INTRODUCTION TO DATA ANALYTICS

Class #1

Introduction to Data

Dr. Sreeja S R

Assistant Professor

Indian Institute of Information Technology

IIIT Sri City

QUOTE OF THE DAY..

You can have data without information,
but you cannot have information without data.

IN TODAY'S DISCUSSION…

- Introduction to data
- Current trend
- Data and Big data
- Big data vs. small data
- Tools and techniques

INTRODUCTION TO DATA

- Example:

  10, 25, …, Sri City, Sreeja, 10CS3002, india@gov.in

  Anything else?

- Data vs. Information

  100.0, 0.0, 250.0, 150.0, 220.0, 300.0, 110.0

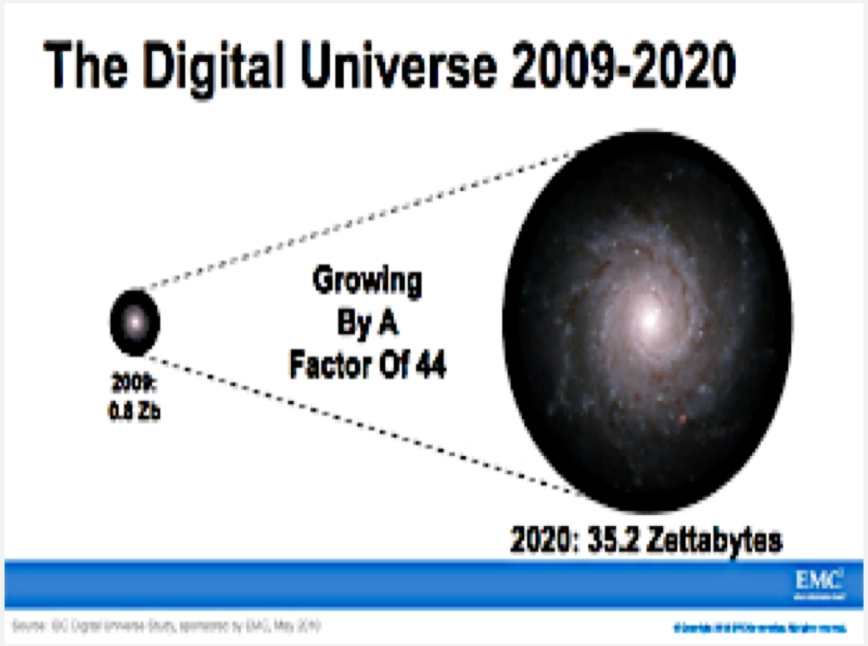
  Is there any information?

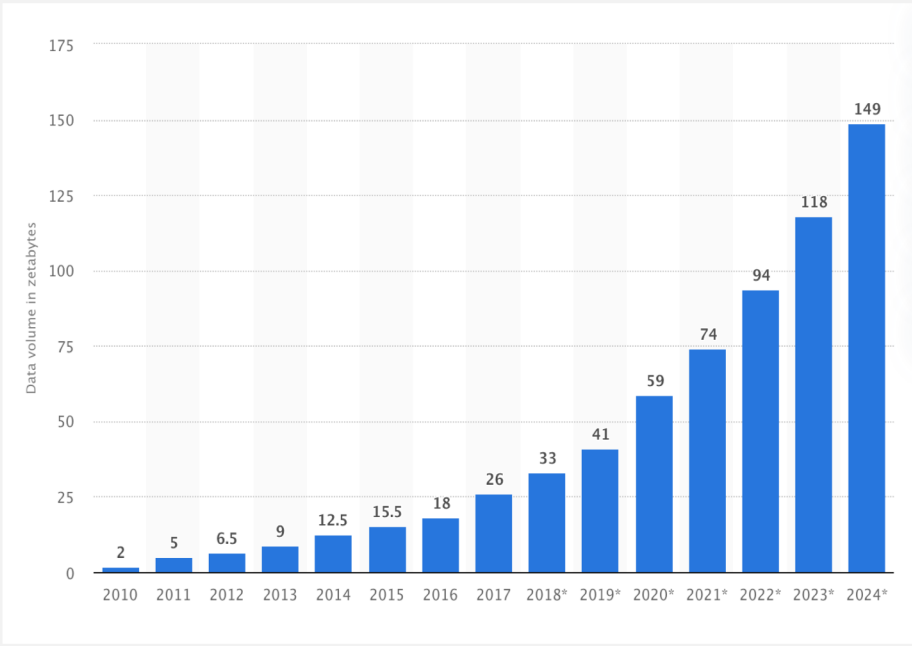
HOW LARGE YOUR DATA IS?

