Introduction to Big Data & Basic Data Analysis

By: Prof. Prakash Patel
IT,Dept. GIT

What’s Big Data?

- ‘Big-data’ is similar to ‘Small-data’, but bigger
- **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 Data which doesn't fit in storage as well as processing capacity then its BigData.

Why Big Data?

Key enablers for the growth of “Big Data” are:

- Increase of storage capacities
- Increase of processing power
- Availability of data

Challenges to Big Data.

- Capture
- Storage
- Search
- Sharing
- Transfer
- Analysis
- Visualization

Benefits of Big Data

- Using the information kept in the social network like Facebook, the marketing agencies are learning about the response for their campaigns, promotions, and other advertising mediums.
- Using the information in the social media like preferences and product perception of their consumers, product companies and retail organizations are planning their production.
- Using the data regarding the previous medical history of patients, hospitals are providing better and quick service.

Big Data EveryWhere!

- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - purchases at department/ grocery stores
  - Bank/Credit Card transactions
  - Social Network

How much data?

- Google processes 20 PB(10^{15} bytes) a day (2008)
- Wayback Machine has 3 PB + 100 TB/month (3/2009)
- Facebook has 2.5 PB of user data + 15 TB/day (4/2009)
- eBay has 6.5 PB of user data + 50 TB/day (5/2009)
- CERN’s Large Hydron Collider (LHC) generates 15 PB a year

640K ought to be enough for anybody.

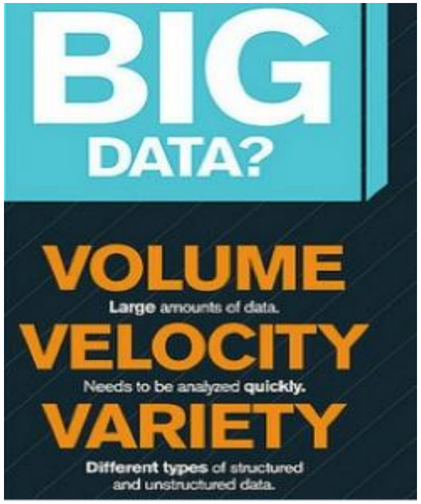
Type of Data

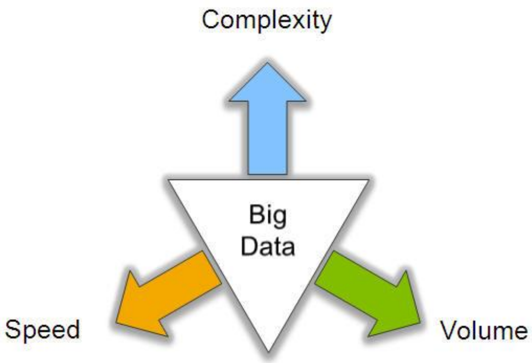
- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
  - Social Network, Semantic Web
- Streaming Data
  - You can only scan the data once

What to do with these data?

- Aggregation and Statistics
  - Data warehouse and OLAP(Online Analytical Processing)
- Indexing, Searching, and Querying
  - Keyword based search
  - Pattern matching
    - (XML(Extensible Markup Language)
    - RDF(Resource Description Framework))
- Knowledge discovery
  - Data Mining
  - Statistical Modeling

