Big Data – Data Science

Srinivas Bhoosarapu
(TechnoEconomist, Former CISO and Chief FinTech & Innovation Officer)

$One Trillion Digital Economy
Viksit Bharat - 2047

Agenda

- Big Data
- Data Science
- LLM
- Use cases and Case Studies
- Conclusion

Big Data – Data Science

Large-Scale Data Management

Big Data Analytics

Data Science and Analytics

- How to manage very large amounts of data and extract value and knowledge from them

Introduction to Big Data

What is Big Data?

What makes data, "Big" Data?

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…

What is Big Data?

- What is Big Data

- Massive sets of unstructured/semi-structured data from Web traffic, social media, sensors, etc
- Petabytes, exabytes of data
    - Volumes too great for typical DBMS
- Information from multiple internal and external sources:
    - Transactions
    - Social media
    - Enterprise content
    - Sensors
    - Mobile devices

- In the last minute there were …….

What is Big Data?

- What is Big Data

- 204 million emails sent
- 61,000 hours of music listened to on Pandora
- 20 million photo views

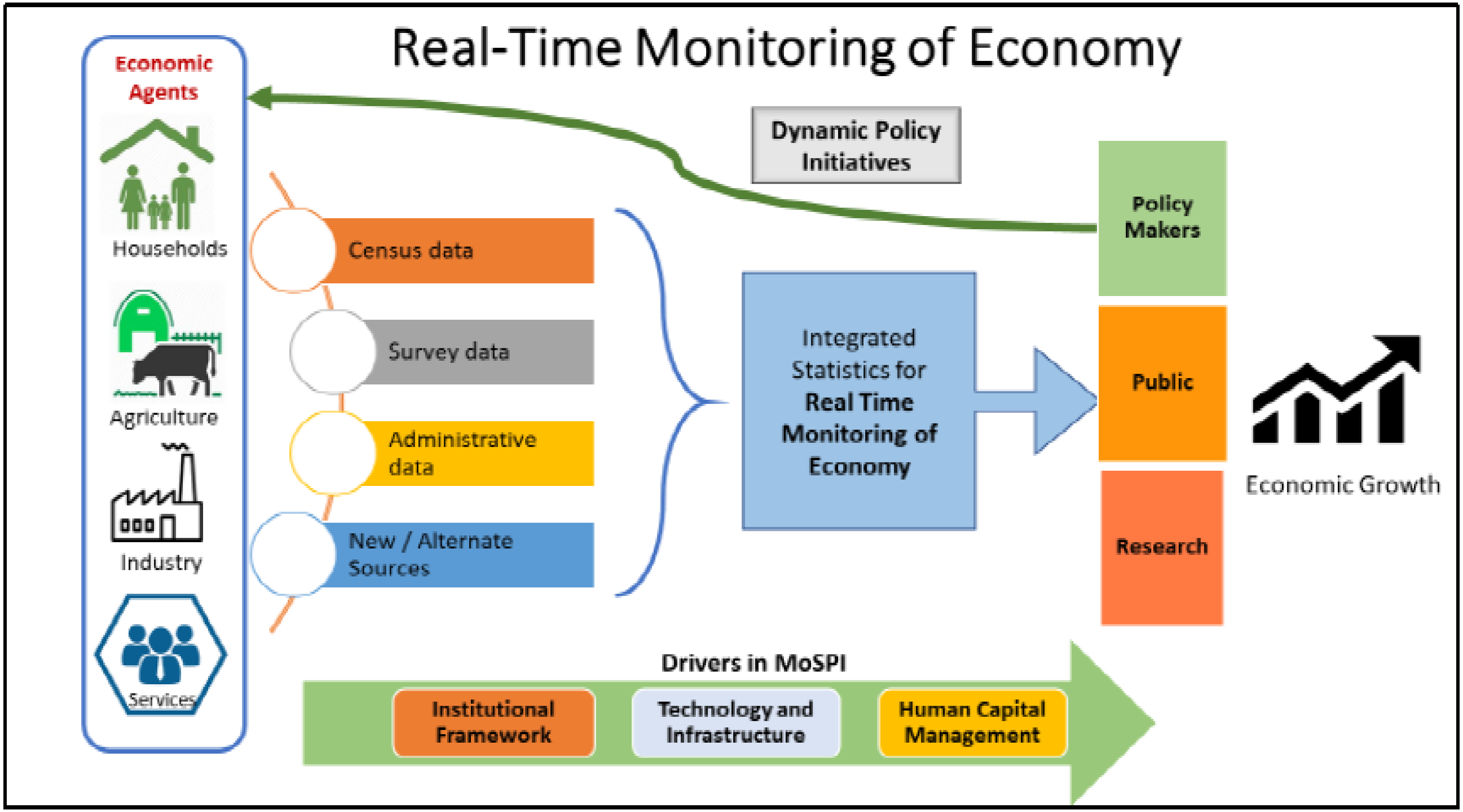
- 100,000 tweets
- 6 million views and 277,000 Facebook Logins
- 2+ million Google searches
- 3 million uploads on Flickr

What is Big Data?

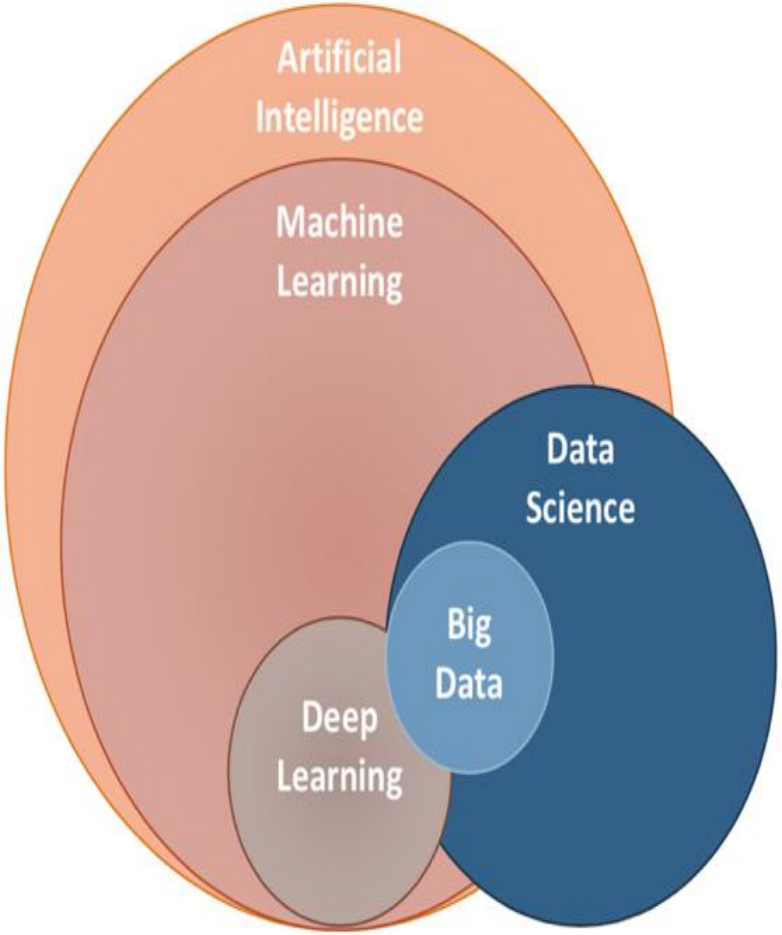
- Companies leverage data to adapt products and services to:
  - Meet customer needs
  - Optimize operations
  - Optimize infrastructure
  - Find new sources of revenue
  - Can reveal more patterns and anomalies

- IBM estimates that by 2015 4.4 million jobs will be created globally to support big data
  - 1.9 million of these jobs will be in the United States

Real Time Monitoring



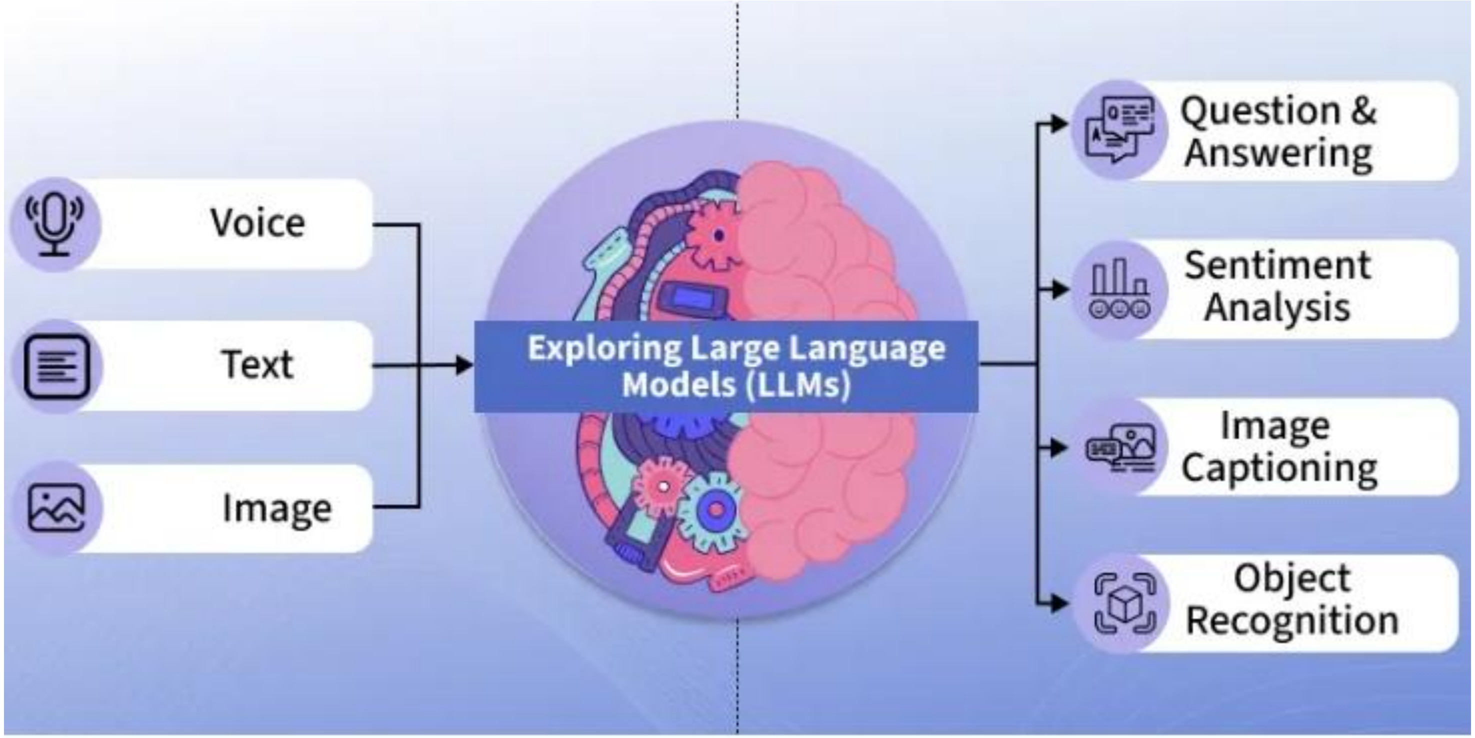
Data Science



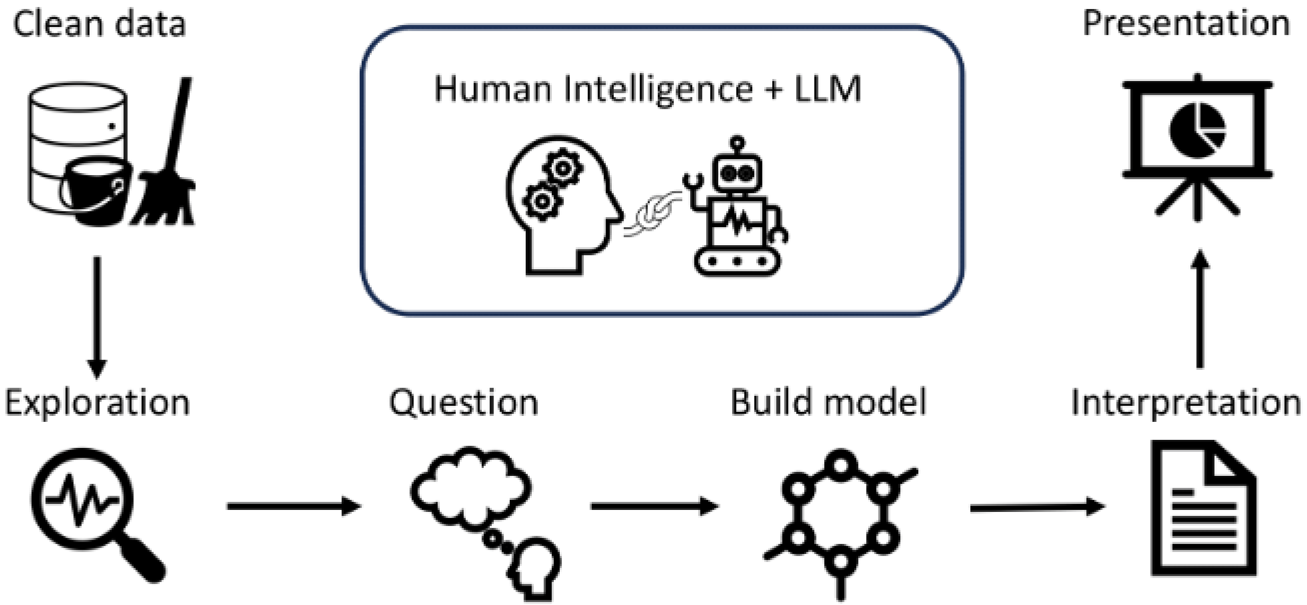
Data science is an **interdisciplinary field** that combines scientific methods, algorithms, and systems to extract knowledge and actionable insights from both structured and unstructured data. It serves as the foundation for modern **Artificial Intelligence (AI)** and **machine learning**, allowing organizations to make informed, data-driven decisions rather than relying on intuition.