- What is the maximum file size you have dealt so far?
  - Movies/files/streaming video that you have used?
- What is the maximum download speed you get?
  - To retrieve data stored in distant locations?
- How fast your computation is?
  - How much time to just transfer from you, process and get result?

Memory unit	Equal to	Binary size
Bit	1 bit	2^1
Nibble	4 bits	2^4
Byte	8 bits	2^8
Kilobyte (KB)	1024 Bytes	2^{10}
Megabyte (MB)	1024 KB	2^{20}
Gigabyte (GB)	1024 MB	2^{30}
Terabyte (TB)	1024 GB	2^{40}
Petabyte (PB)	1024 TB	2^{50}
Exabyte (EB)	1024 PB	2^{60}
Zettabyte (ZB)	1024 EB	2^{70}
Yottabyte (YB)	1024 ZB	2^{80}

GROWTH OF DATA





SOURCES OF DATA

- "Every day, we create 2.5 quintillion bytes of data
  - So much that 90% of the data in the world today has been created in the last two years alone.

  - The data come from several sources
    - sensors used to gather climate information
    - posts to social media sites,
    - digital pictures and videos
    - purchase transaction records
    - cell phone GPS signals

etc. …… to name a few!

EXAMPLES

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)

NOW DATA IS BIG DATA!

- No single standard definition!

- 'Big-data' is similar to 'Small-data', but bigger

  …but having data bigger consequently requires different approaches

  - techniques, tools and architectures

  …to solve: new problems

  …and, of course, in a better way

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…

NOW DATA IS BIG DATA!

- No single standard definition!