Big Data: 3V's

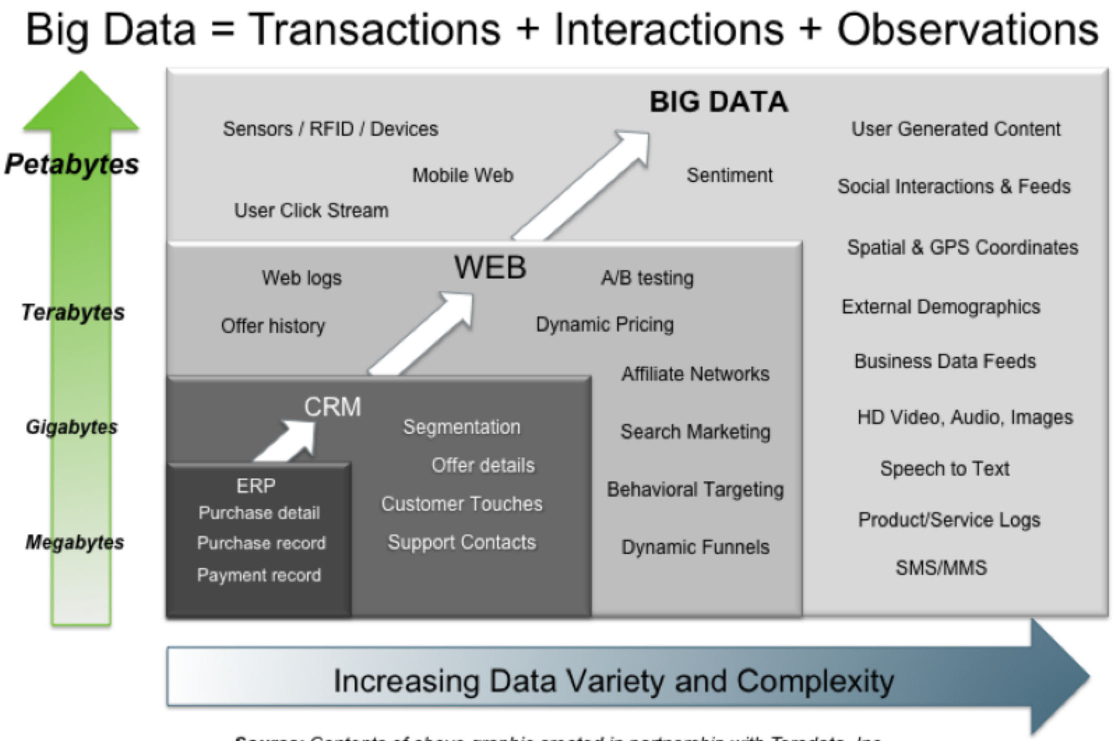

BIG DATA?
VOLUME
Large amounts of data.
VELOCITY
Needs to be analyzed quickly.
VARIETY
Different types of structured and unstructured data.


Complexity

Big Data

Speed

Volume


Big Data = Transactions + Interactions + Observations

Petabytes

Terabytes

Gigabytes

Megabytes

BIG DATA

Sensors / RFID / Devices

Mobile Web

User Click Stream

Sentiment

User Generated Content

Social Interactions & Feeds

Spatial & GPS Coordinates

External Demographics

Business Data Feeds

HD Video, Audio, Images

Speech to Text

Product/Service Logs

SMS/MMS

WEB

Web logs

Offer history

A/B testing

Dynamic Pricing

Affiliate Networks

Search Marketing

Behavioral Targeting

Dynamic Funnels

CRM

Segmentation

Offer details

Customer Touches

Support Contacts

ERP

Purchase detail

Purchase record

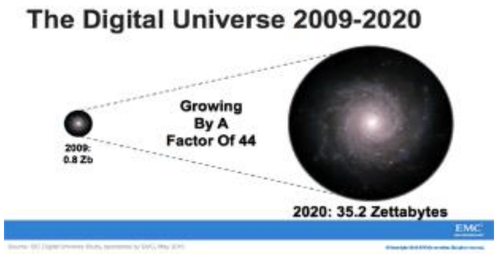
Payment record

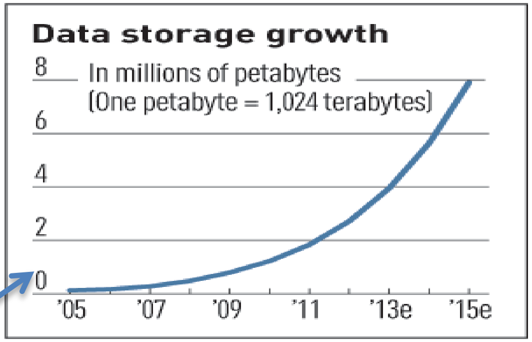
Increasing Data Variety and Complexity

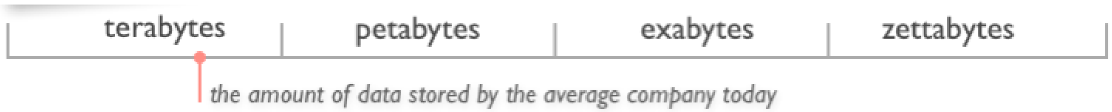
***Source**: Contents of above graphic created in partnership with Teradata, Inc.*

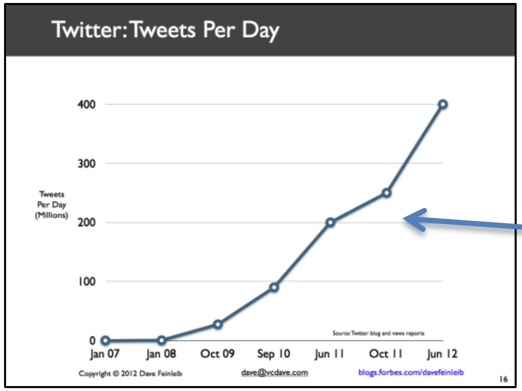
Volume (Scale)

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









Exponential increase in collected/generated data

12+ TBs of tweet data every day

? TBs of 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

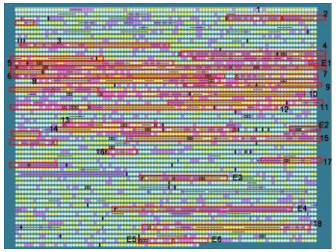
Variety (Complexity)

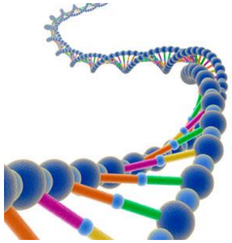
- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
  - Social Network, Semantic Web (RDF), …

