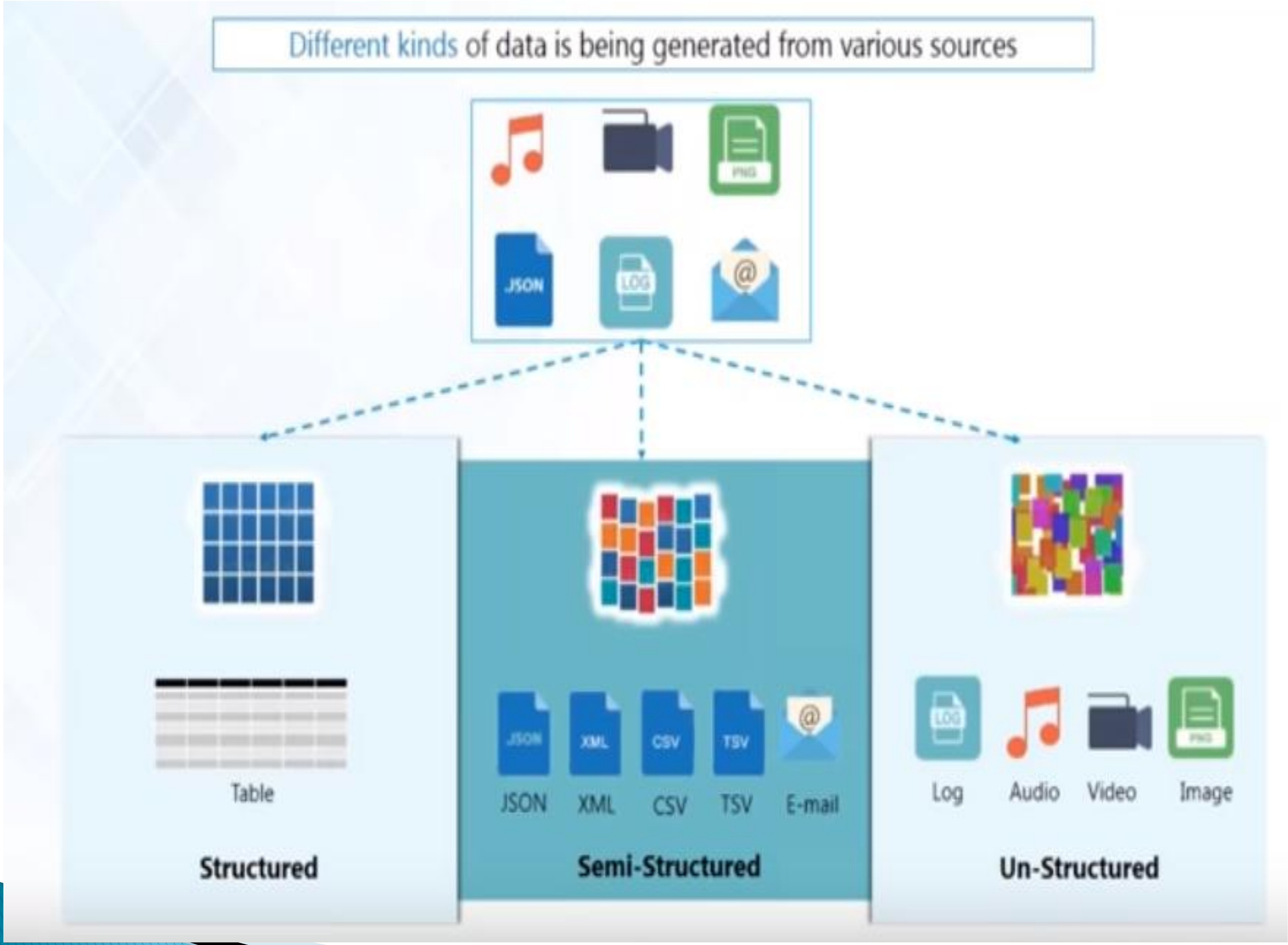





For area (Lat: 33.5N–34.5N, Lon: 77W–76W)
Chlorophyll a concentration [mg/m+3]
Time