- 'Big-data' is similar to 'Small-data', but bigger

  …but having data bigger consequently requires different approaches

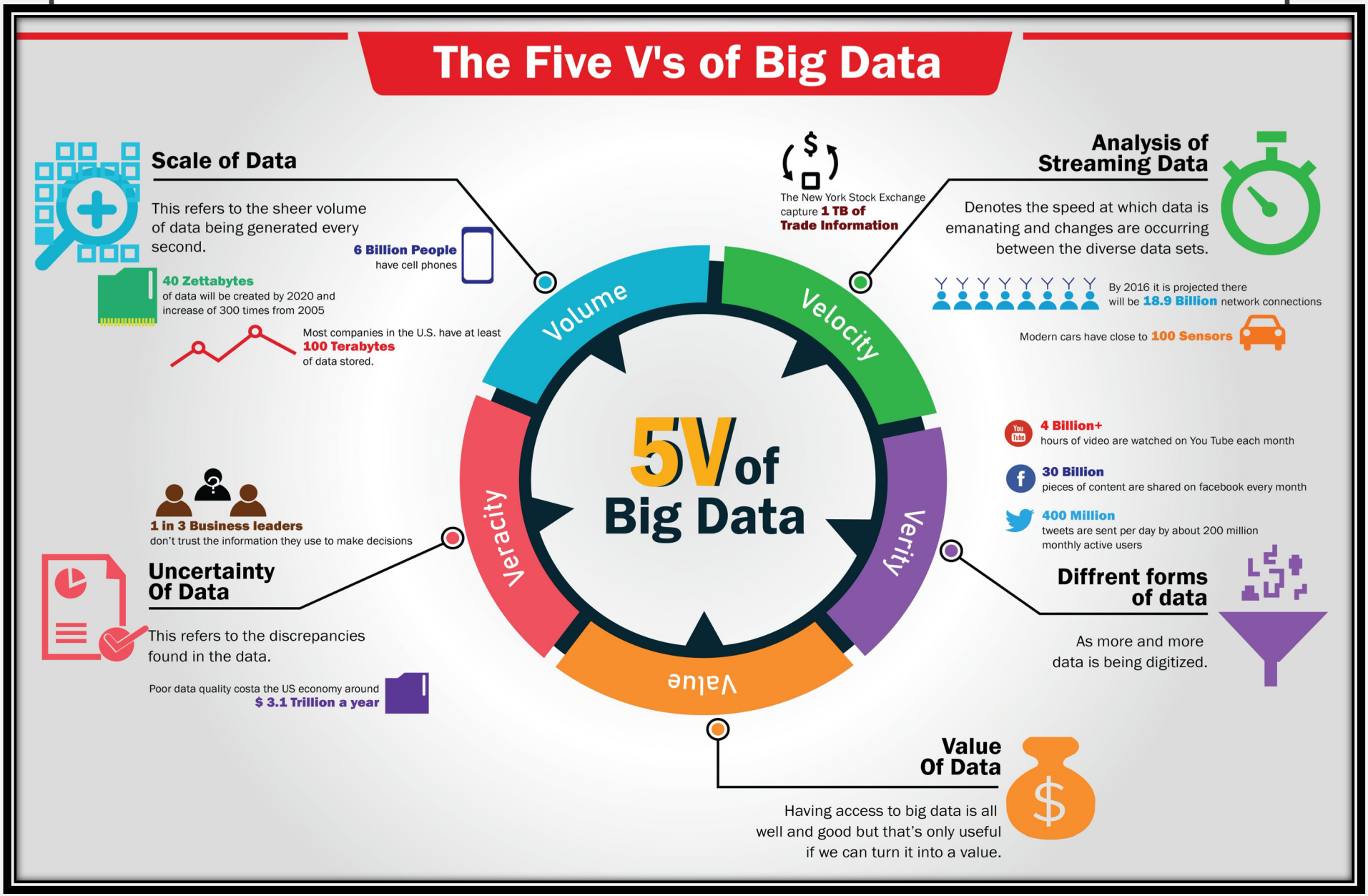
  - techniques, tools and architectures

  …to solve: new problems

  …and, of course, in a better way

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…

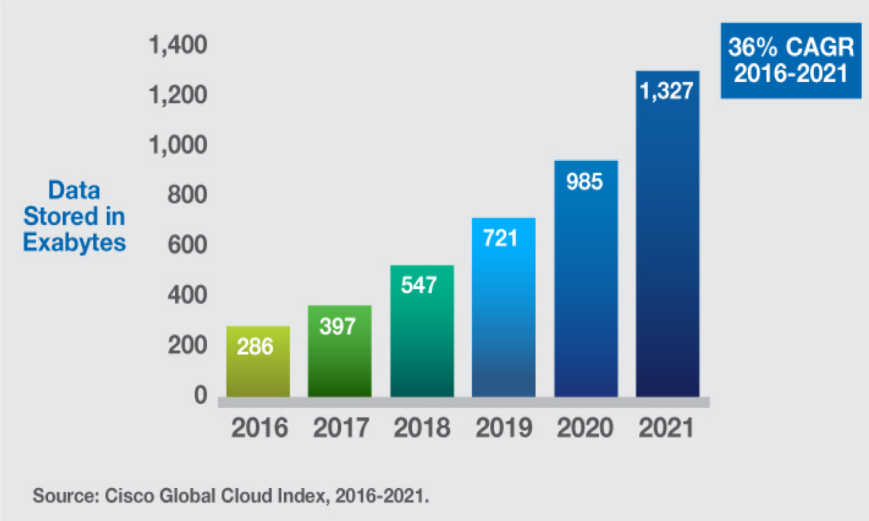
CHARACTERISTICS OF BIG DATA: **V5**

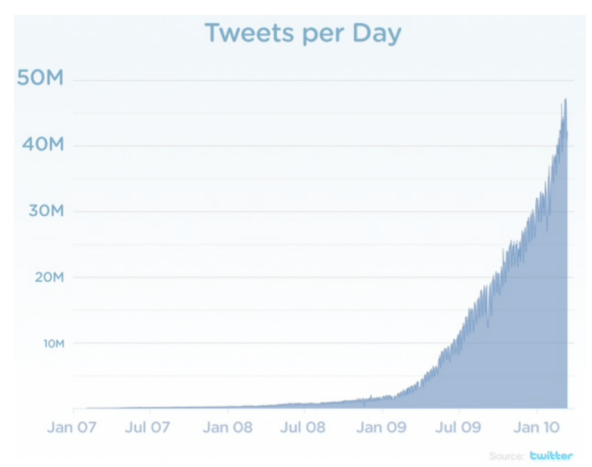


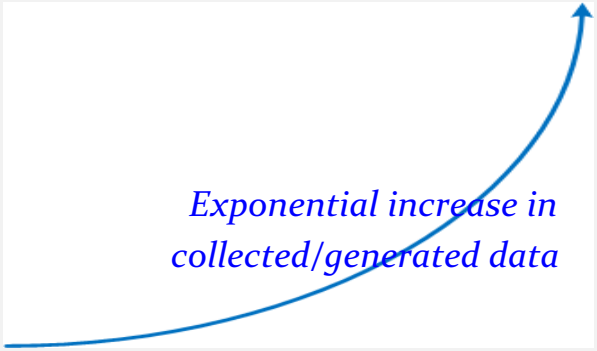
Source: Twitter

5V : V FOR VOLUME

- Volume of data, which needs to be processed is increasing rapidly
  - More storage capacity