LLM

Large Language Models (LLMs) are advanced AI systems built on deep neural networks designed to process, understand and generate human-like text



LLM



Clean data

Human Intelligence + LLM

Presentation

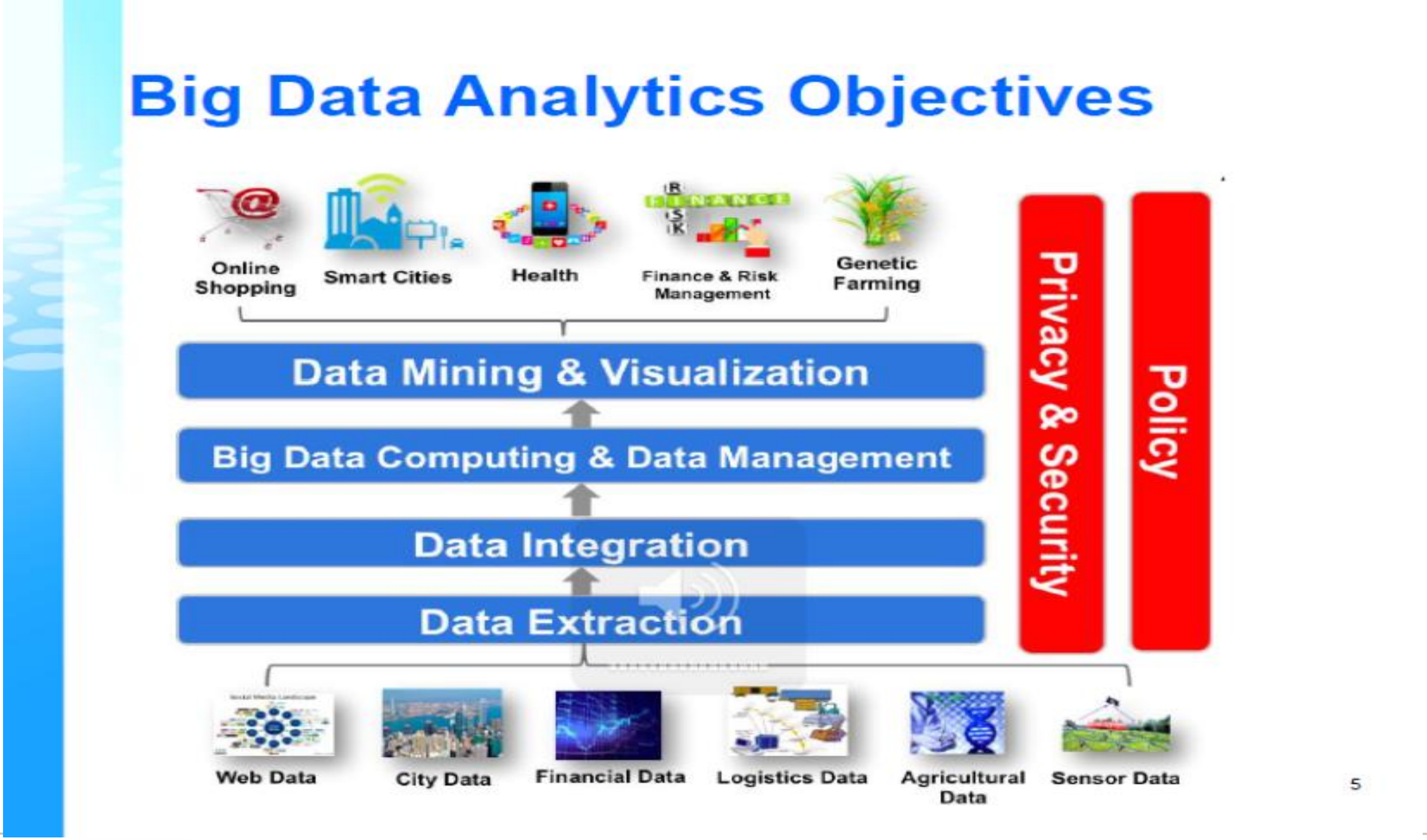
Exploration

Question

Build model

Interpretation

Big Data Analytics



Big Data Analytics Objectives

Online Shopping

Smart Cities

Health

Finance & Risk Management

Genetic Farming

Data Mining & Visualization

Big Data Computing & Data Management

Data Integration

Data Extraction

Privacy & Security

Policy

Web Data

City Data

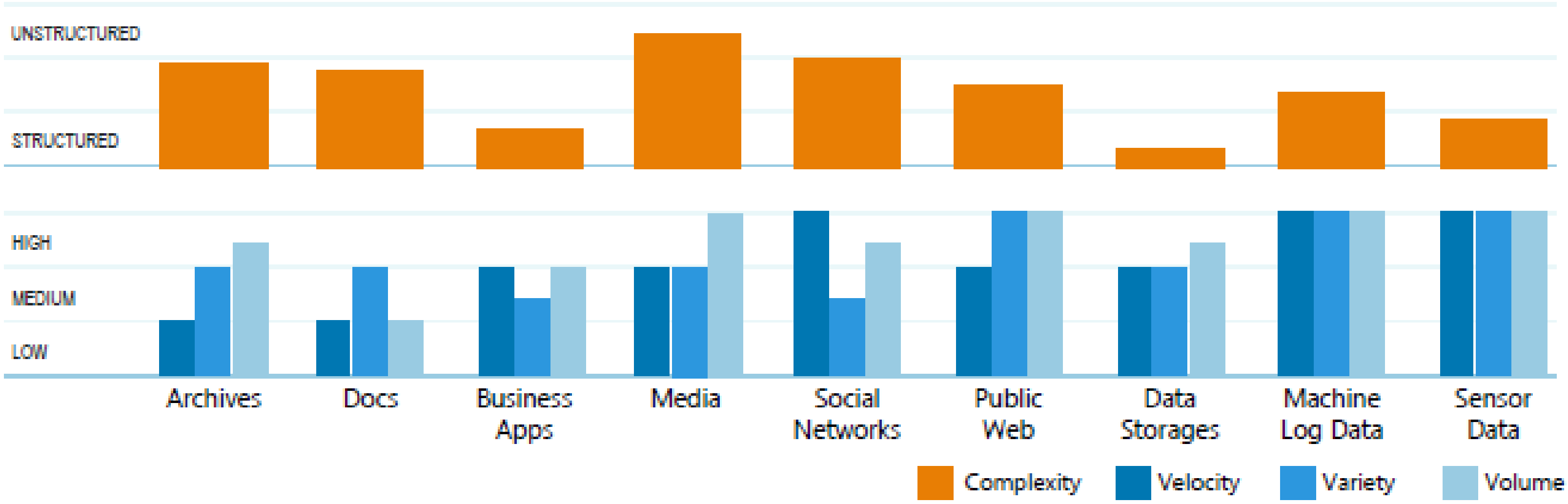
Financial Data

Logistics Data

Agricultural Data

Sensor Data

Big Data Challenges



Archives
Scanned documents, statements, medical records, e-mails etc..

Docs
XLS, PDF, CSV, HTML, JSON etc.

Business Apps
CRM, ERP systems, HR, project management etc.

Media
Images, video, audio etc.

Social Networks
Twitter, Facebook, Google+, LinkedIn etc.

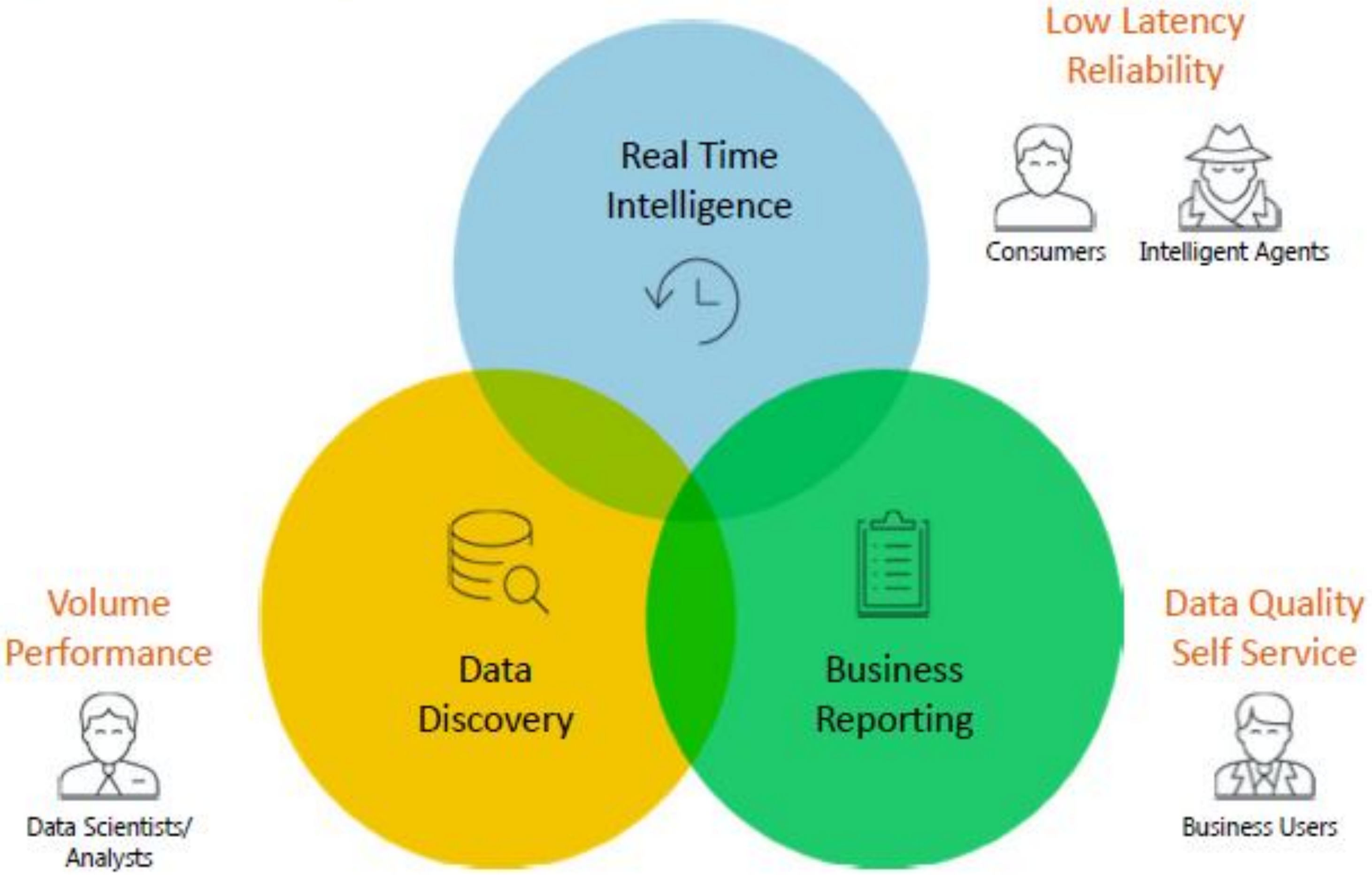
Public Web
Wikipedia, news, weather, public finance etc

Data Storages
RDBMS, NoSQL, Hadoop, file systems etc.