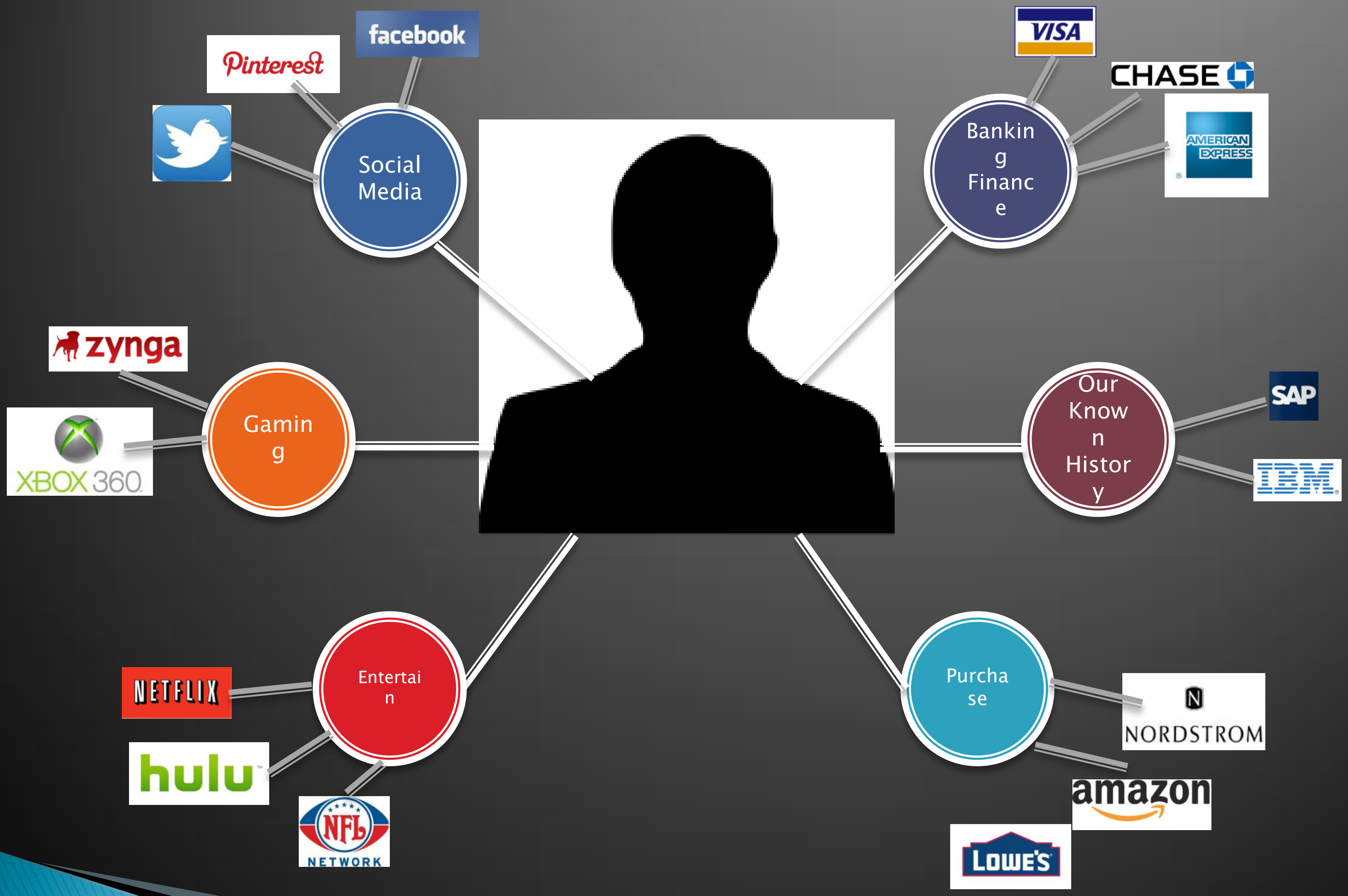


To extract knowledge→ all these types of data need to linked together



Different kinds of data is being generated from various sources

# A Single View to the Customer



Social Media

Banking Finance

Gaming

Our Known History

Entertain

Purchase

# Velocity (Speed)

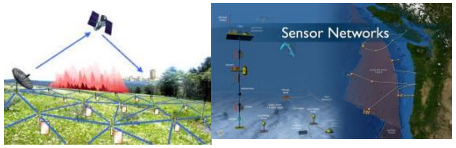
- Data is begin generated fast and need to be processed fast
- Online Data Analytics
- Late decisions ➔ missing opportunities
- Examples
  - **E-Promotions:** Based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you
  - **Healthcare monitoring:** sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction

# Real-time/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 tim

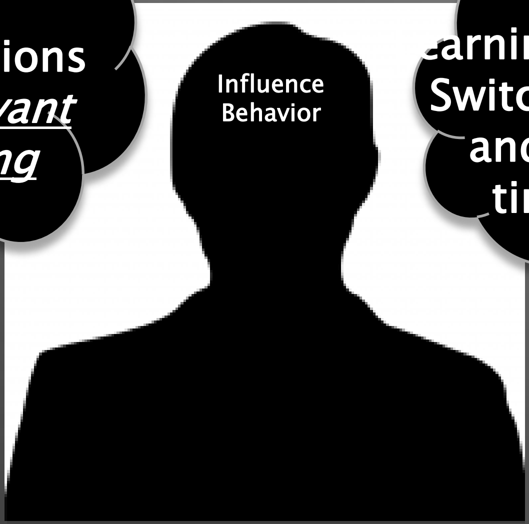


**Sensor technology and networks**
(measuring all kinds of data)

- The progress and innovation is 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