  - More computation
  - More tools and techniques

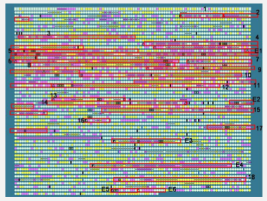


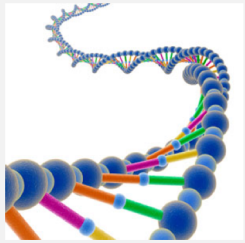


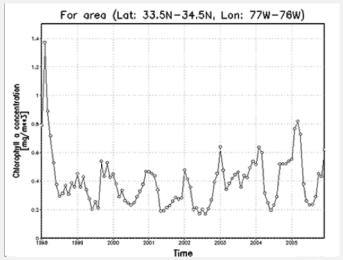


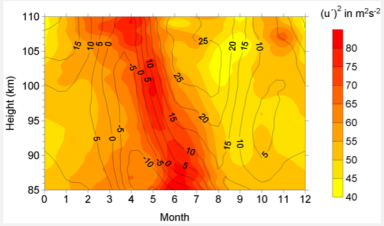
5V: V FOR VARIETY

- Various formats, types, and structures
  - Text, numerical, images, audio, video, sequences, time series, social media data, multi-dimensional arrays, etc…
- Static data vs. streaming data
- A single application can be generating/collecting many types of data









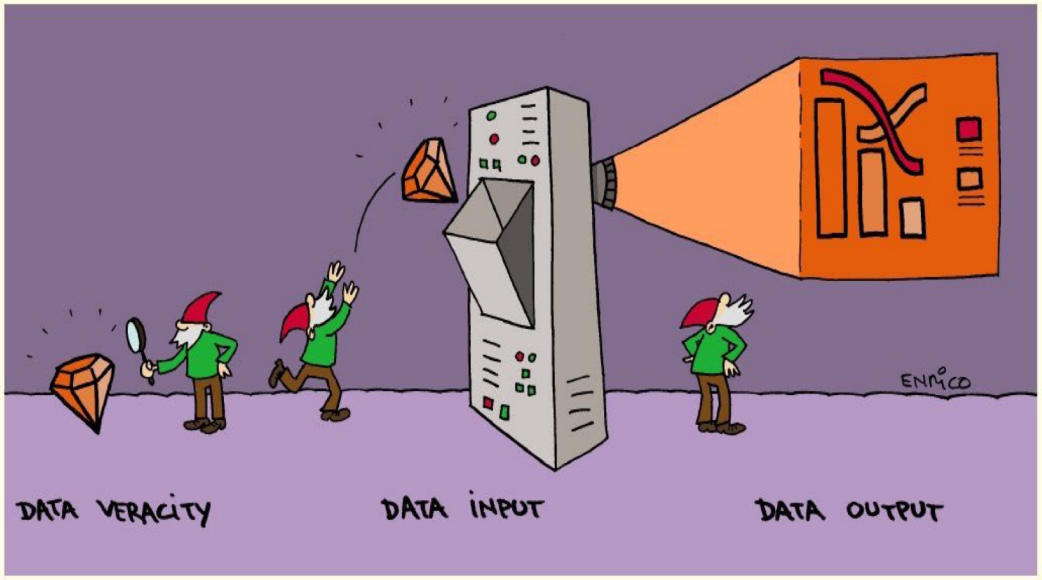
To extract knowledge➔ all these types of data need to be linked together