Machine Log Data
Application logs, event logs, server data, CDRs, clickstream data etc.

Sensor Data
Smart electric meters, medical devices, car sensors, road cameras etc.

Big Data Analytics Use Cases



Real Time Intelligence

Low Latency
Reliability

Consumers

Intelligent Agents

Data Discovery

Volume
Performance

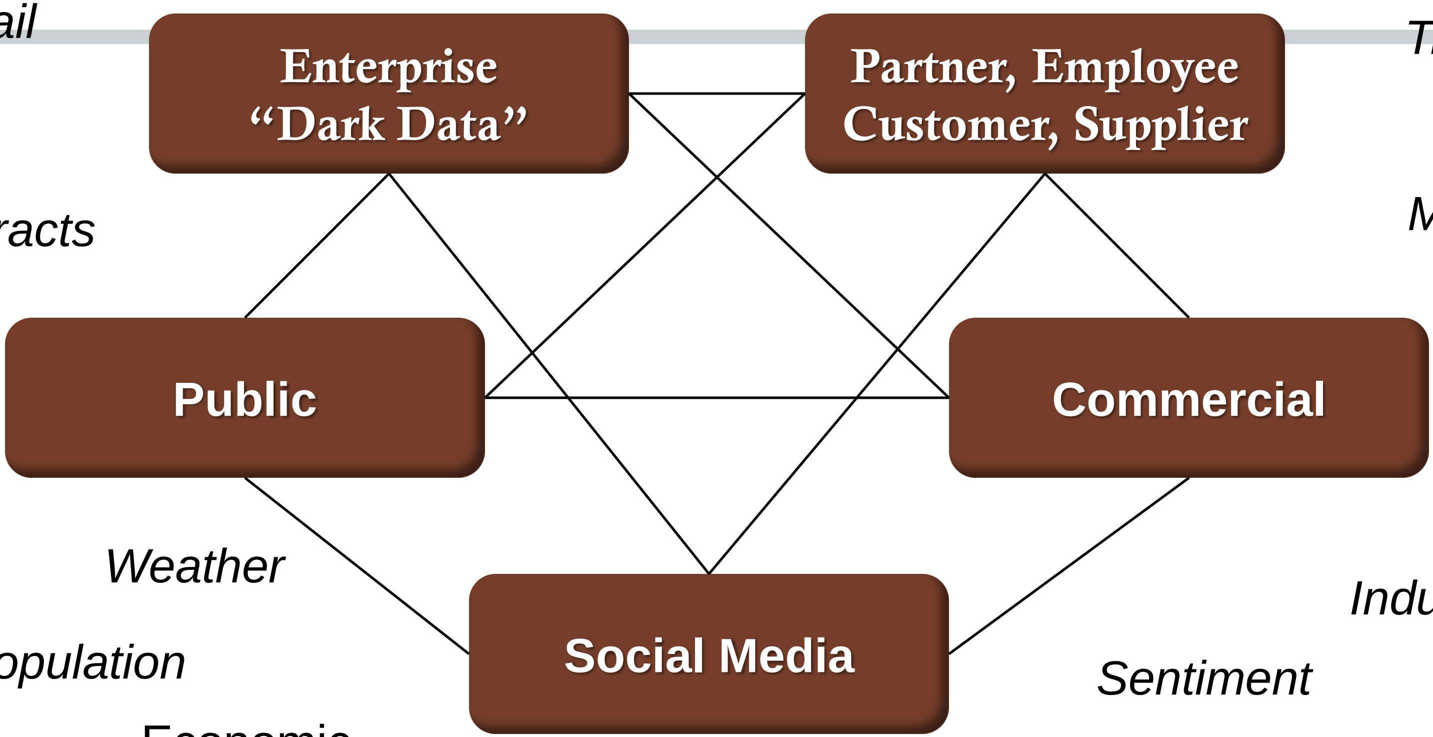
Data Scientists/ Analysts

Business Reporting

Data Quality
Self Service

Business Users

Where does Big Data come from?



Enterprise
“Dark Data”

Partner, Employee
Customer, Supplier

Public

Commercial

Social Media

Email

Transactions

Contracts

Monitoring

Credit

Sensor

Weather

Industry

Population

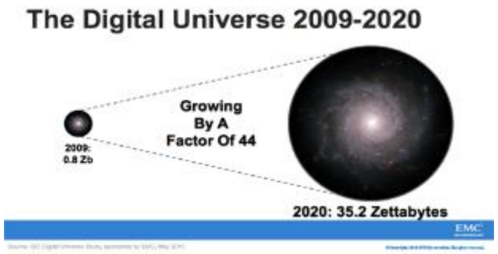
Sentiment

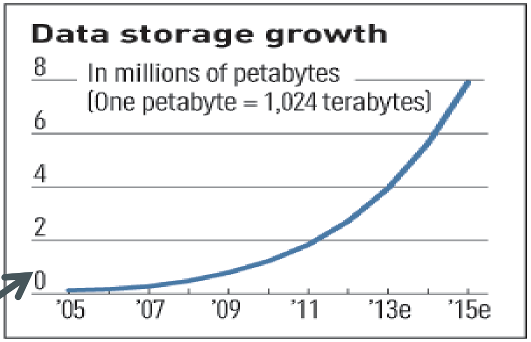
Economic

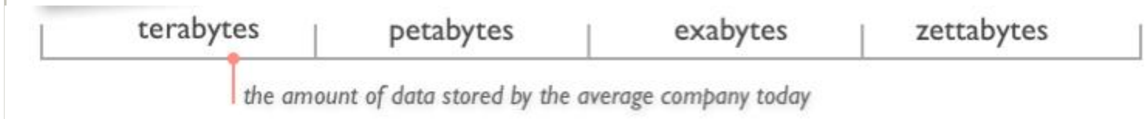
Network

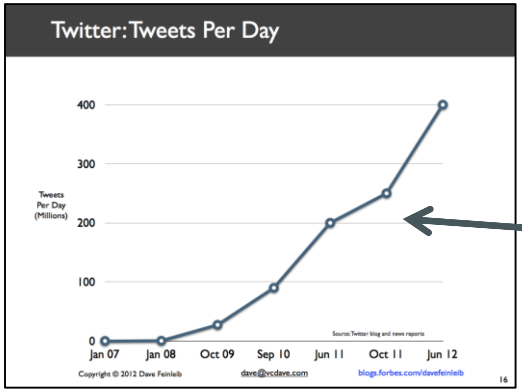
Characteristics of Big Data:
1-Scale (Volume)

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









Source: Twitter blog and news reports

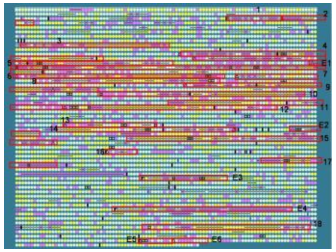
Copyright © 2012 Dave Feinleib dave@vcdave.com blogs.forbes.com/davefeinleib

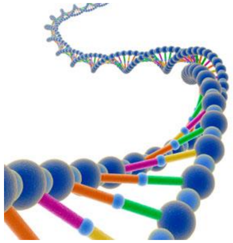
Exponential increase in collected/generated data

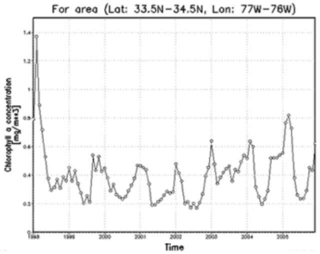
Characteristics of Big Data: 2-Complexity (Variety)

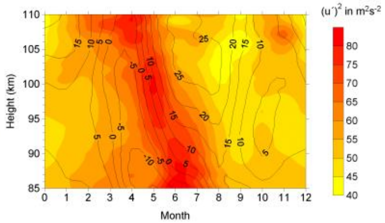
- Various formats, types, and structures
- Text, numerical, images, audio, video, sequences, time series, social media data, multi-dim arrays, etc…
- Static data vs. streaming data
- A single application can be generating/collecting many types of data

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





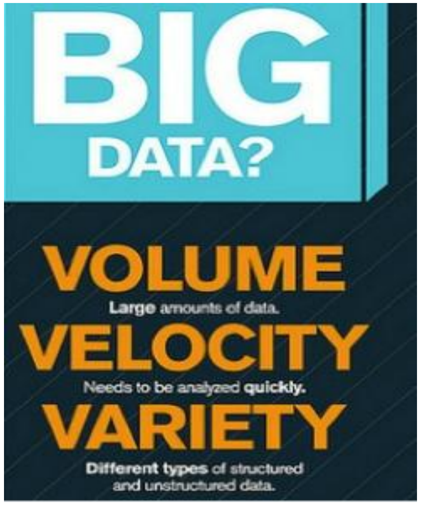

For area (Lat: 33.5N–34.5N, Lon: 77W–76W)
Chlorophyll a concentration [mg/m3]
Time

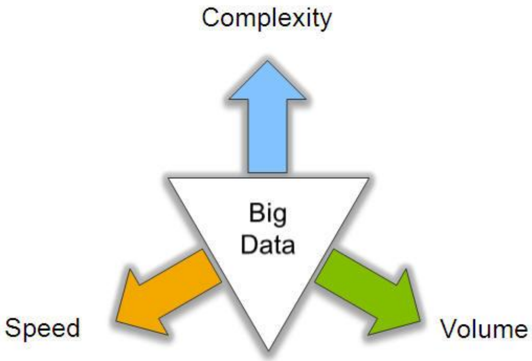


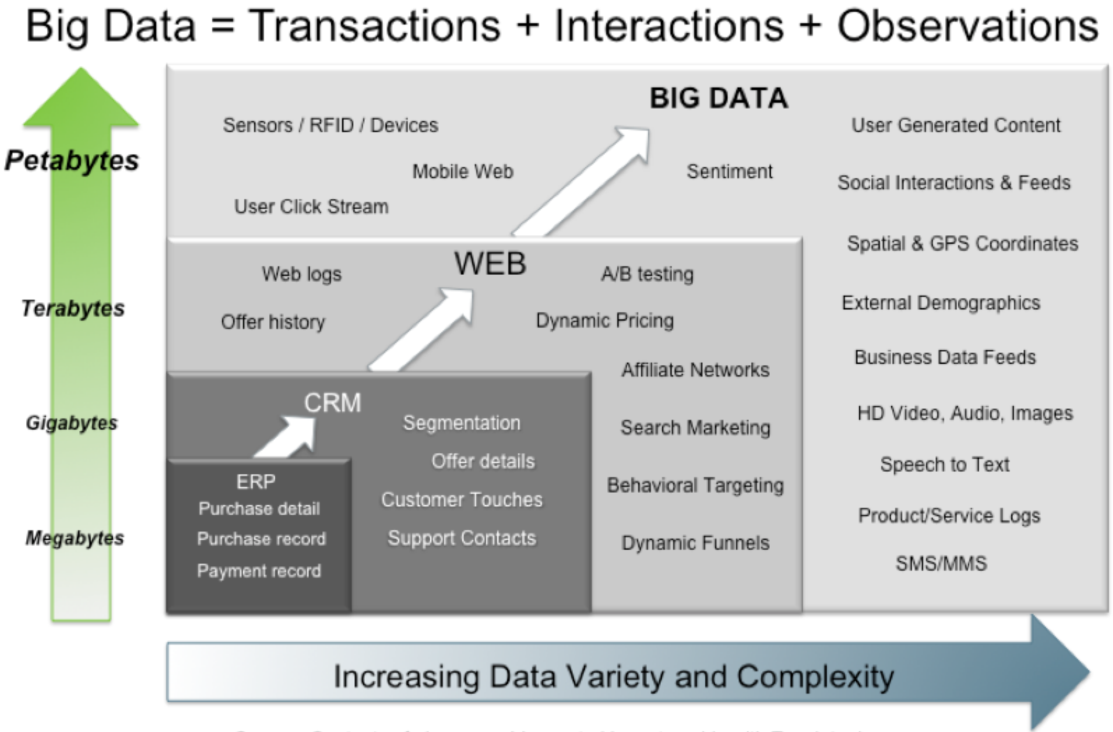
Characteristics of Big Data: 3-Speed (Velocity)

- 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

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.