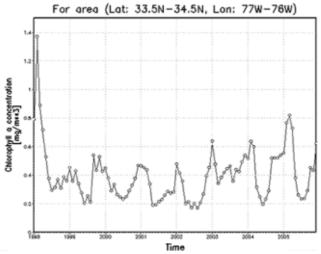
- Streaming Data
  - You can only scan the data once

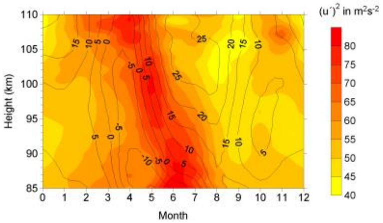
- A single application can be generating/collecting many types of data

- Big Public Data (online, weather, finance, etc)



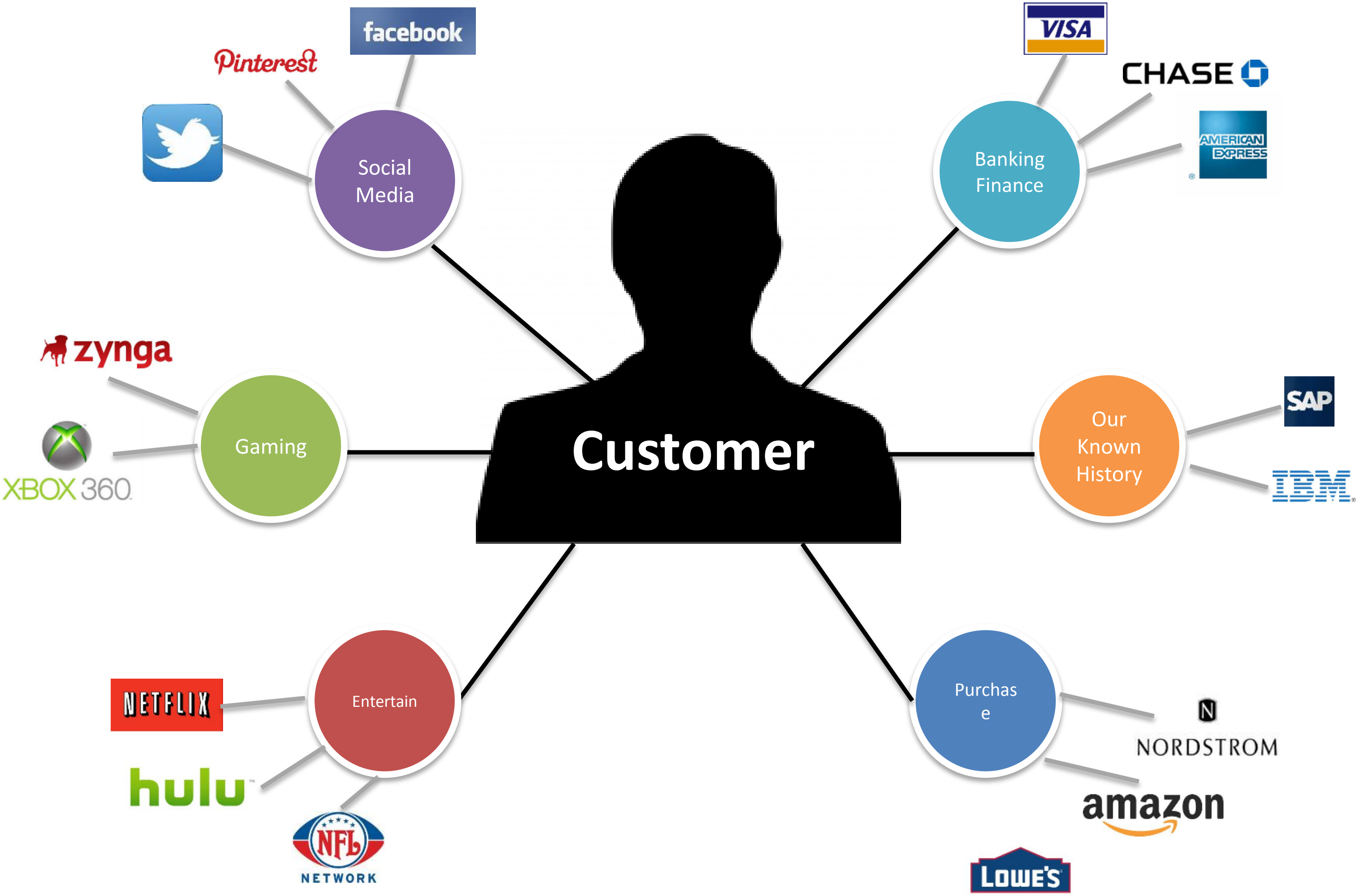



For area (Lat: 33.5N–34.5N, Lon: 77W–76W)
Chlorophyll a concentration [mg/m**3]
Time



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

A Single View to the Customer



Customer

Social Media

Banking Finance

Gaming

Our Known History

Entertain

Purchas e

Velocity (Speed)

- Data is begin generated fast and need to be processed fast
- Online Data Analytics
- Late decisions ➔ missing opportunities