5V: V FOR VELOCITY

- Data is being generated fast and need to be processed fast
  - For time-sensitive processes such as catching fraud, big data must be used as it streams into your enterprise in order to maximize its value
  - Scrutinize 5 million trade events created each day to identify potential fraud
  - Analyze 500 million daily call detail records in real-time to predict customer churn faster
- Sometimes, 2 minutes is too late!
  - The latest we have heard is 10 ns (nano seconds) delay is too much

5V: V FOR VERACITY

• In a world of (big) data and IoT (data), the veracity of data, *i.e.* the trustworthiness of data (including the related data quality), is more important than ever!



s+a+oo

42

It refers to the quality and accuracy of data. Gathered data could have missing pieces, may be inaccurate or may not be able to provide real, valuable insight. Veracity, overall, refers to the level of trust there is in the collected data.

Data can sometimes become messy and difficult to use. A large amount of data can cause more confusion than insights if it's incomplete.

5V: V FOR VALUE

The last V in the 5 V's of big data is value. This refers to the value that big data can provide, and it relates directly to what organizations can do with that collected data. Being able to pull value from big data is a requirement, as the value of big data increases significantly depending on the insights that can be gained from them.

Organizations can use the same big data tools to gather and analyze the data, but how they derive value from that data should be unique to them.

Healthcare and Medical discoveries

- Prediction and preventive analysis bird flu – prevent mass outbreaks of the virus

Retail and service industries

- Purchase behaviour and preference tracking, real-time personalisations

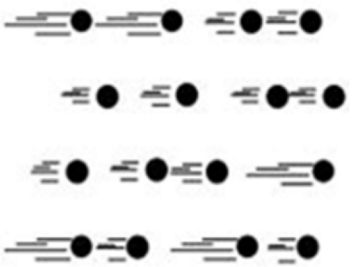
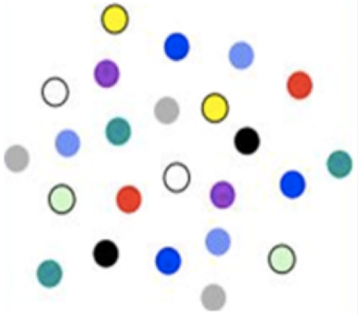
Transportation and Logistics