Big Data = Transactions + Interactions + Observations

Petabytes

Terabytes

Gigabytes

Megabytes

BIG DATA

Sensors / RFID / Devices

Mobile Web

Sentiment

User Click Stream

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

A/B testing

Offer history

Dynamic Pricing

Affiliate Networks

Search Marketing

Behavioral Targeting

Dynamic Funnels

CRM

Segmentation

Offer details

Customer Touches

Support Contacts

ERP

Purchase detail

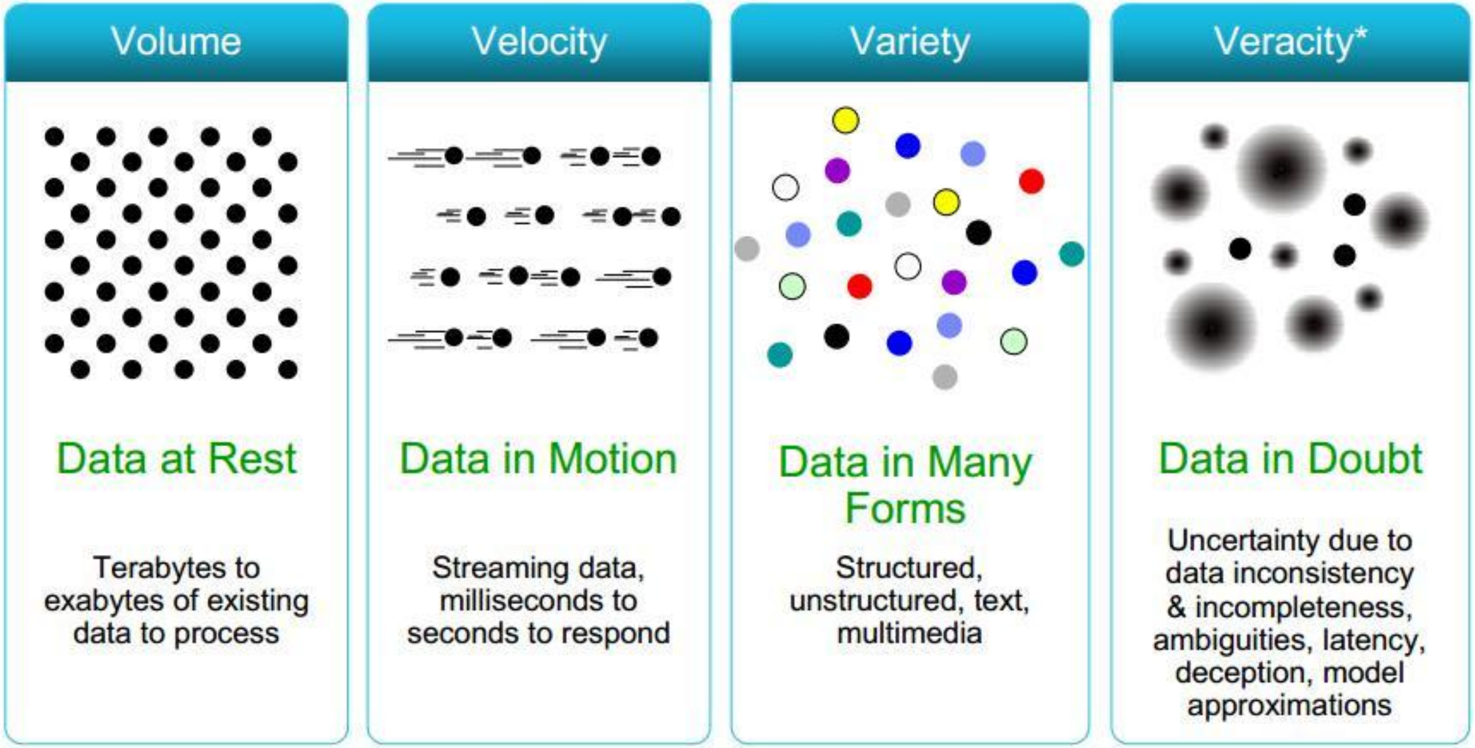
Purchase record

Payment record

Increasing Data Variety and Complexity

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

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

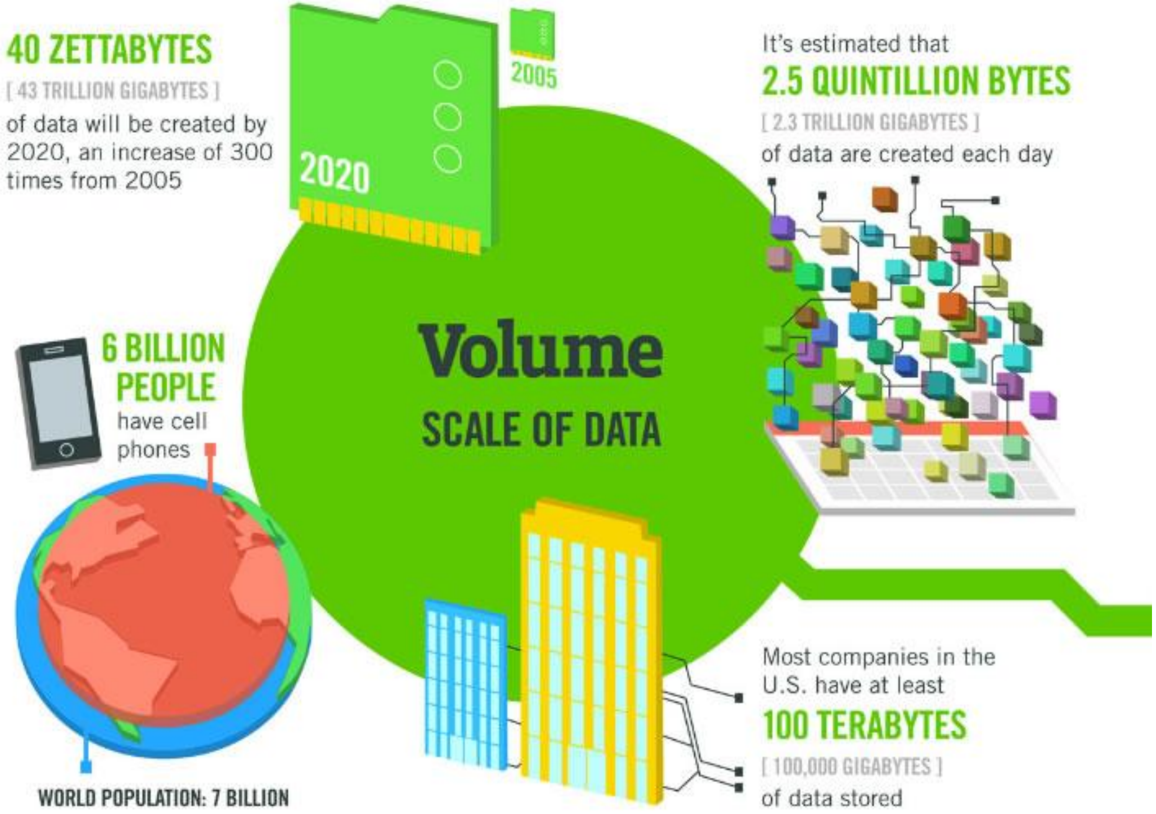
Structured, unstructured, text, multimedia

Veracity*

Data in Doubt

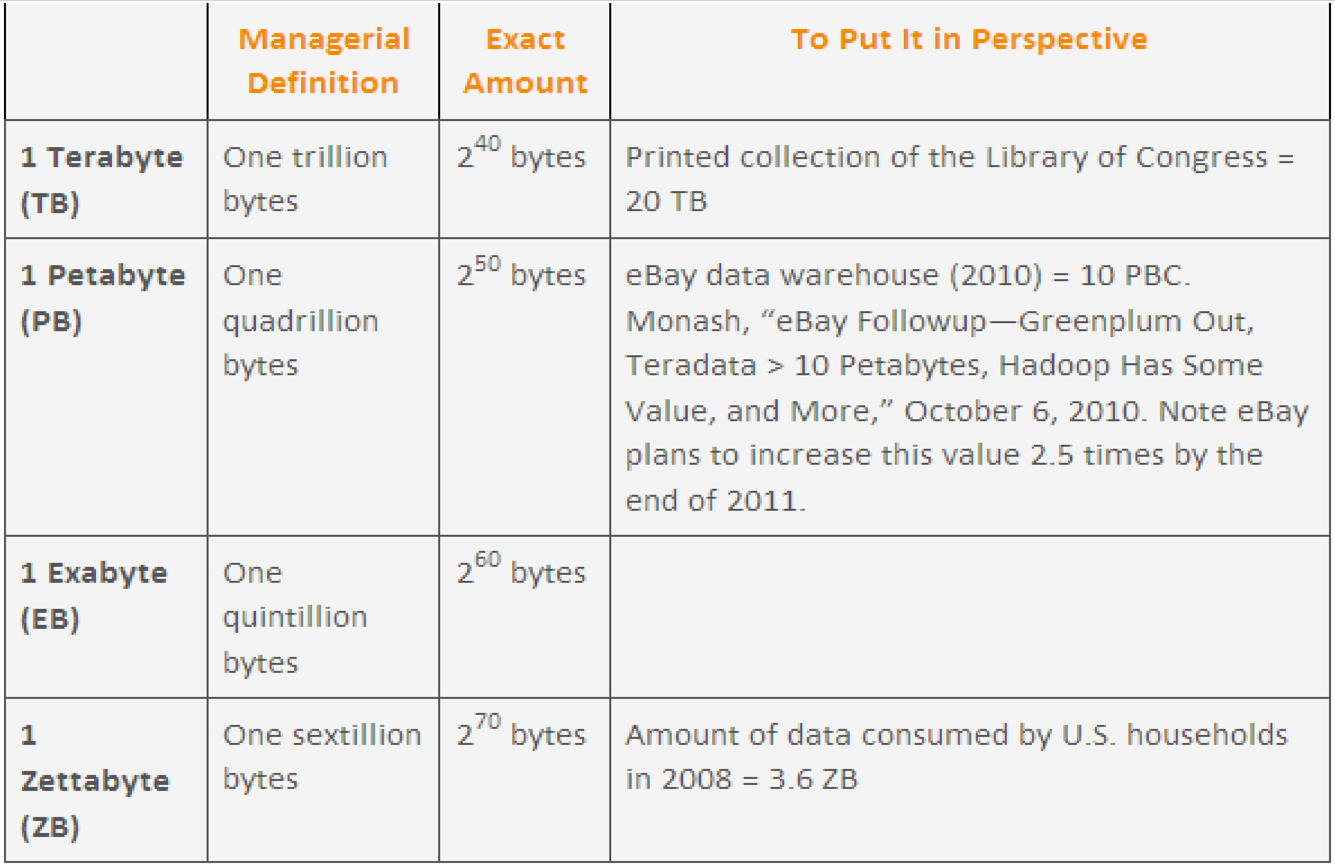
Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Volume


40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2005
2020
It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day
6 BILLION PEOPLE
have cell phones
Volume
SCALE OF DATA
WORLD POPULATION: 7 BILLION
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

