DATA ANALYTICS

Class #1

Introduction to Data

Dr. Sreeja S R

Assistant Professor

**Indian Institute of Information Technology
IIIT Sri City**

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	
Nibble	4 bits	
Byte	8 bits	
Kilobyte (KB)	1024 Bytes	
Megabyte (MB)	1024 KB	
Gigabyte (GB)	1024 MB	
Terabyte (TB)	1024 GB	
Petabyte (PB)	1024 TB	
Exabyte (EB)	1024 PB	
Zettabyte (ZB)	1024 EB	
Yottabyte (YB)	1024 ZB	

GROWTH OF DATA
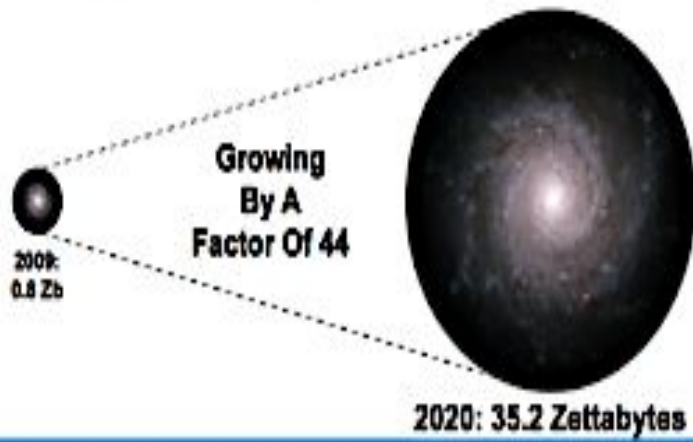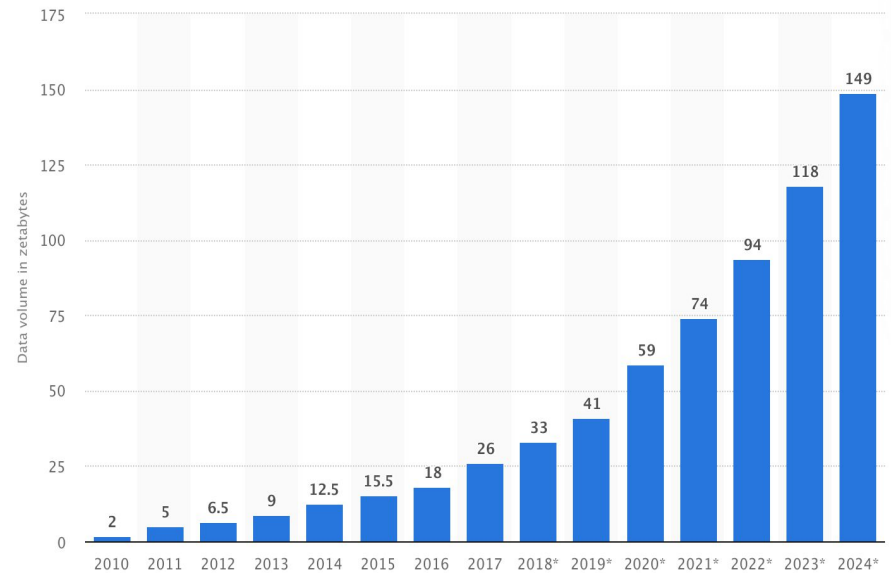
SOURCES OF DATA