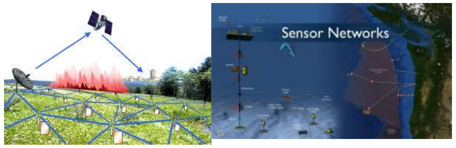
- **Examples**
  - **E-Promotions:** Based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you
  - **Healthcare monitoring:** sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction

Real-time/Fast Data

Social media and networks
(all of us are generating data)

Scientific instruments
(collecting all sorts of data)

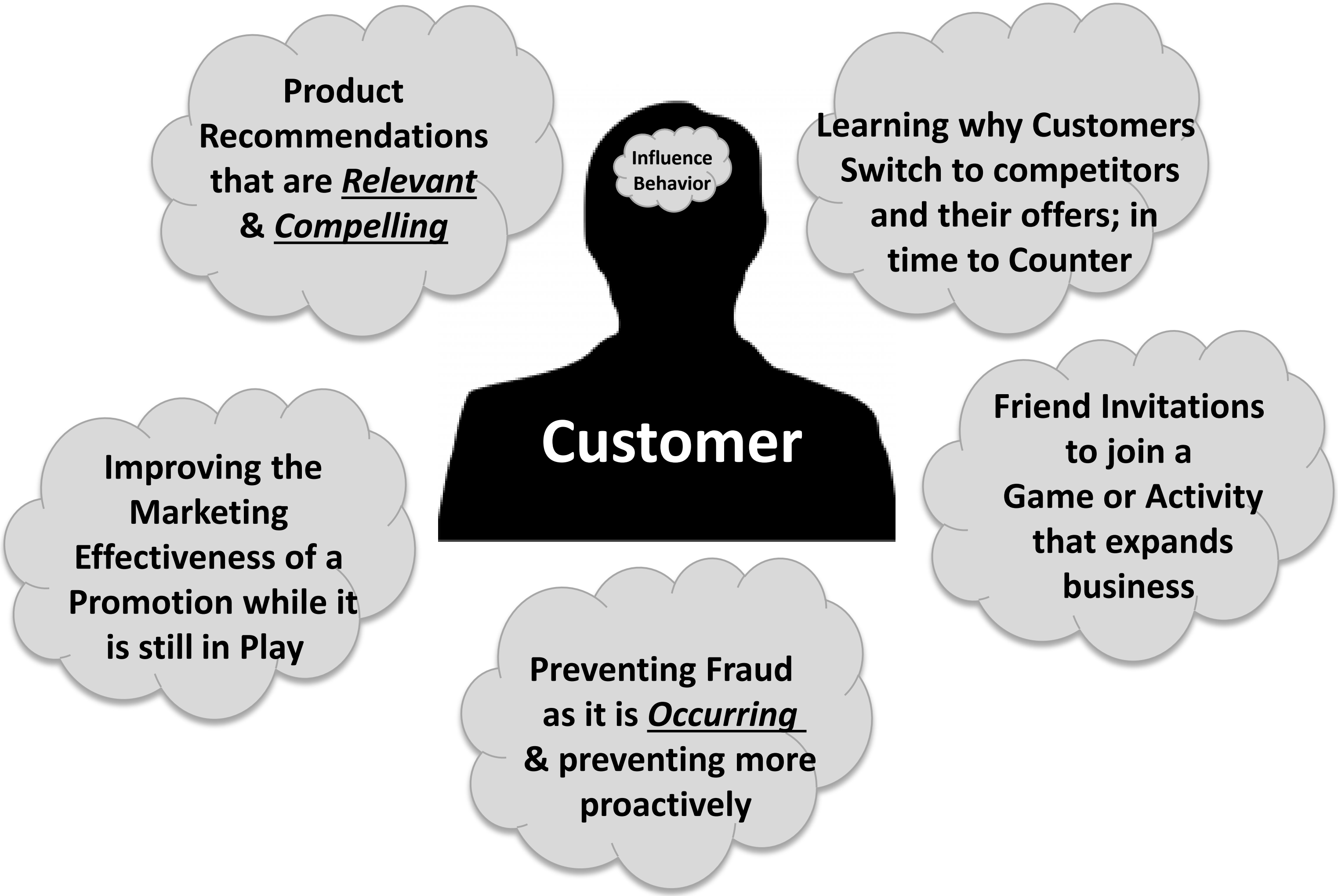
Mobile devices
(tracking all objects all the time)