- Volume
  - Petabytes, exabytes of data
  - Volumes too great for typical DBMS

Volume - Bytes Defined

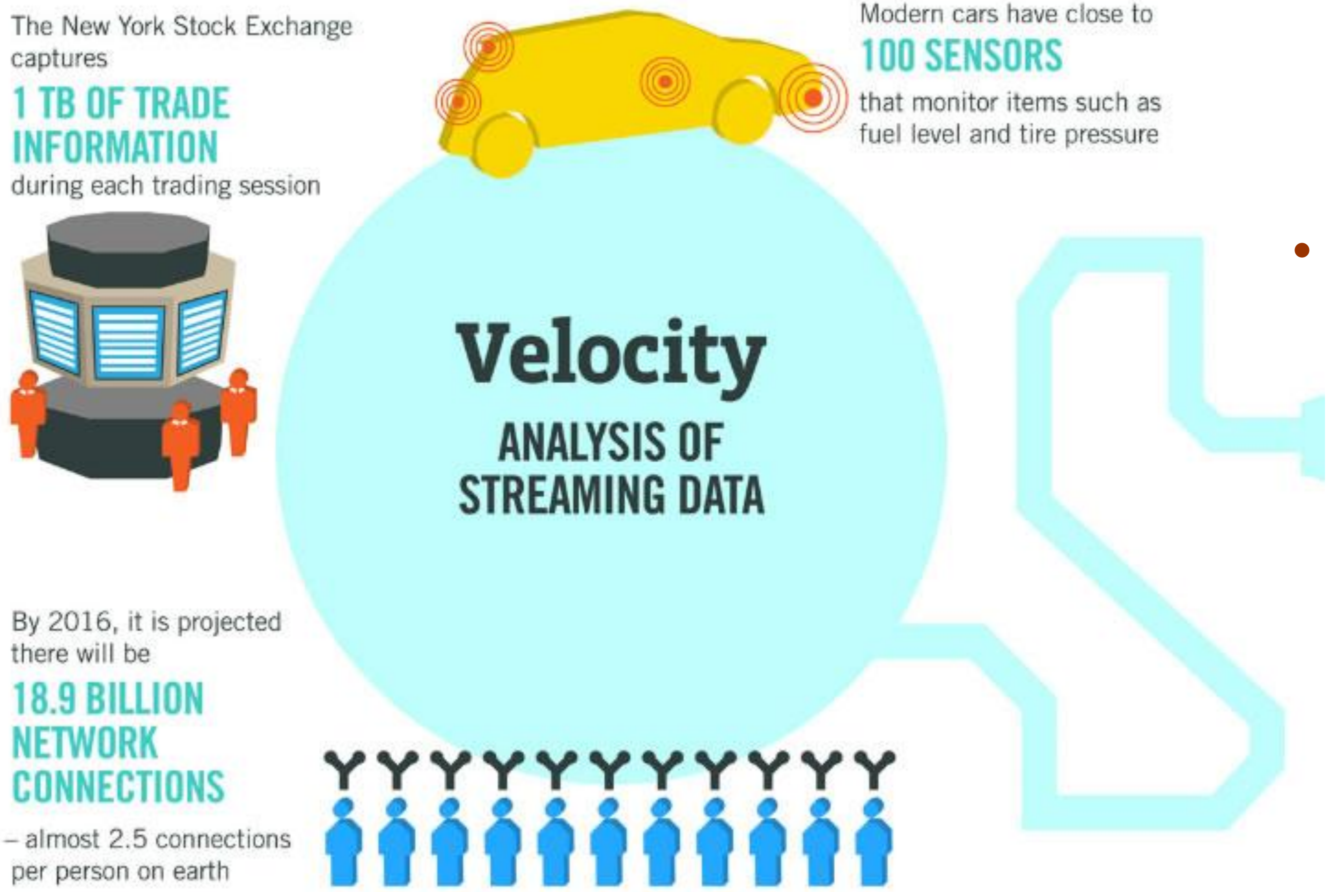


	Managerial Definition	Exact Amount	To Put It in Perspective
1 Terabyte (TB)	One trillion bytes	2^{40} bytes	Printed collection of the Library of Congress = 20 TB
1 Petabyte (PB)	One quadrillion bytes	2^{50} bytes	eBay data warehouse (2010) = 10 PBC. Monash, "eBay Followup—Greenplum Out, Teradata > 10 Petabytes, Hadoop Has Some Value, and More," October 6, 2010. Note eBay plans to increase this value 2.5 times by the end of 2011.
1 Exabyte (EB)	One quintillion bytes	2^{60} bytes	
1 Zettabyte (ZB)	One sextillion bytes	2^{70} bytes	Amount of data consumed by U.S. households in 2008 = 3.6 ZB

Megabyte: 2^{20} bytes or, loosely, one million bytes

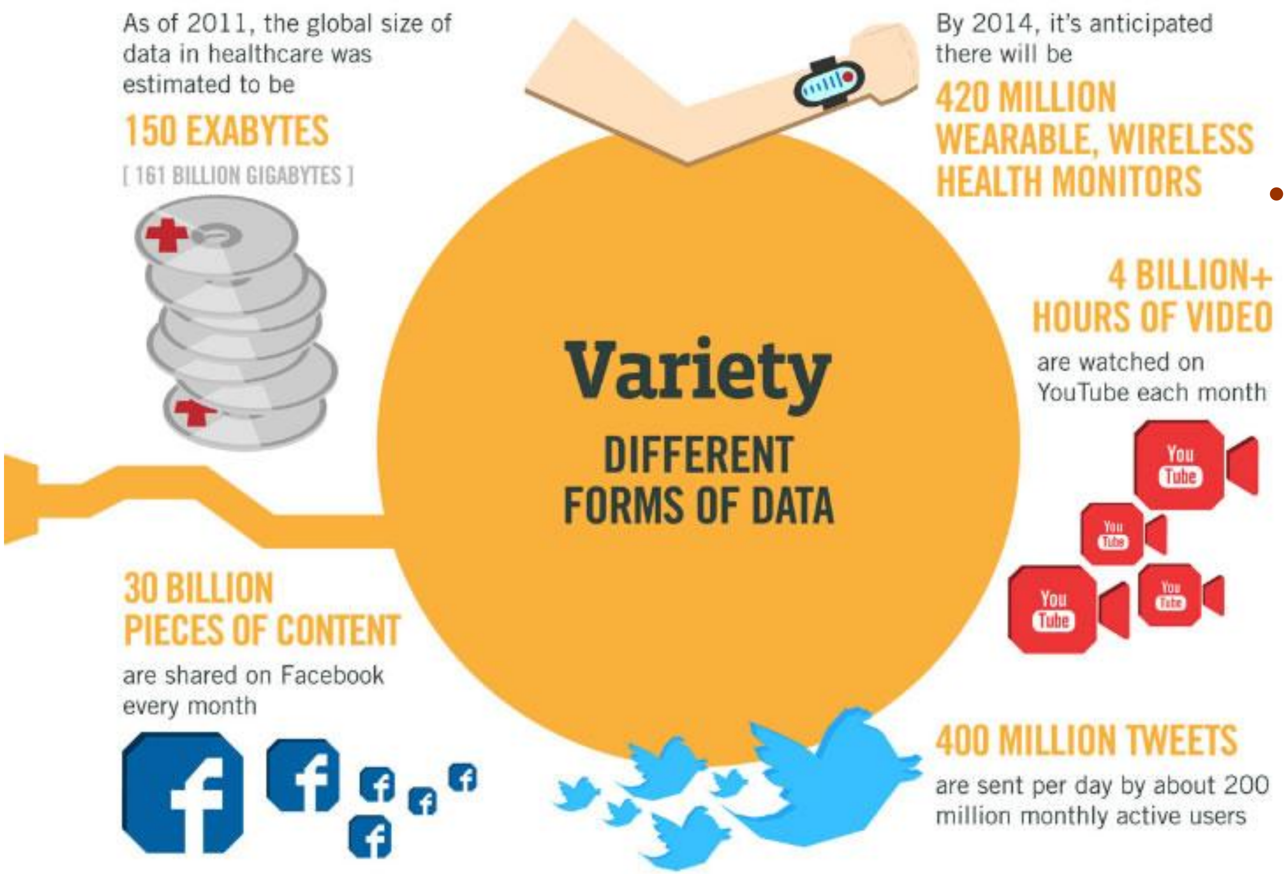
Gigabyte: 2^{30} bytes or, loosely one billion bytes

Velocity


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
Velocity
ANALYSIS OF STREAMING DATA
By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

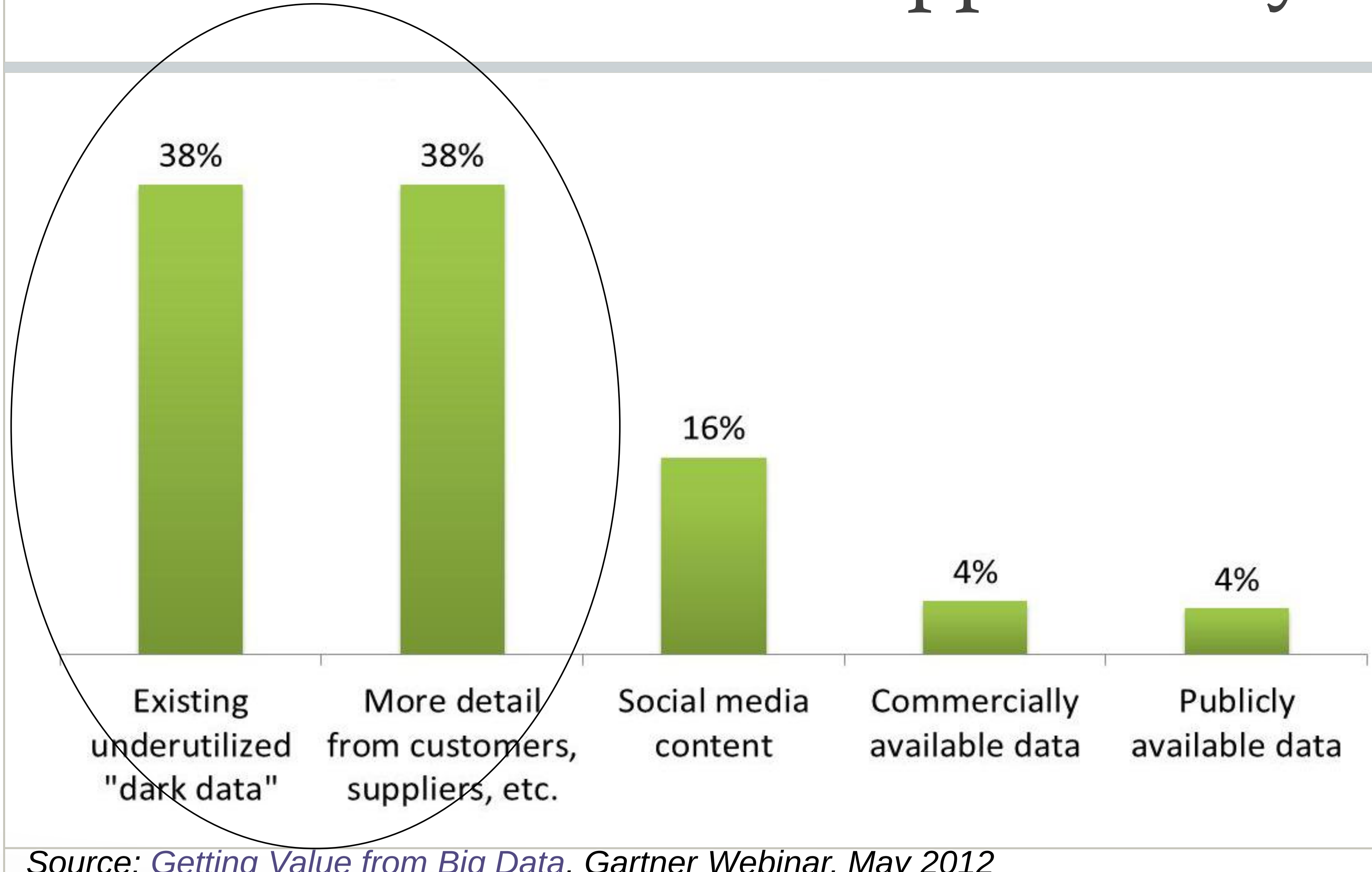
- Velocity
  - •Massive amount of streaming data

Variety


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
Variety
DIFFERENT FORMS OF DATA
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
You Tube
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

- Variety
  - Massive sets of unstructured/semi-structured data from Web traffic, social media, sensors, and so on

Which source of data represents the most immediate opportunity?