Product Recommendations that are *Relevant* & *Compelling*

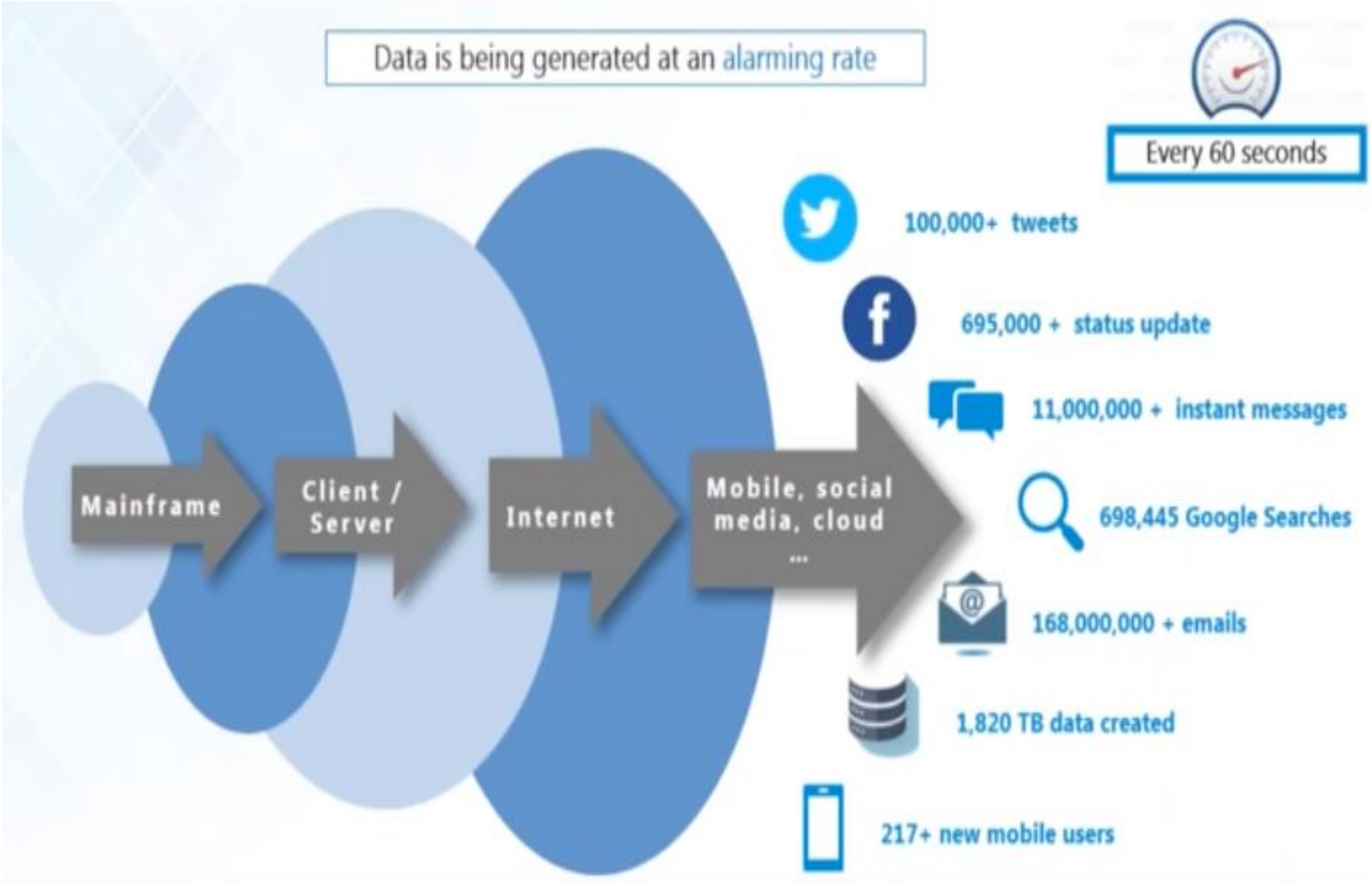


earning why Customers Switch to competitors and their offers; in time to Counter

Friend Invitations to join a Game or Activity that expands business

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

Preventing Fraud as it is *Occurring* & preventing more proactively

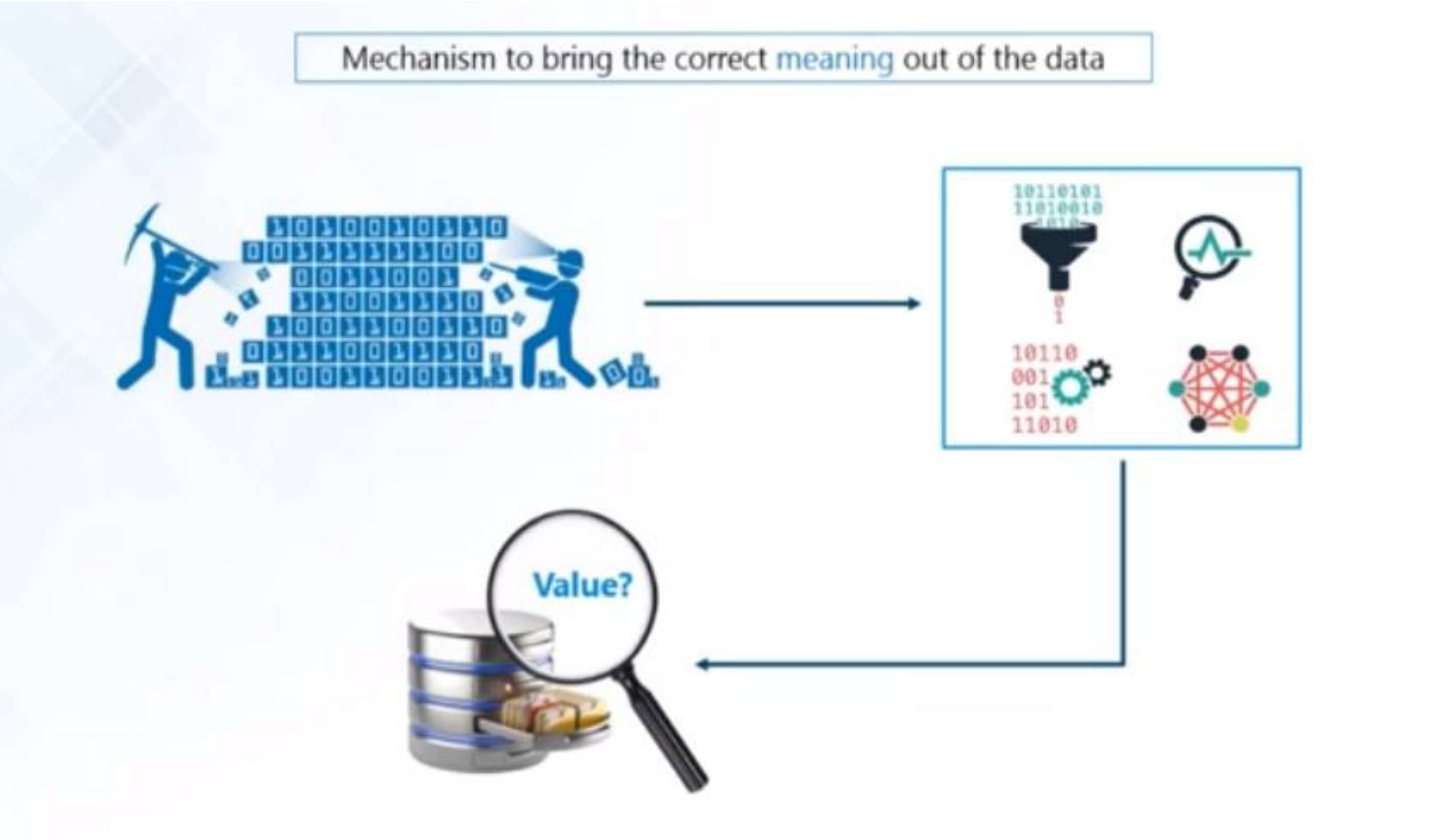


## Data is being generated at an alarming rate

Every 60 seconds

- 100,000+ tweets
- 695,000 + status update
- 11,000,000 + instant messages
- 698,445 Google Searches
- 168,000,000 + emails
- 1,820 TB data created
- 217+ new mobile users