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



Influence Behavior

Customer

Product Recommendations that are *Relevant* & *Compelling*

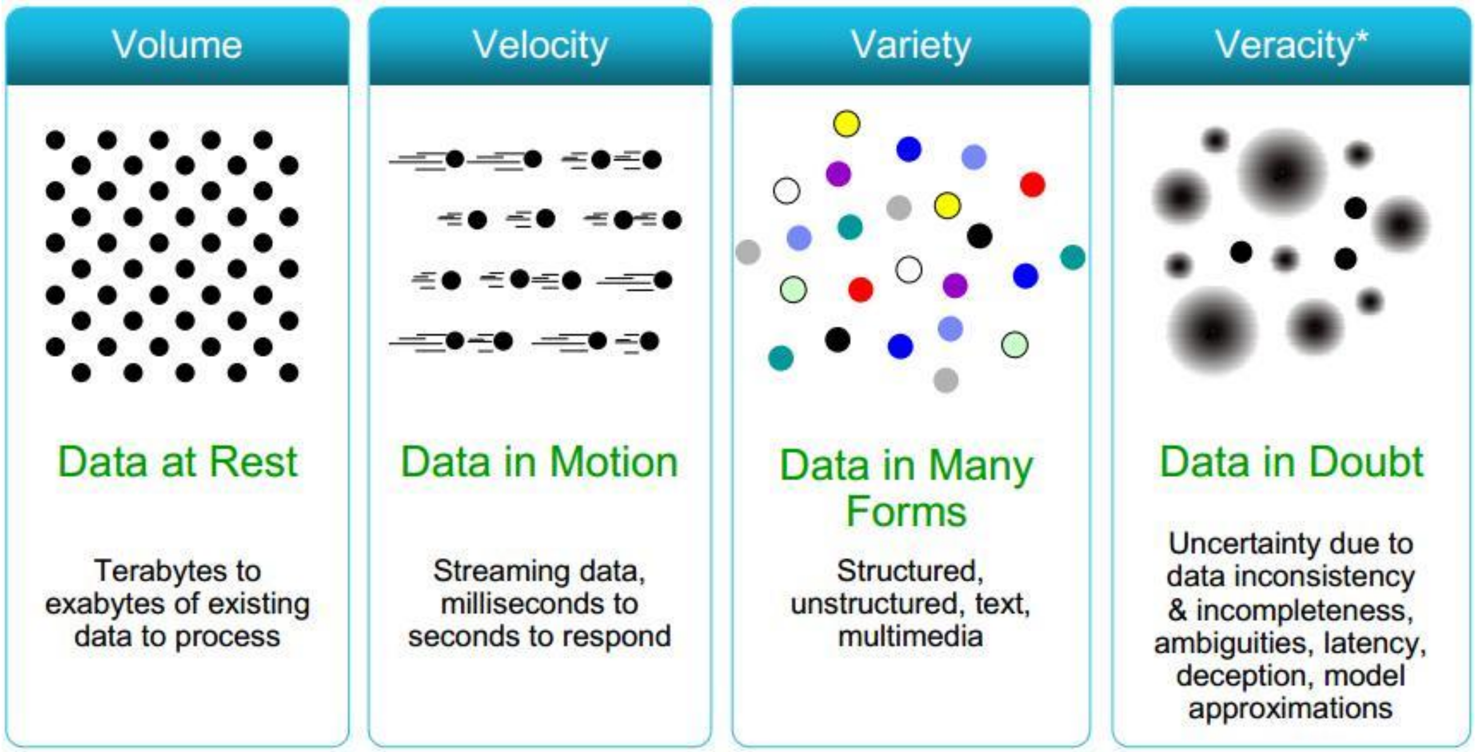
Learning why Customers Switch to competitors and their offers; in time to Counter

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

Friend Invitations to join a Game or Activity that expands business

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

Some Make it 4V's



Volume

Data at Rest

Terabytes to exabytes of existing data to process

Velocity

Data in Motion

Streaming data, milliseconds to seconds to respond

Variety

Data in Many Forms

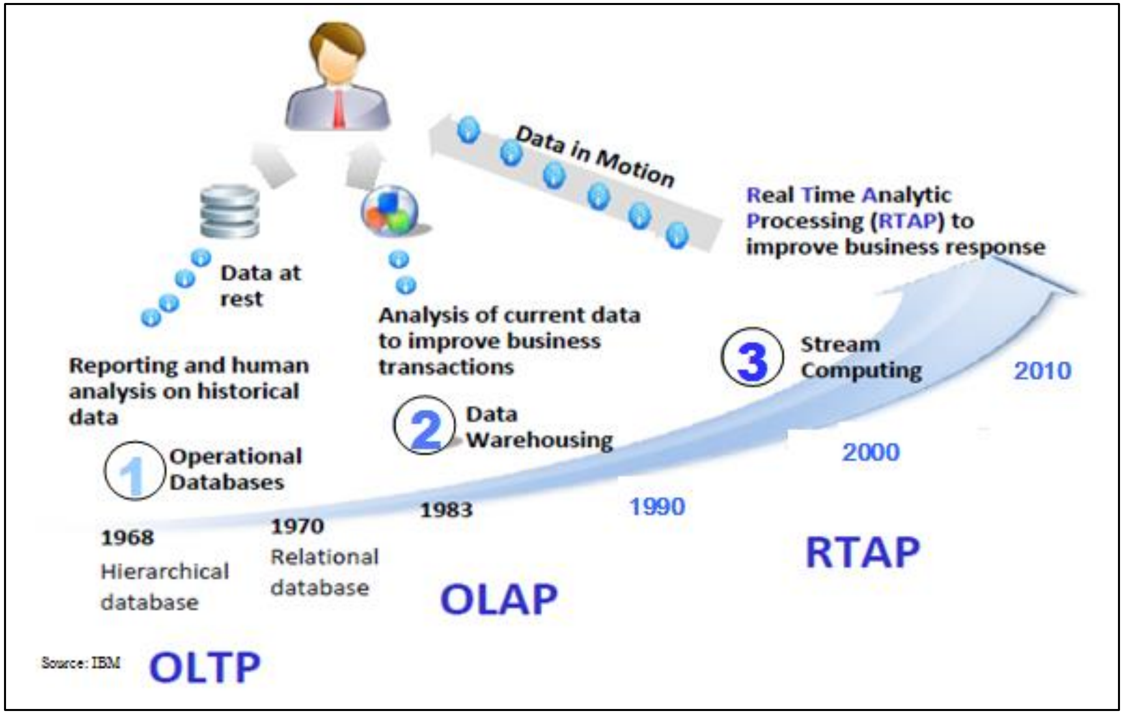
Structured, unstructured, text, multimedia