Source: Getting Value from Big Data, Gartner Webinar, May 2012

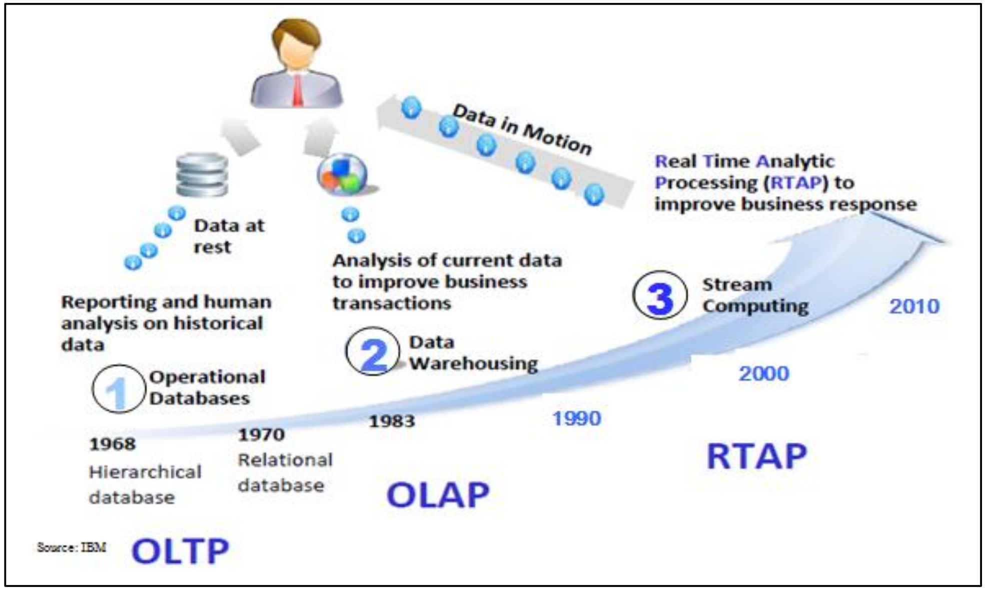
Big Data Opportunities

Making better informed decisions
e.g. strategies, recommendations

Discovering hidden insights
e.g. anomalies forensics, patterns, trends

Automating business processes
e.g. complex events, translation

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
OLTP

Source: IBM

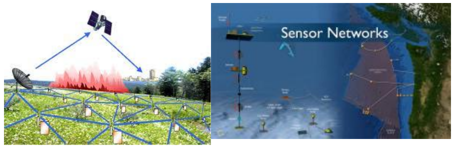
- **OLTP:** Online Transaction Processing (DBMSs)
- **OLAP:** Online Analytical Processing (Data Warehousing)
- **RTAP:** Real-Time Analytics Processing (Big Data Architecture & technology)

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 time)



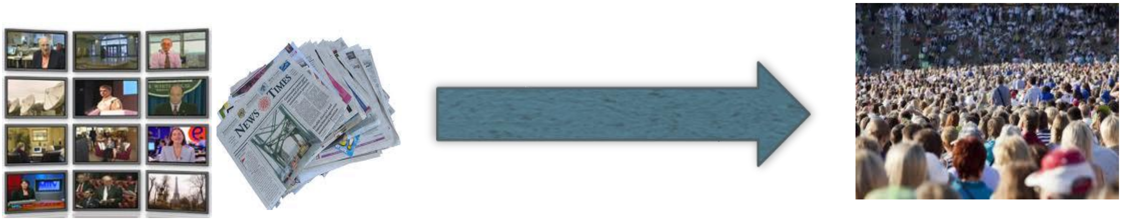
Sensor technology and networks
(measuring all kinds of data)

- 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

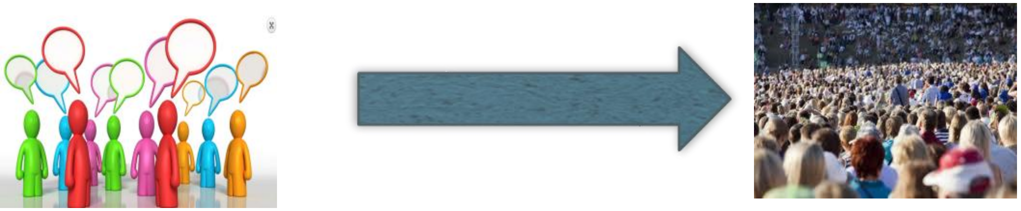
The Model Has Changed…

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

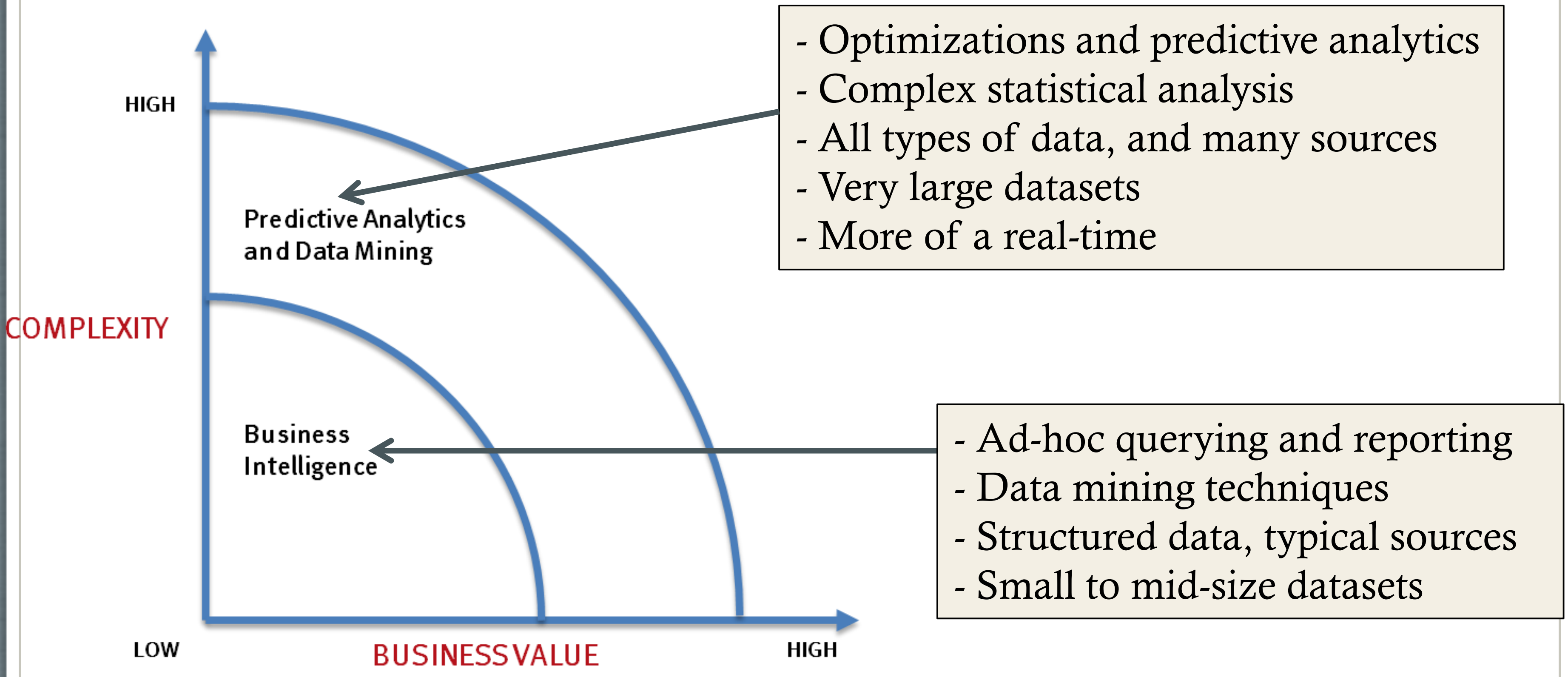
Old Model: Few companies are generating data, all others are consuming data



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



What’s driving Big Data

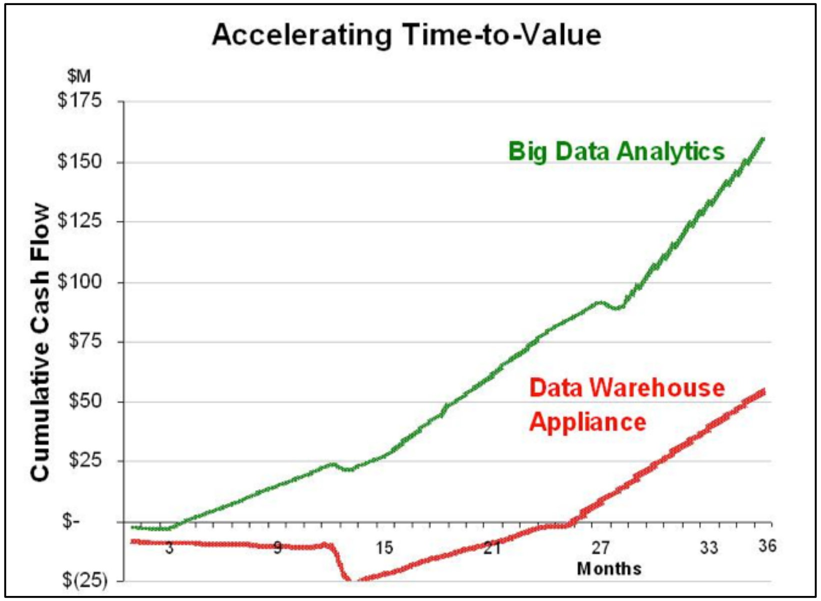


- Optimizations and predictive analytics
- Complex statistical analysis
- All types of data, and many sources
- Very large datasets
- More of a real-time

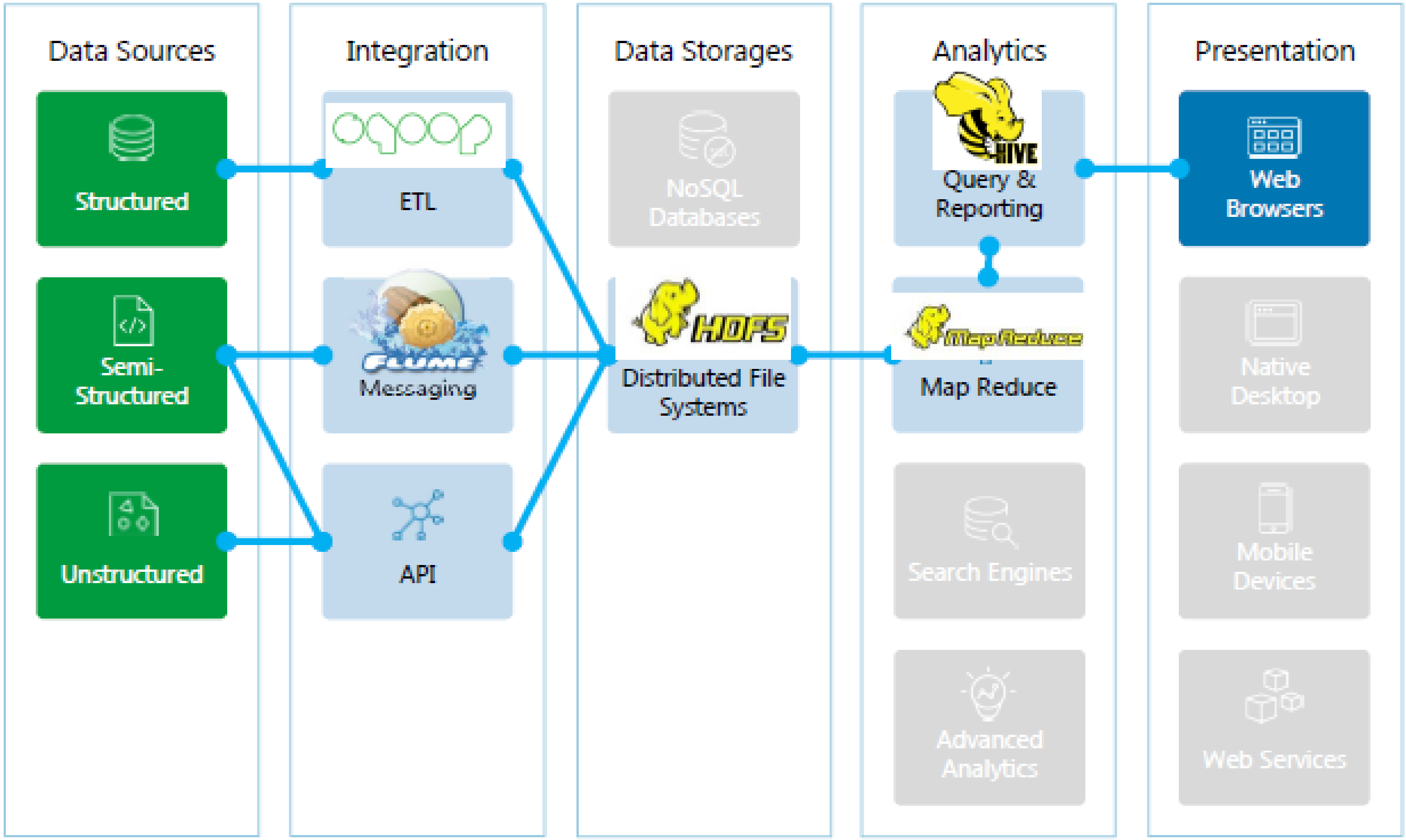
- Ad-hoc querying and reporting
- Data mining techniques
- Structured data, typical sources
- Small to mid-size datasets