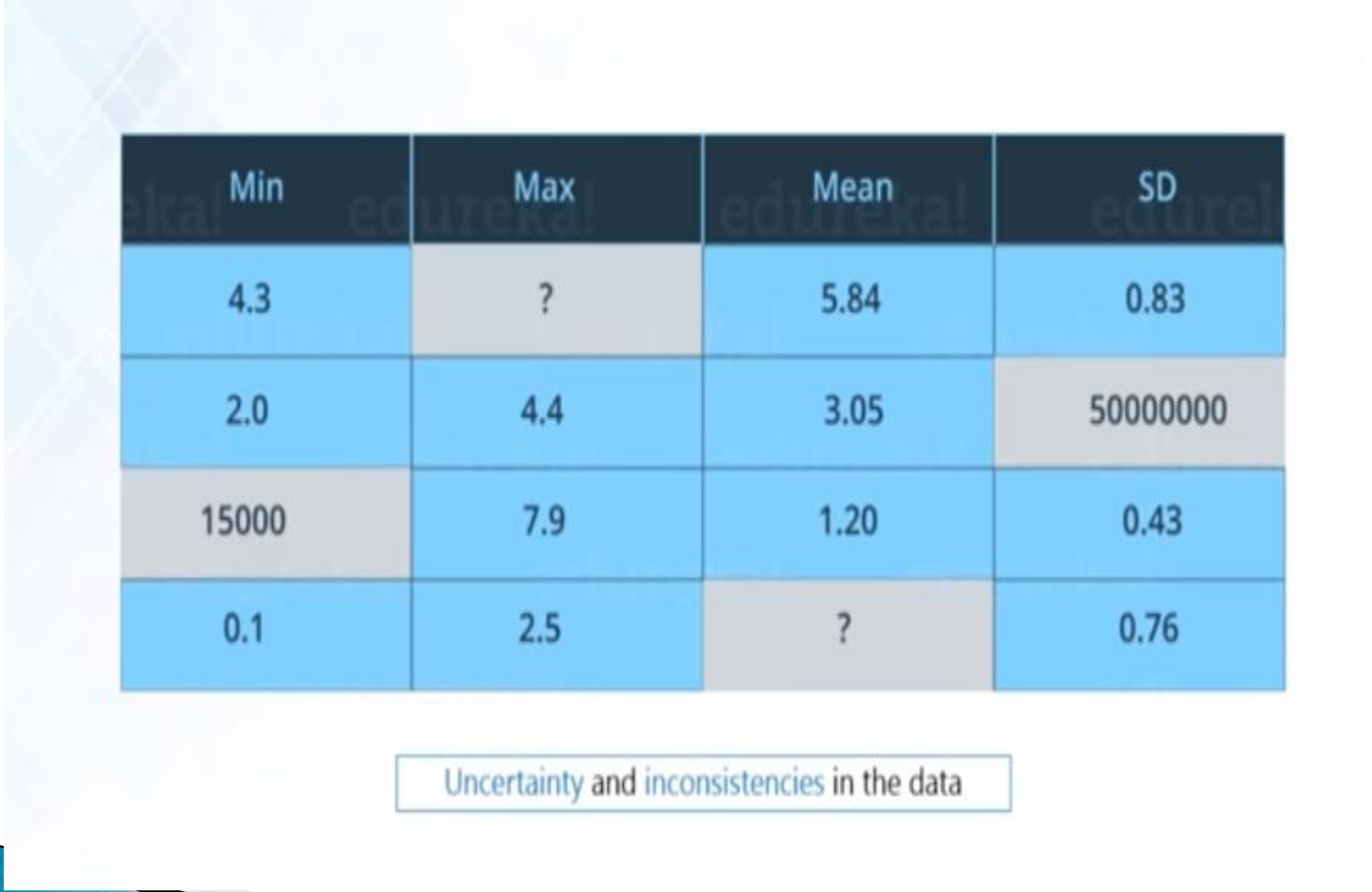
# Value



Mechanism to bring the correct meaning out of the data

- The bulk of Data having no Value is of no good to the company, unless you turn it into something useful.
- Data in itself is of no use or importance but it needs to be converted into something valuable to extract Information. Hence, you can state that Value! is the most important V of all the 5V’s.

# Verasity

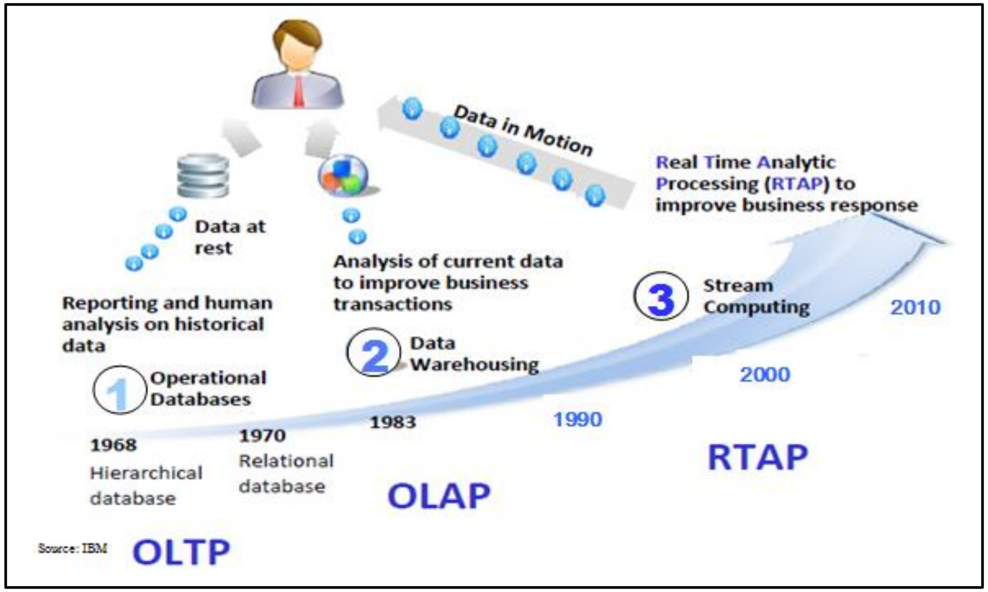


| Min | Max | Mean | SD |
|---|---|---|---|
| 4.3 | ? | 5.84 | 0.83 |
| 2.0 | 4.4 | 3.05 | 50000000 |
| 15000 | 7.9 | 1.20 | 0.43 |
| 0.1 | 2.5 | ? | 0.76 |

Uncertainty and inconsistencies in the data

- It refers to inconsistencies and uncertainty in data, that is data which is available can sometimes get messy and quality and accuracy are difficult to control.
- Big Data is also variable because of the multitude of data dimensions resulting from multiple disparate data types and sources.
- *Example:* Data in bulk could create confusion whereas less amount of data could convey half or Incomplete Information.