Veracity*

Data in Doubt

Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Connecting 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
1990
1983
1970 Relational database
1968 Hierarchical database
RTAP
OLAP
OLTP

Source: IBM

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

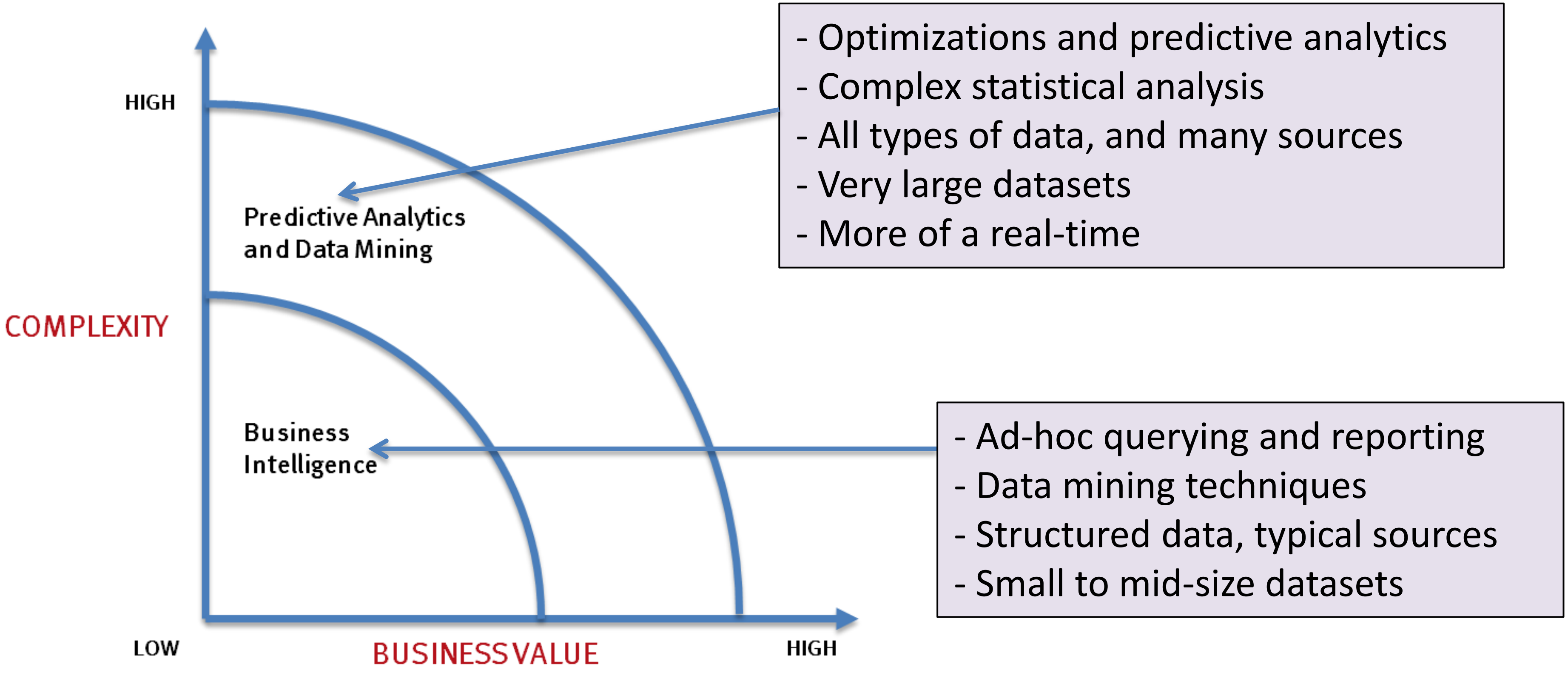