- Mapping and route planning

Government

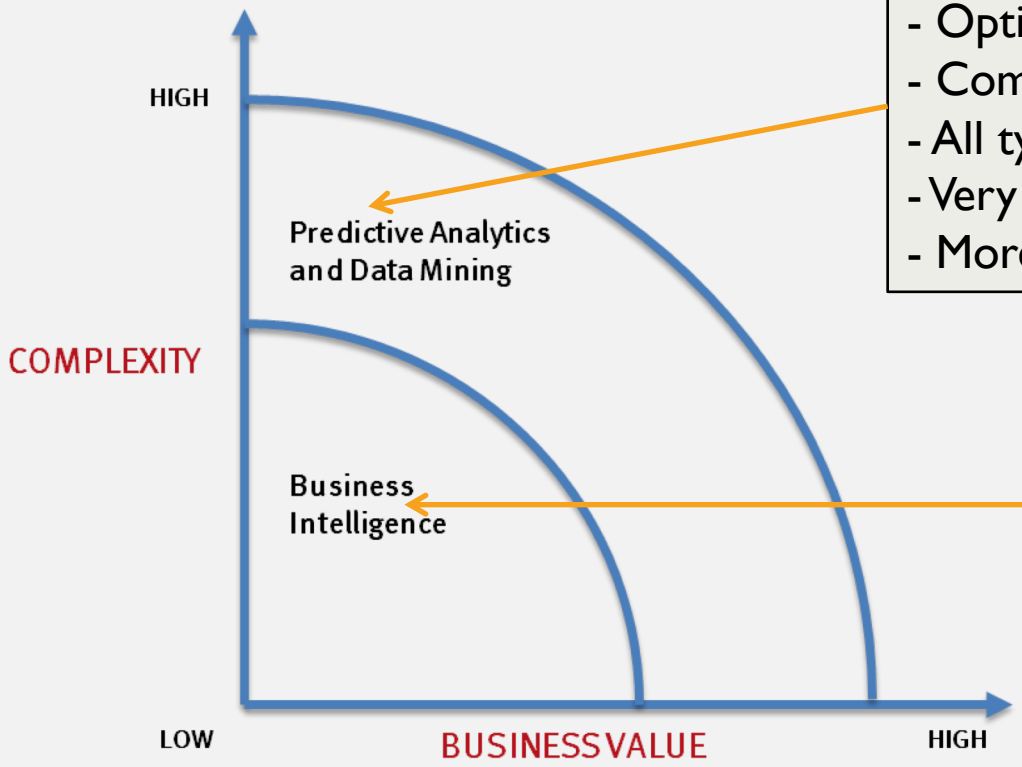
- Cost savings

CHARACTERISTICS OF BIG DATA: **V5**

Volume	Velocity	Variety	Veracity	Value
				
Data at Rest	**Data in Motion**	**Data in Many Forms**	**Data in Doubt**	**Data into Money**
Terabytes to Exabytes of existing data to process	Streaming data, requiring milliseconds to seconds to respond	Structured, unstructured, text, multimedia,...	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations	Business models can be associated to the data

Adapted by a post of Michael Walker on 28 November 2012

BIG DATA VS. SMALL DATA

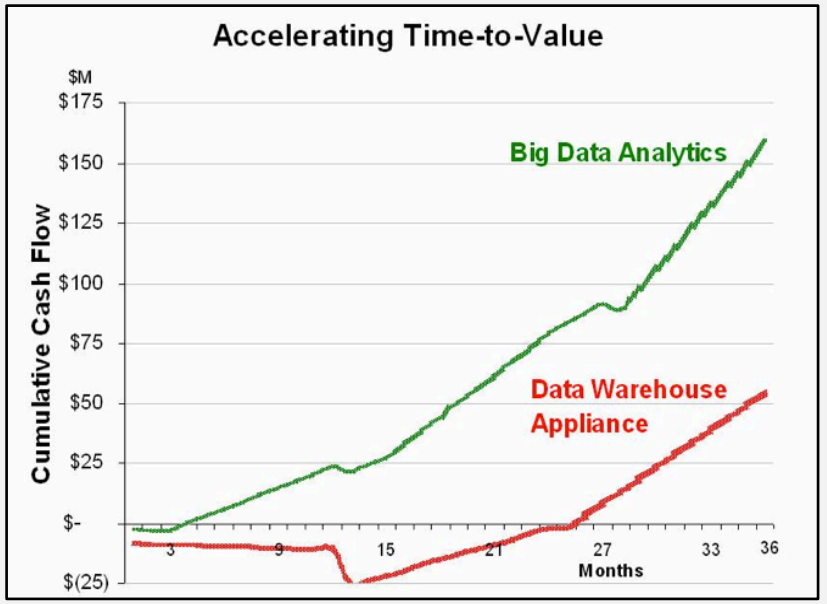


- Optimizations and predictive analytics
- Complex statistical analysis
- All types of data, and many sources
- Very large datasets
- More of a real-time

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

BIG DATA VS. SMALL DATA

- Big data is more real-time in nature than traditional applications
- Big data architecture
  - Traditional architectures are not well-suited for big data applications (e.g. Exa-data, Tera-data)
  - Massively parallel processing, scale out architectures are well-suited for big data applications



CHALLENGES AHEAD…

- **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

Who are the major players in the world of Big data?