# Harnessing Big Data



Data in Motion
Real Time Analytic Processing (RTAP) to improve business response
Data at rest
Analysis of current data to improve business transactions
Reporting and human analysis on historical data
3 Stream Computing
2010
2 Data Warehousing
2000
1 Operational Databases
1983
1990
1968 Hierarchical database
1970 Relational database
RTAP
OLAP
Source: IBM
OLTP

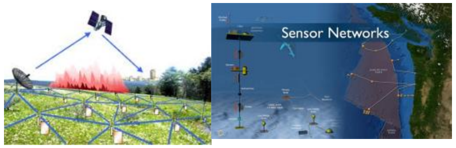
- OLTP: Online Transaction Processing (DBMSs)
- OLAP: Online Analytical Processing (Data Warehousing)
- RTAP: Real-Time Analytics Processing (Big Data Architecture & technology)

# Who’s Generating Big 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 tim



**Sensor technology and networks**
(measuring all kinds of data)

- The progress and innovation is 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

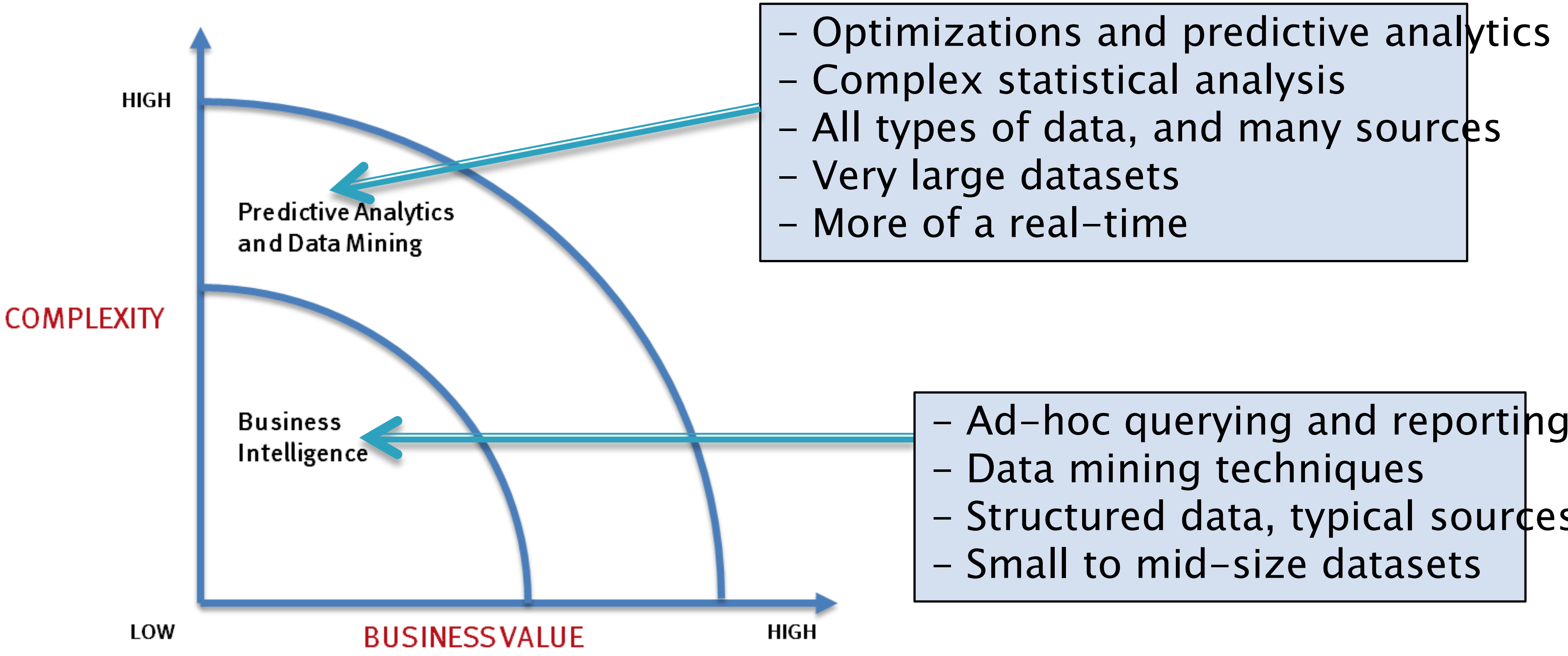
# The Model Has Changed…

- **The Model of Generating/Consuming Data has Changed**

**Old Model:** Few companies are generating data, all others are consuming data

**New Model:** all of us are generating data, and all of us are consuming data

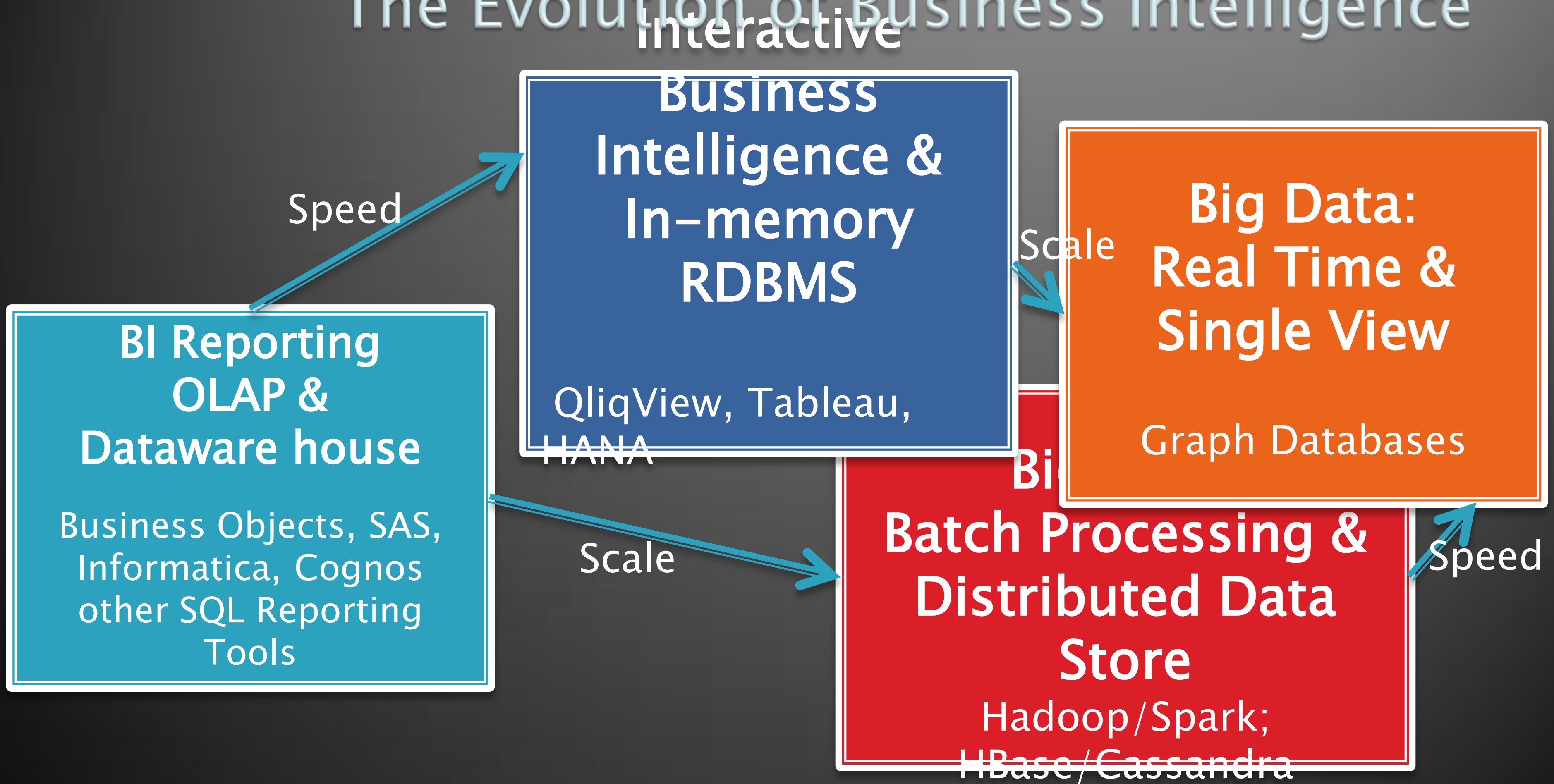
# What’s driving Big Data



- Optimizations and predictive analytics
- Complex statistical analysis
- All types of data, and many sources
- Very large datasets
- More of a real-time

- Ad-hoc querying and reporting
- Data mining techniques
- Structured data, typical sources
- Small to mid-size datasets