The Model Has Changed…

- **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

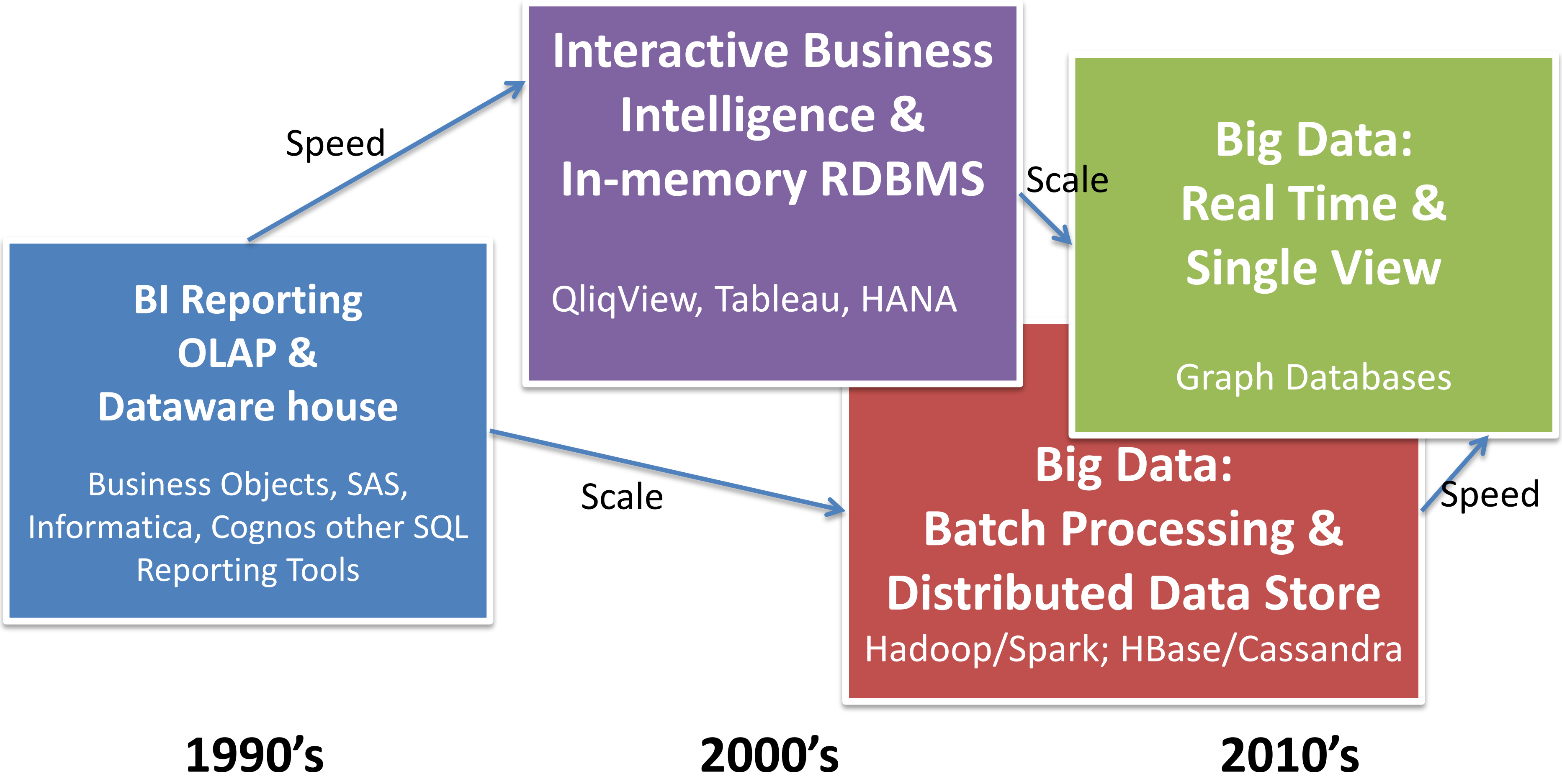
What’s driving Big Data



- 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

Interactive Business Intelligence & In-memory RDBMS

QliqView, Tableau, HANA

Scale

Big Data: Batch Processing & Distributed Data Store