- "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

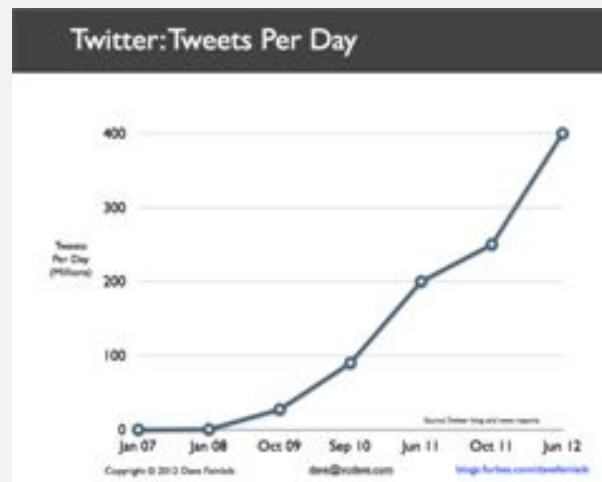   …to solve: new problems

        …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…

CHARACTERISTICS OF BIG DATA: **V3**
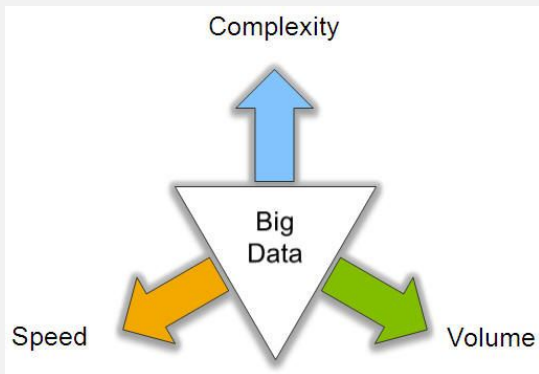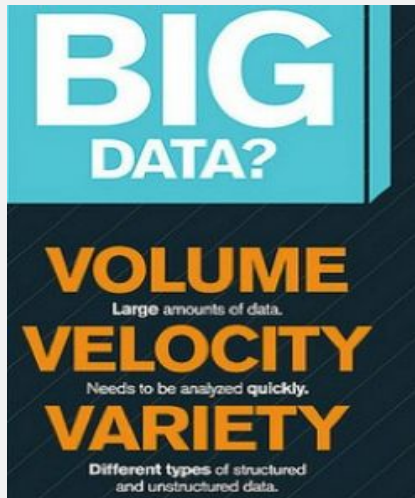
V3 : V FOR VOLUME

- Volume of data, which needs to be processed is increasing rapidly

  - More storage capacity

  - More computation

  - More tools and techniques

Data storage growth
In millions of petabytes
(One petabyte = 1,024 terabytes)

Twitter: Tweets Per Day
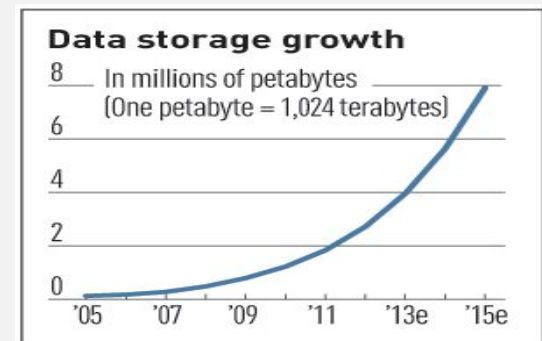
Exponential increase in collected/generated data

V3: 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
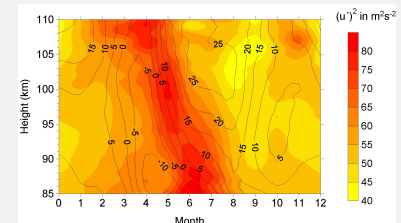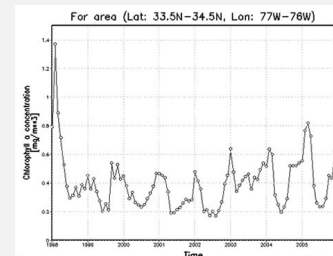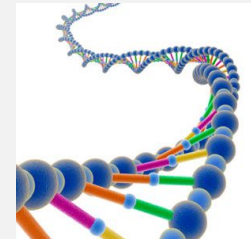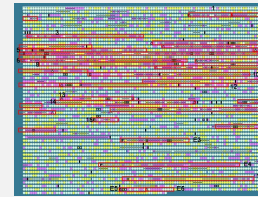
To extract knowledge☐ all these types of data need to be linked together