# The Evolution of Business Intelligence



BI Reporting
OLAP &
Dataware house

Business Objects, SAS, Informatica, Cognos other SQL Reporting Tools

Speed

Scale

Interactive Business Intelligence & In-memory RDBMS

QliqView, Tableau, HANA

Scale

Big Data:
Real Time &
Single View

Graph Databases

Bi
Batch Processing & Distributed Data Store

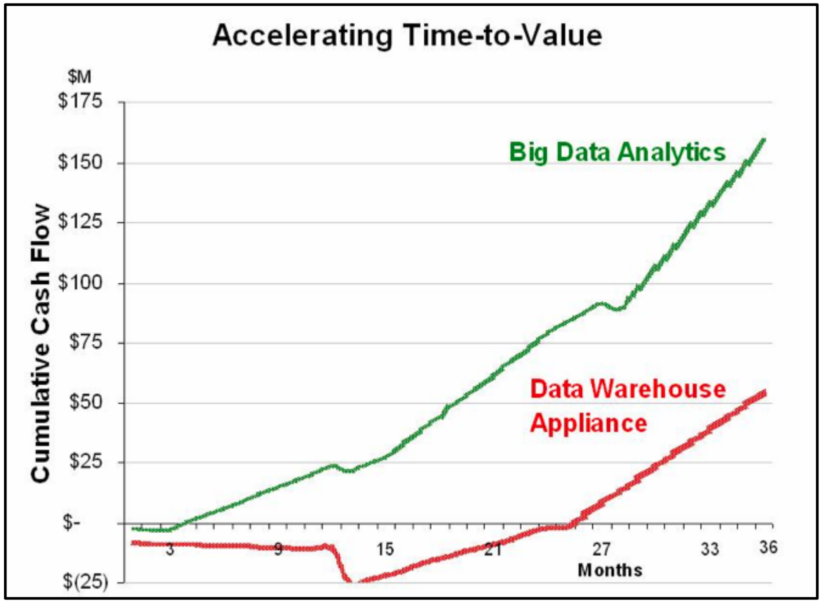
Hadoop/Spark; HBase/Cassandra

Speed

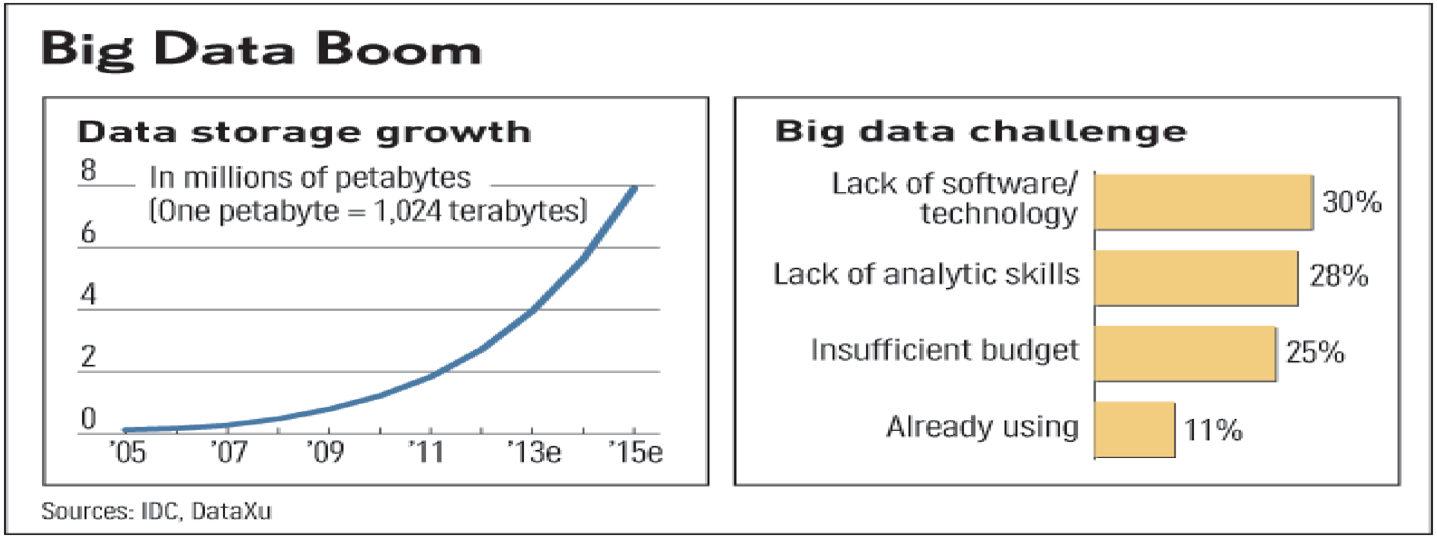
1990's 2000's 2010's

# Value of Big Data Analytics

- Big data is more real-time in nature than traditional DW applications
- Traditional DW architectures (e.g. Exadata, Teradata) are not well-suited for big data apps
- Shared nothing, massively parallel processing, scale out architectures are well-suited for big data apps



# Challenges in Handling Big Data



- **The Bottleneck is in technology**
  - New architecture, algorithms, techniques are needed
- **Also in technical skills**
  - Experts in using the new technology and dealing with big data

# The Big Data Landscape

## *Apps*

### Vertical Apps

- Atigeo
- ellucian.
- MYRRIX
- Placed.
- PREDICTIVE POLICING
- Quantivo

### Operational Intelligence

- VITRIA
- loggly
- splunk>
- sumologic

### Ad / Media Apps

- IPONWEB JAPAN
- bloomreach
- bluefin
- collective[i]
- DataXu Data. Insight. Action.
- LuckySort
- Media Science
- Recorded Future
- rocketfuel
- TURN

### Business Intelligence

- ATTIVIO
- Autonomy
- bime
- birst
- Business Objects
- Chart.io
- COGNOS
- DOMO
- GoodData
- IBM
- JASPERSOFT
- MicroStrategy
- pentaho
- SiSense

### Analytics And Visualization

- 1010data
- alteryx
- AYATA
- centrifuge
- CIRRO
- ClearStory
- Datameer
- emcien
- KARMASPHERE
- metaLayer
- OPERA
- Palantir
- panopticon
- platfora
- QlikView
- RJMetrics
- saffron
- sas
- tableau
- TIBCO
- visual.ly

### Data As A Service

- DATASIFT
- GNIP
- factual.
- FICO
- GNIP
- INRIX
- kaggle
- knoema beta
- LexisNexis
- LOQATE
- SPACE CURVE

## *Infrastructure*

### Analytics Infrastructure

- calpont
- cloudera
- DATASTAX
- EXASOL
- GREENPLUM
- HADAPT
- Hortonworks
- INFOBRIGHT
- kognitio
- MAPR TECHNOLOGIES
- PARACCEL
- VERTICA