Hadoop/Spark; HBase/Cassandra

Scale

Big Data: Real Time & Single View

Graph Databases

Speed

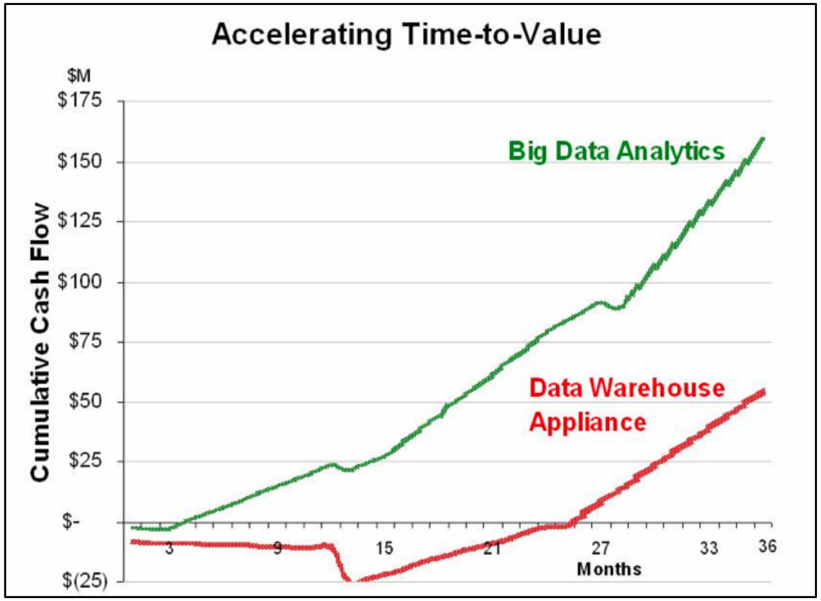
1990's

2000's

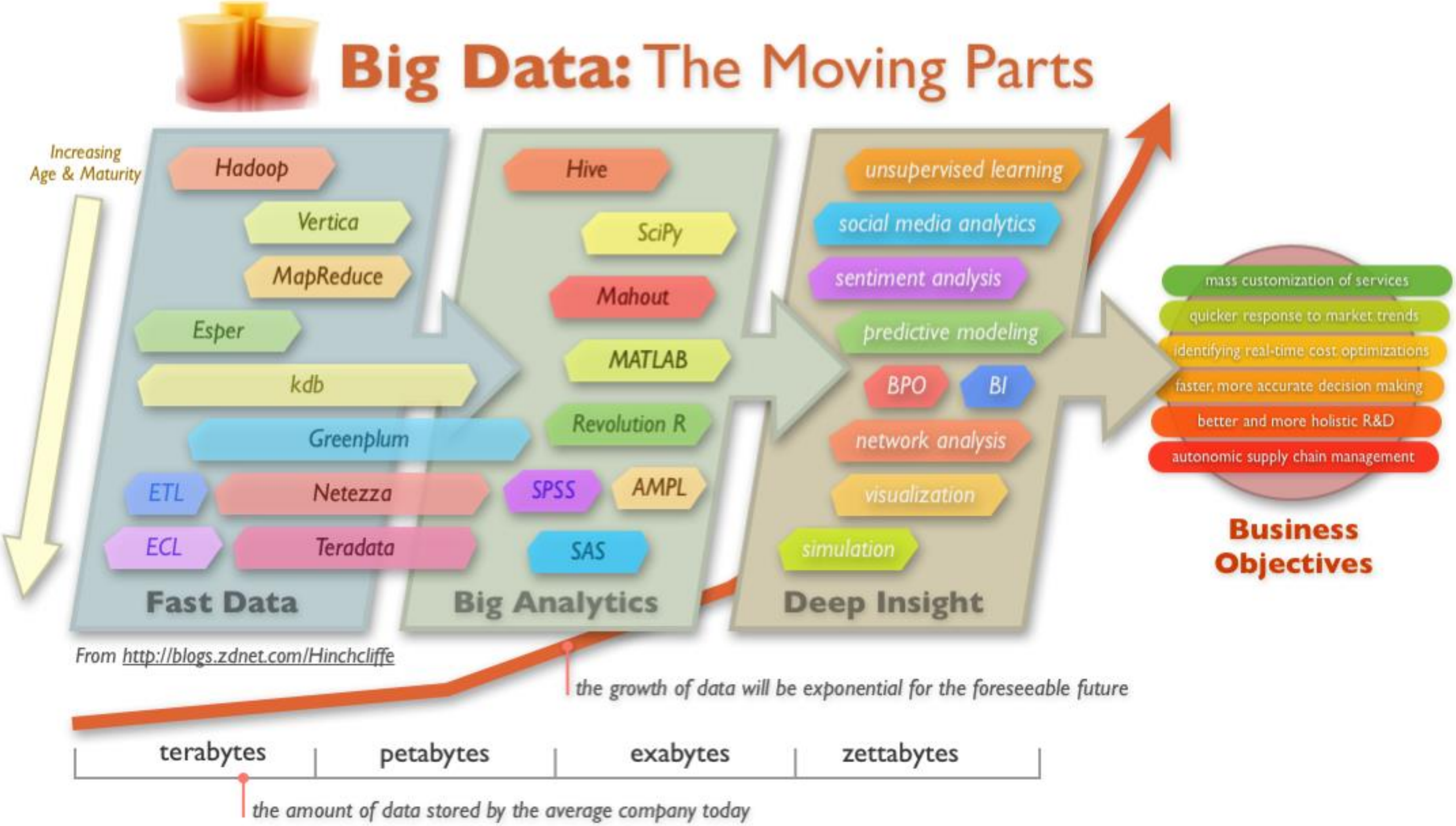
2010's

Big Data Analytics

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



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