V3: 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

BIG DATA VS. SMALL DATA

HIGH

COMPLEXITY

Predictive Analytics
and Data Mining

Business
Intelligence

LOW   BUSINESS VALUE   HIGH

- 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
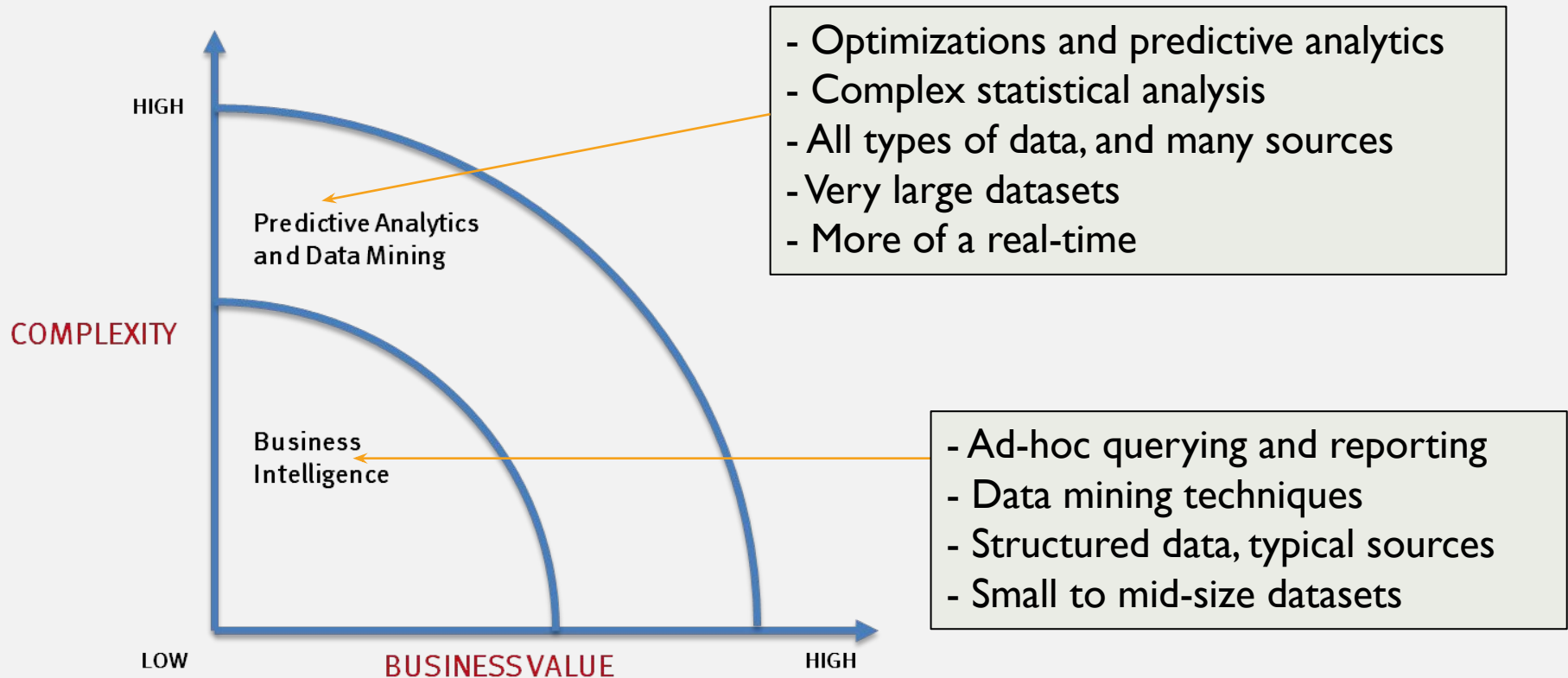
BIG DATA VS. SMALL DATA

- 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
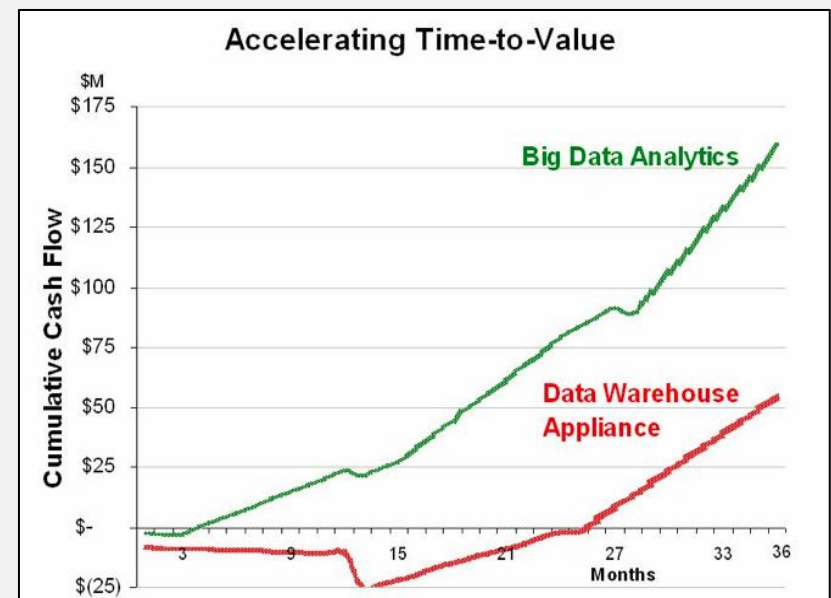
CHALLENGES AHEAD…

- **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. GET FOUND.
MYRRIX

Log Data Apps
splunk> loggly sumologic

Data As A Service
factual.
GNIP DATASIFT Windows Azure Marketplace INRIX LexisNexis SPACE CURVE
kaggle
knoema beta
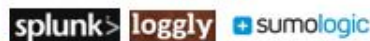LOQATE Everything Location

Ad/Media Apps
rocketfuel
bluefin
Media Science TURN
collective[i]
Recorded Future
LuckySort
DataXu Data. Insight. Action.

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

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

Analytics Infrastructure
Hortonworks VERTICA An HP Company MAPR TECHNOLOGIES
cloudera INFOBRIGHT PARACCEL
EMC² GREENPLUM
NETEZZA kognitio
DATASTAX EXASOL calpont

Operational Infrastructure
COUCHBASE 10gen the MongoDB company
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
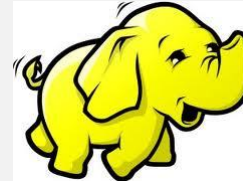
MAJOR PLAYERS…

- 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