### Operational Infrastructure

- 10gen
- Couchbase
- MarkLogic
- TERRACOTTA
- VoltDB

### Infrastructure As A Service

- CONTINUUITY
- infochimps
- MORTAR
- Qubole

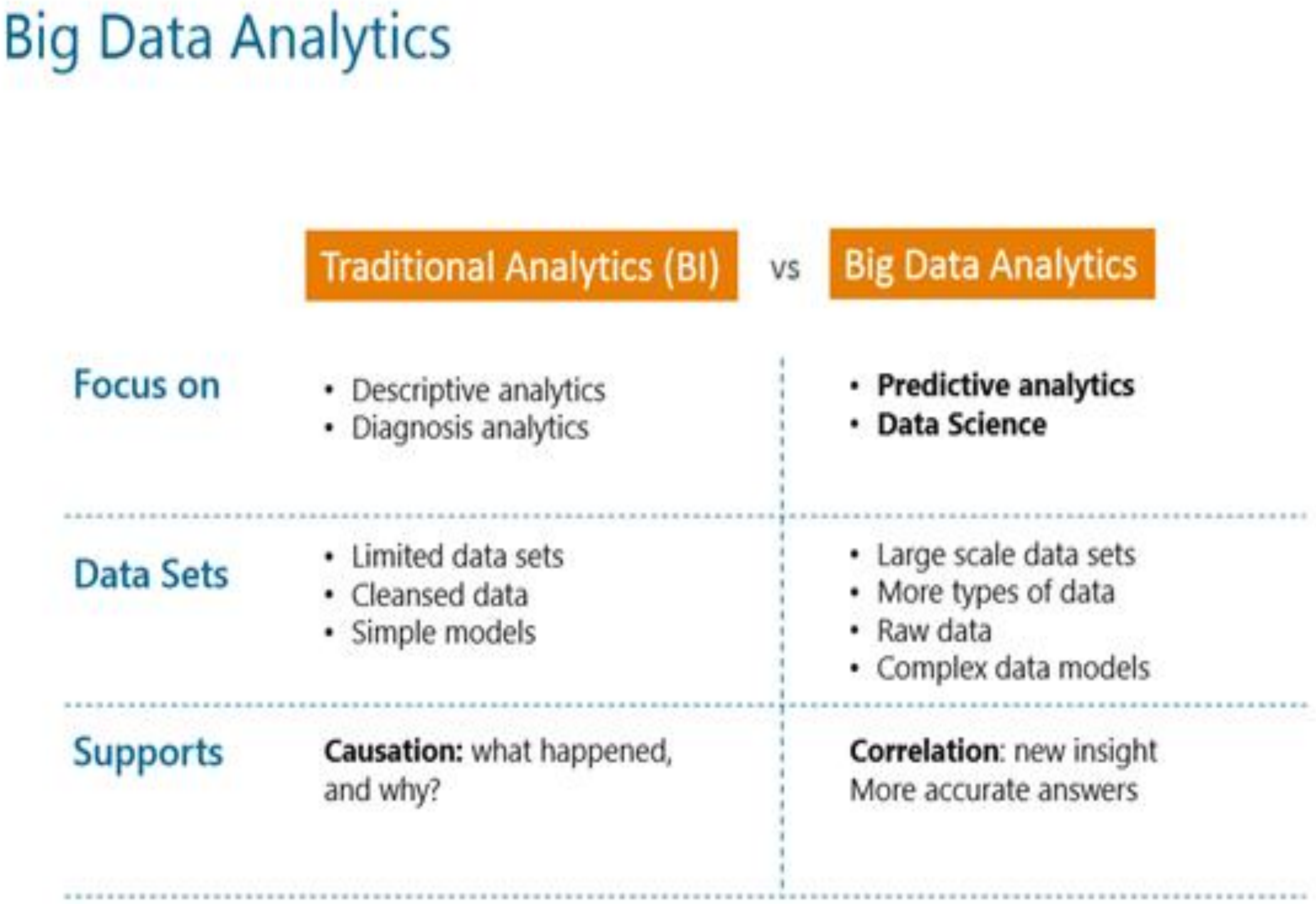
### Structured Databases

- IBM DB2.
- SQL Server
- MySQL
- ORACLE
- PostgreSQL
- SYBASE

### Technologies

- APACHE HBASE
- Cassandra
- hadoop

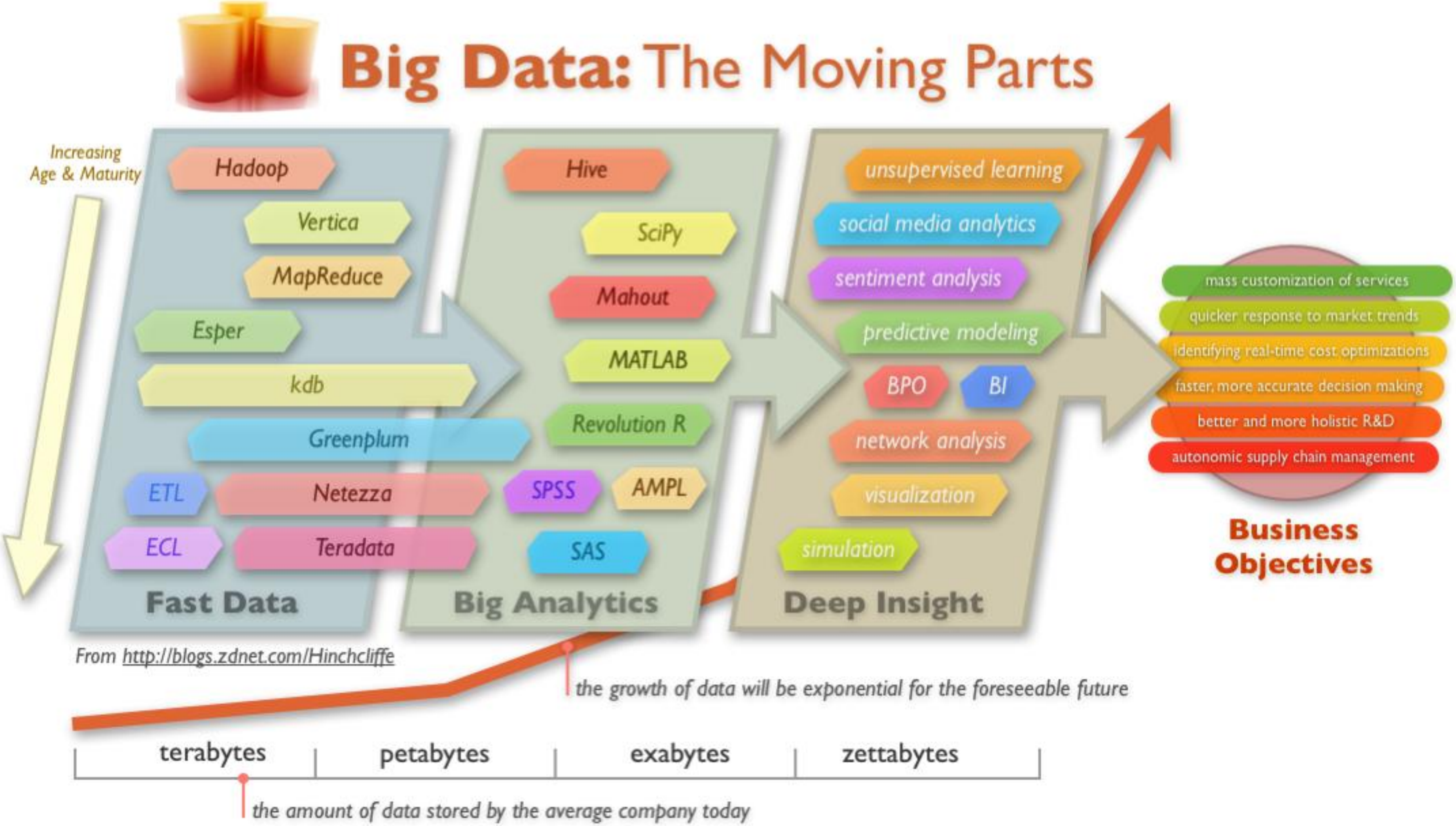
# Big Data Analytics



## Big Data Analytics

|  | Traditional Analytics (BI) | vs Big Data Analytics |
|---|---|---|
| Focus on | • Descriptive analytics<br>• Diagnosis analytics | • **Predictive analytics**<br>• **Data Science** |
| Data Sets | • Limited data sets<br>• Cleansed data<br>• Simple models | • Large scale data sets<br>• More types of data<br>• Raw data<br>• Complex data models |
| Supports | **Causation:** what happened, and why? | **Correlation**: new insight<br>More accurate answers |

# Big Data Technology


Big Data: The Moving Parts
Increasing Age & Maturity
Hadoop
Vertica
MapReduce
Esper
kdb
Greenplum
ETL
Netezza
ECL
Teradata
Fast Data
Hive
SciPy
Mahout
MATLAB
Revolution R
SPSS
AMPL
SAS
Big Analytics
unsupervised learning
social media analytics
sentiment analysis
predictive modeling
BPO
BI
network analysis
visualization
simulation
Deep Insight
mass customization of services
quicker response to market trends
identifying real-time cost optimizations
faster, more accurate decision making
better and more holistic R&D
autonomic supply chain management
Business Objectives
the growth of data will be exponential for the foreseeable future
terabytes
petabytes
exabytes
zettabytes
the amount of data stored by the average company today

From http://blogs.zdnet.com/Hinchcliffe

# Challenges

- capture,
- cleaning,
- curation,
- integration,
- storage,
- processing,
- indexing,
- search,
- sharing,
- transfer,
- mining,
- analysis,
- visualization