Big Data Landscape

Vertical Apps

PREDICTIVE POLICING
bloomreach
MYRRIX

Log Data Apps

splunk> loggly sumologic

Ad/Media Apps

rocketfuel
collective[i]
bluefin
Recorded Future
LuckySort
Media Science
TURN
DataXu

Business Intelligence

ORACLE | Hyperion
SAP Business Objects RJMetrics
Microsoft | Business Intelligence
IBM COGNOS birst
MicroStrategy
Autonomy
bime
QlikView
Chart.io
GoodData

Analytics and Visualization

tableau
Palantir
OPERA
metaLayer
METAMARKETS
dataspora
centrifuge
TERADATA ASTER
SAS TIBCO
KARMASPHERE
panopticon
Real-Time Visual Data Analysis
pentaho
Datameer
platfora
ClearStory
CIRRO
alteryx
visual.ly
AYATA

Data As A Service

factual
kaggle
knoema
GNIP
DATASIFT
Windows Azure Marketplace
INRIX
LexisNexis
LOQATE

Analytics Infrastructure

Hortonworks
VERTICA
MAPR
cloudera
INFOBRIGHT
ParAccel
EMC²
GREENPLUM
NETEZZA
kognitio
DATASTAX
EXASOL
calpont

Operational Infrastructure

CouchBase
10gen
TERADATA
HADAPT
TERRACOTTA
VoltDB
MarkLogic
INFORMATICA

Infrastructure As A Service

amazon web services
Windows Azure
infochimps
Google BigQuery

Structured Databases

ORACLE
MySQL
Microsoft SQL Server
PostgreSQL
IBM DB2
SYBASE
memsql

Technologies

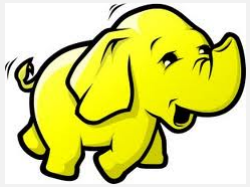
hadoop
hadoop MapReduce
mahout
APACHE HBASE
Cassandra

dave@vcdave.com

blogs.forbes.com/davefeinleib

MAJOR PLAYERS…

- Google
- Hadoop
- MapReduce
- Mahout
- Apache Hbase
- Cassandra



TOOLS AVAILABLE

- **NoSQL**
  - DatabasesMongoDB, CouchDB, Cassandra, Redis, BigTable, Hbase, Hypertable, Voldemort, Riak, ZooKeeper
- **MapReduce**
  - Hadoop, Hive, Pig, Cascading, Cascalog, mrjob, Caffeine, S4, MapR, Acunu, Flume, Kafka, Azkaban, Oozie, Greenplum
- **Storage**
  - S3, HDFS, GDFS
- **Servers**
  - EC2, Google App Engine, Elastic, Beanstalk, Heroku
- **Processing**
  - R, Yahoo! Pipes, Mechanical Turk, Solr/Lucene, ElasticSearch, Datameer, BigSheets, Tinkerpop

Any question?

QUESTIONS OF THE DAY…

1. What is the smallest and largest units of measuring size of data?

2. How big a Quintillion measure is?

3. Give the examples of a smallest the largest entities of data.

4. Give FIVE parameters with which data can be categorized as i) simple, ii) Moderately complex and iii) complex?

QUESTIONS OF THE DAY…

5. What type of data are involved in the following applications?
    1. Weather forecasting
    2. Mobile usage of all customers of a service provider
    3. Anomaly (e.g. fraud) detection in a bank organization
    4. Person categorization, that is, identifying a human
    5. Air traffic control in an airport
    6. Streaming data from all flying aircrafts of Boeing