Value of Big Data Analytics

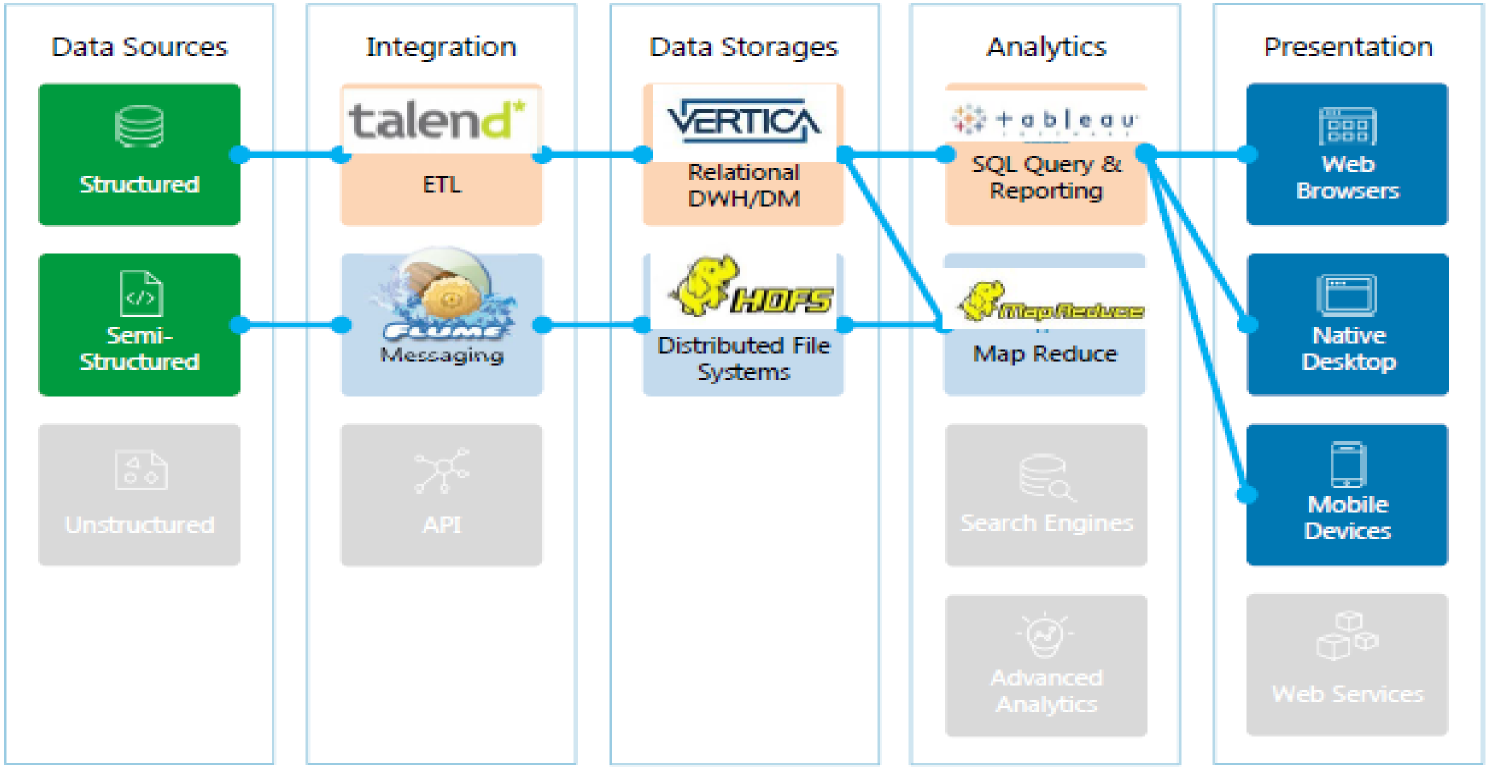
- Big data is more real-time in nature than traditional DW applications
- Traditional DW architectures (e.g. Exadata, Teradata) are not well-suited for big data apps
- Shared nothing, massively parallel processing, scale out architectures are well-suited for big data apps



Data Discovery: Non-Relational Architecture

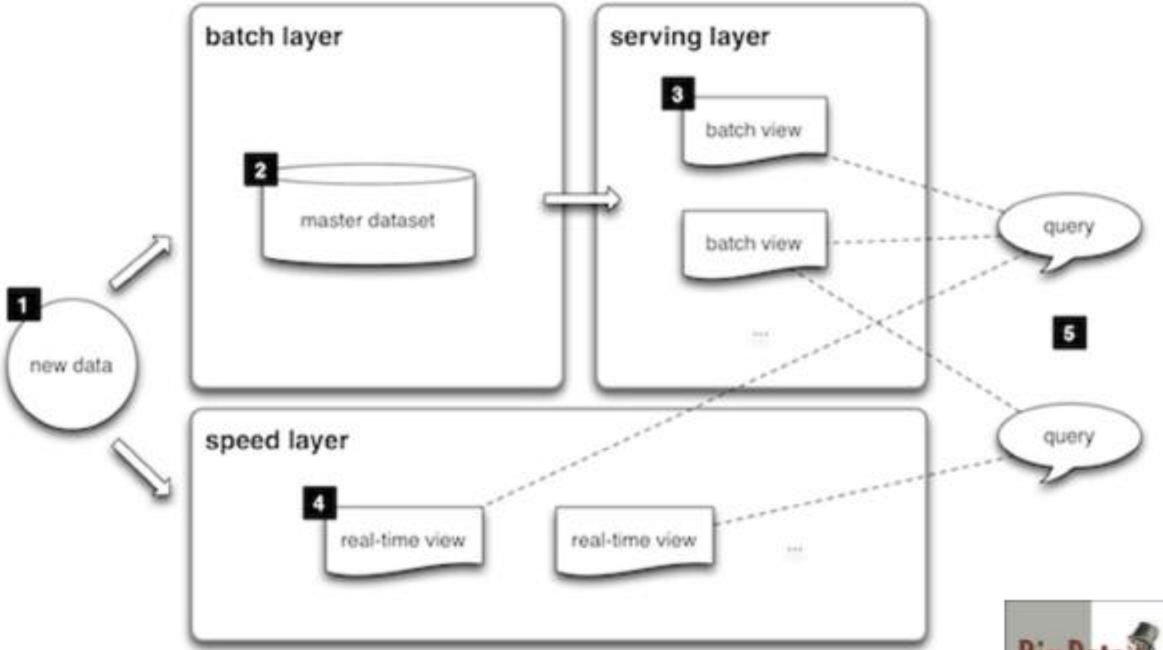


Business Reporting: Hybrid Architecture



🟧 Extended Relational components 🟦 Non-relational components

Lambda Architecture



Source:

Architectural Decisions

Architecture Drivers:

- Volume (45 TB)
- Sources (Semi-structured - JSON)
- Throughput (> 20K/sec)
- Latency (1 hour)
- Extensibility (Custom tags)
- Data Quality (Not critical)
- Reliability (24/7)
- Security (Multitenancy)
- Self-Service (Canned reports, Data science)
- Cost (The less the better ☺)
- Constraints (Public Cloud)

Trade-off:

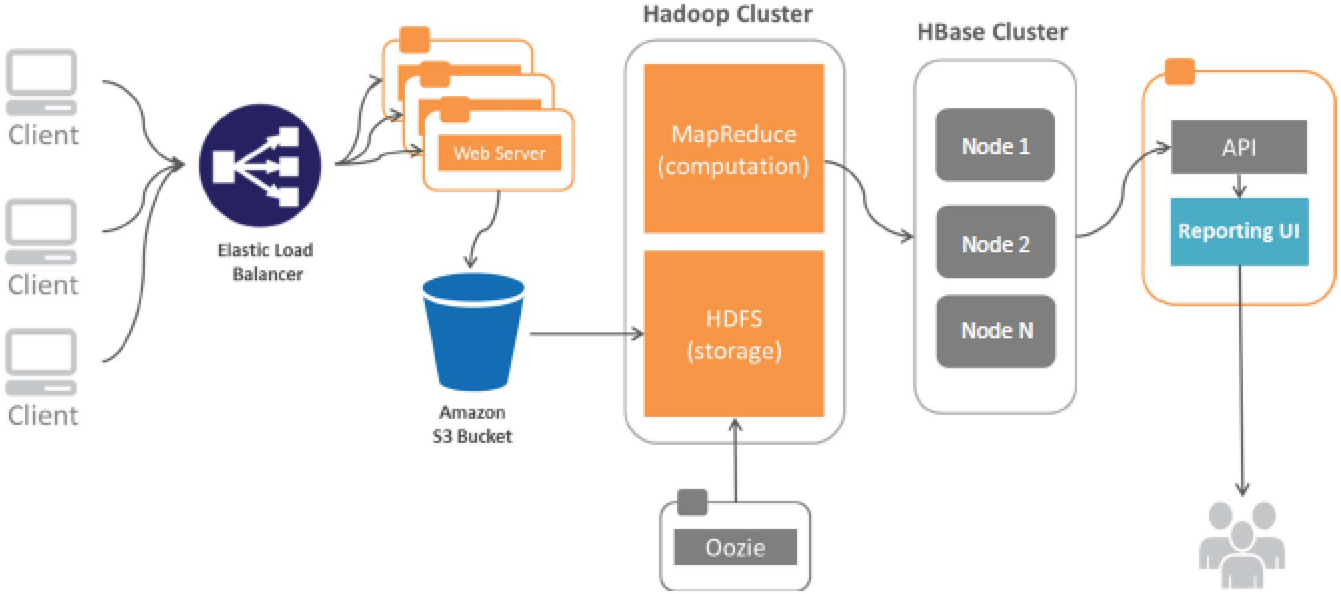
	Extended Relational	Non-Relational
Volume/Scalability	+/-	+
Throughput	+	+
Self-Service	+	+/-
Extensibility	-	+

- ✓ **Non-Relational Architecture**
- ✓ Reporting via **Materialized View** pattern

Solution Architecture

Technologies:

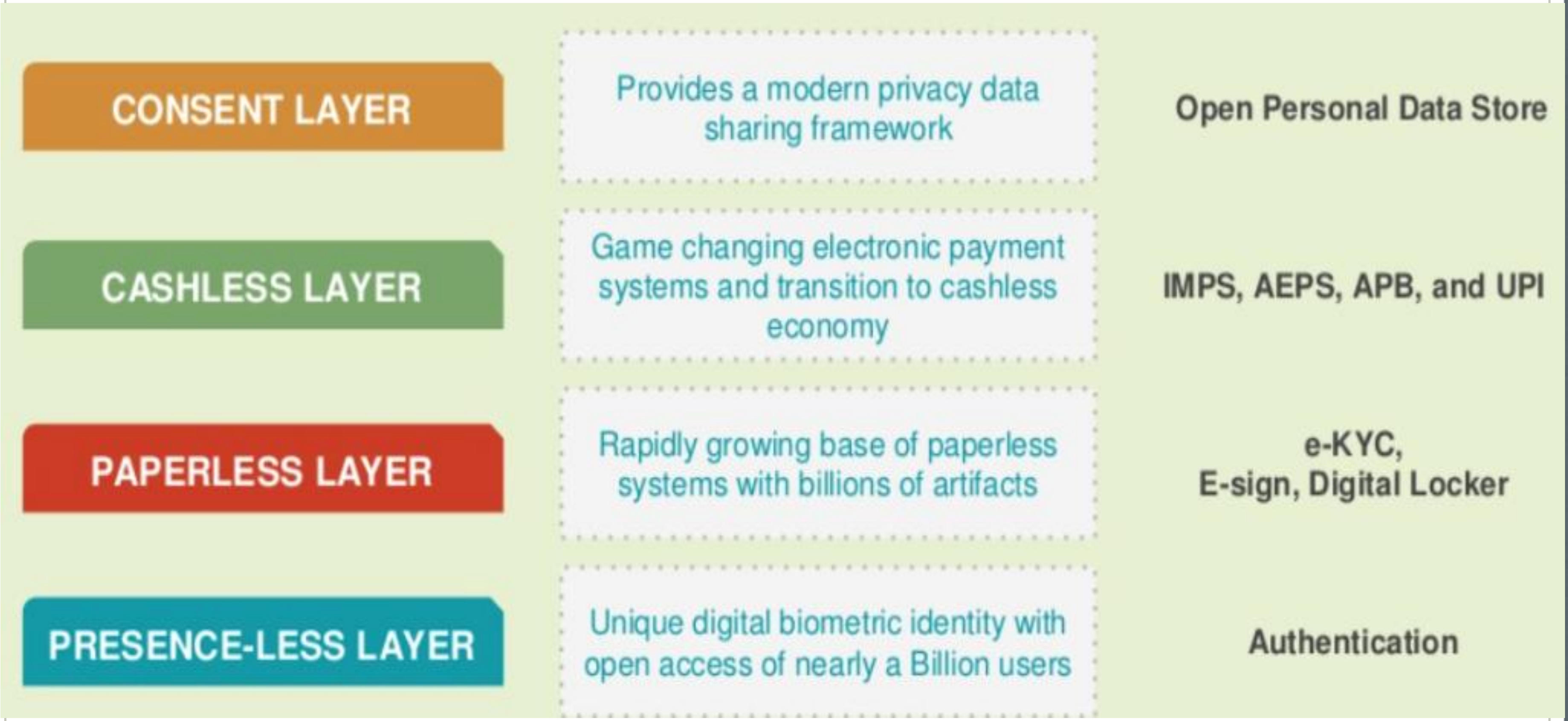
- Amazon S3
- Flume
- Hadoop/HDFS, MapReduce
- HBase
- Oozie
- Hive



Tips for Designing Big Data Solutions

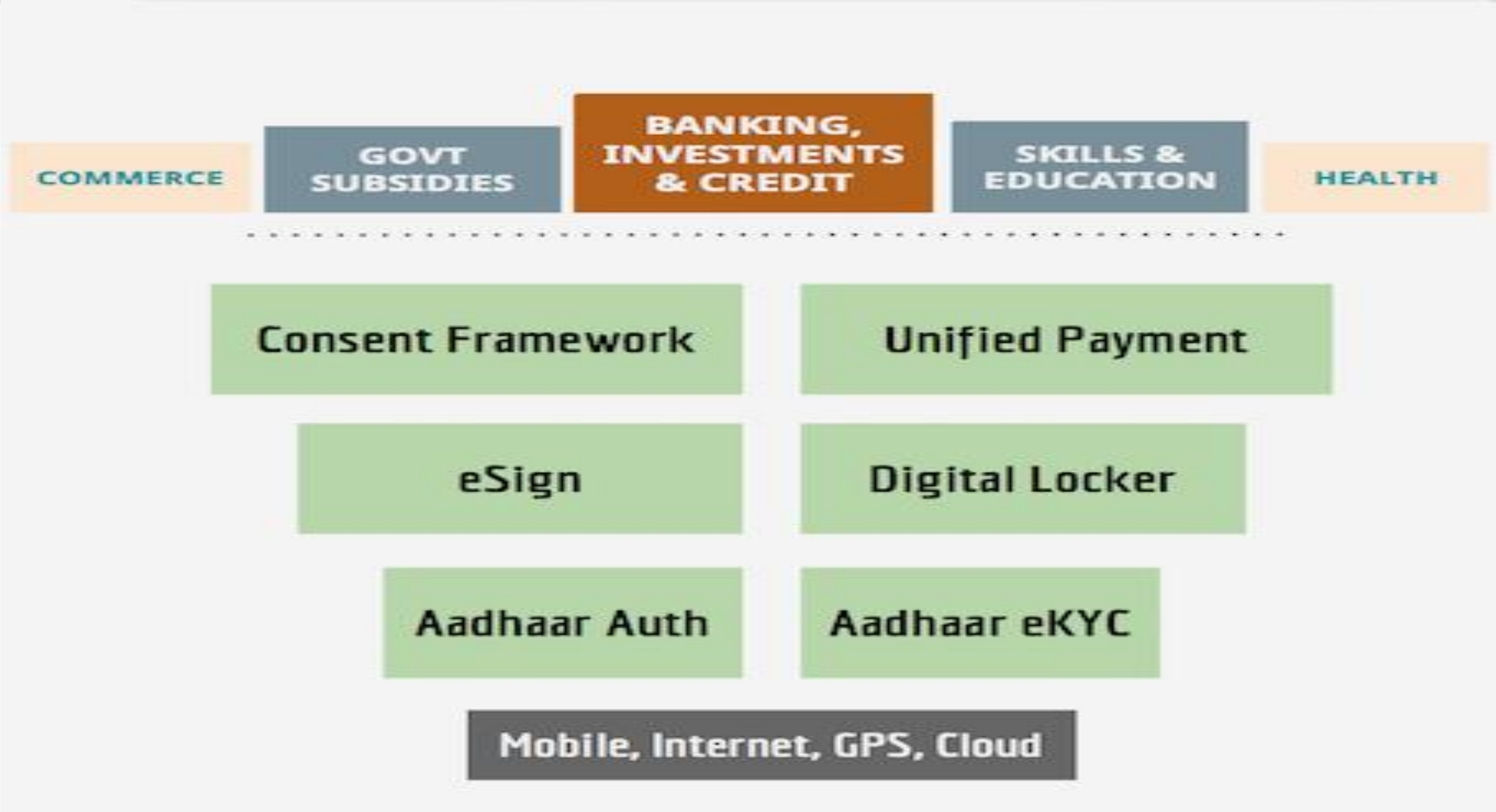
- Understand data users and sources
- Discover architecture drivers
- Select proper reference architecture
- Do trade-off analysis, address cons
- Map reference architecture to technology stack
- Prototype, re-evaluate architecture
- Estimate implementation efforts
- Set up devops practices from the very beginning
- Advance in solution development through “small wins”
- Be ready for changes, big data technologies are evolving rapidly

India Stack



CONSENT LAYER	Provides a modern privacy data sharing framework	Open Personal Data Store
CASHLESS LAYER	Game changing electronic payment systems and transition to cashless economy	IMPS, AEPS, APB, and UPI
PAPERLESS LAYER	Rapidly growing base of paperless systems with billions of artifacts	e-KYC, E-sign, Digital Locker
PRESENCE-LESS LAYER	Unique digital biometric identity with open access of nearly a Billion users	Authentication

India Stack



COMMERCE

GOVT SUBSIDIES

BANKING, INVESTMENTS & CREDIT

SKILLS & EDUCATION

HEALTH

Consent Framework

Unified Payment

eSign

Digital Locker

Aadhaar Auth

Aadhaar eKYC

Mobile, Internet, GPS, Cloud

Identifying Insurance Fraud

Auto Insurance

INFINITY

- Opportunity
  - Save and make money by reducing fraudulent auto insurance claims
- Data & Analytics
  - Predictive analytics against years of historical claims and coverage data
  - Text mining adjuster reports for hidden clues, e.g. missing facts, inconsistencies, changed stories
- Results
  - Improved success rate in pursuing fraudulent claims from 50% to 88%; reduced fraudulent claim investigation time by 95%
  - Marketing to individuals with low propensity for fraud

Identifying Insurance Fraud

Auto Insurance

INFINITY

What **“dark data” is just laying around that can transform business processes?

Operational **data that is not being used. Consulting and market research company Gartner Inc. describes **dark data** as "information assets that organizations collect, process and store in the course of their regular business activity, but generally fail to use for other purposes."

Quality Improvemen

- Opportunity
  - Move from manual to automated inspection of burger bu production to ensure and improve quality
- Data & Analytics
  - Photo-analyze over 1000 buns-per-minute for color, shape and seed distribution
  - Continually adjust ovens and process automatically

Quality Improvemen

- Result
  - Eliminate 1000s of pounds of wasted product per year; speed production; save energy; Reduce manual labor costs

Is the company using all of its “senses” to observe, measure and optimize business processes?

Improving Corporate Image

MORTON'S
THE STEAKHOUSE

- Opportunity
  - Improve reputation, brand and buzz by tapping social media
- Data & Analytics
  - Continually scanning twitterverse for mentions of their business
  - Integrating tweeters with their robust customer management system

Improving Corporate Image

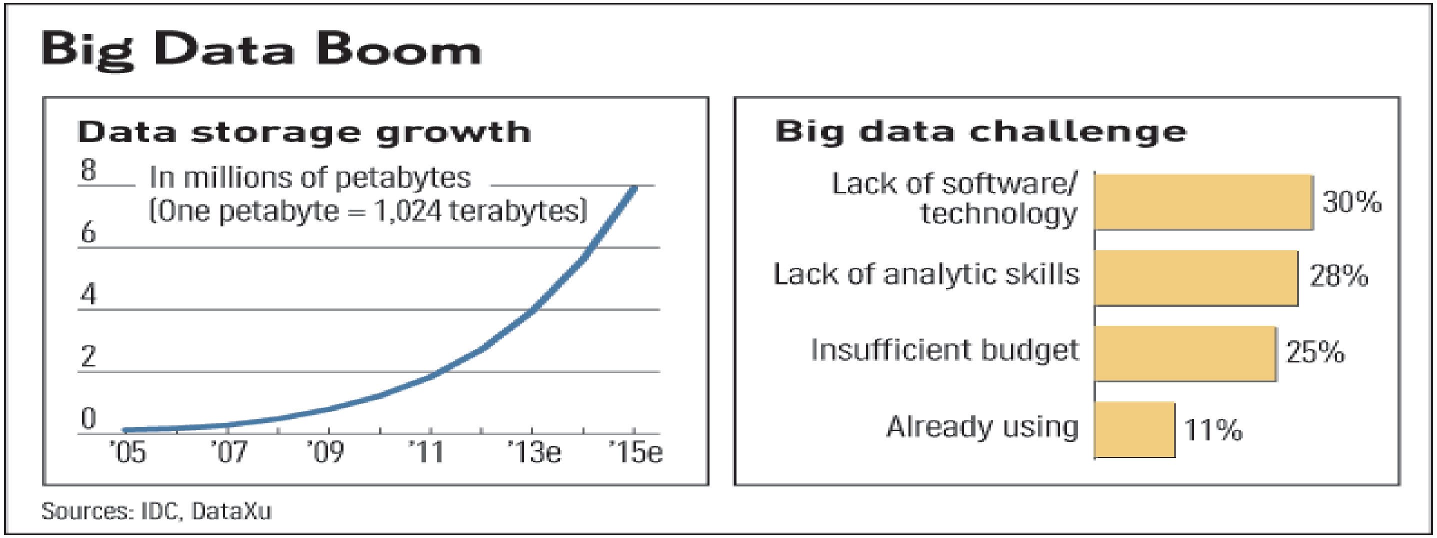
MORTON'S

THE STEAKHOUSE

- Results
  - Saw tweet from a top customer lamenting late flight—no time to dine at Morton’s
  - Tuxedo-clad waiter waiting for him when he landed with a bag containing his favorite steak, prepared the way he normally likes it with all the fixin’s

How can the company listen, analyze and respond in real-time?

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

What Technology Do We Have For Big Data ??

Data Mining Techniques…

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]

Big Data Landscape

Vertical Apps

- PREDICTIVE POLICING
- bloomreach GET FOUND.
- MYRRIX

Log Data Apps

- splunk>
- loggly
- sumologic

Ad/Media Apps

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

Business Intelligence

- ORACLE | Hyperion
- SAP Business Objects
- RJMetrics
- Microsoft | Business Intelligence
- IBM
- COGNOS
- birst
- MicroStrategy
- Autonomy
- QlikView
- bime
- DOMO
- 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
- ClearStory DATA
- CIRRO
- platfora
- alteryx
- visual.ly
- AYATA

Data As A Service

- factual.
- kaggle
- knoema beta
- GNIP
- DATASIFT
- Windows Azure Marketplace
- INRIX
- LexisNexis
- SPACE CURVE
- LOCATE Everything Location

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

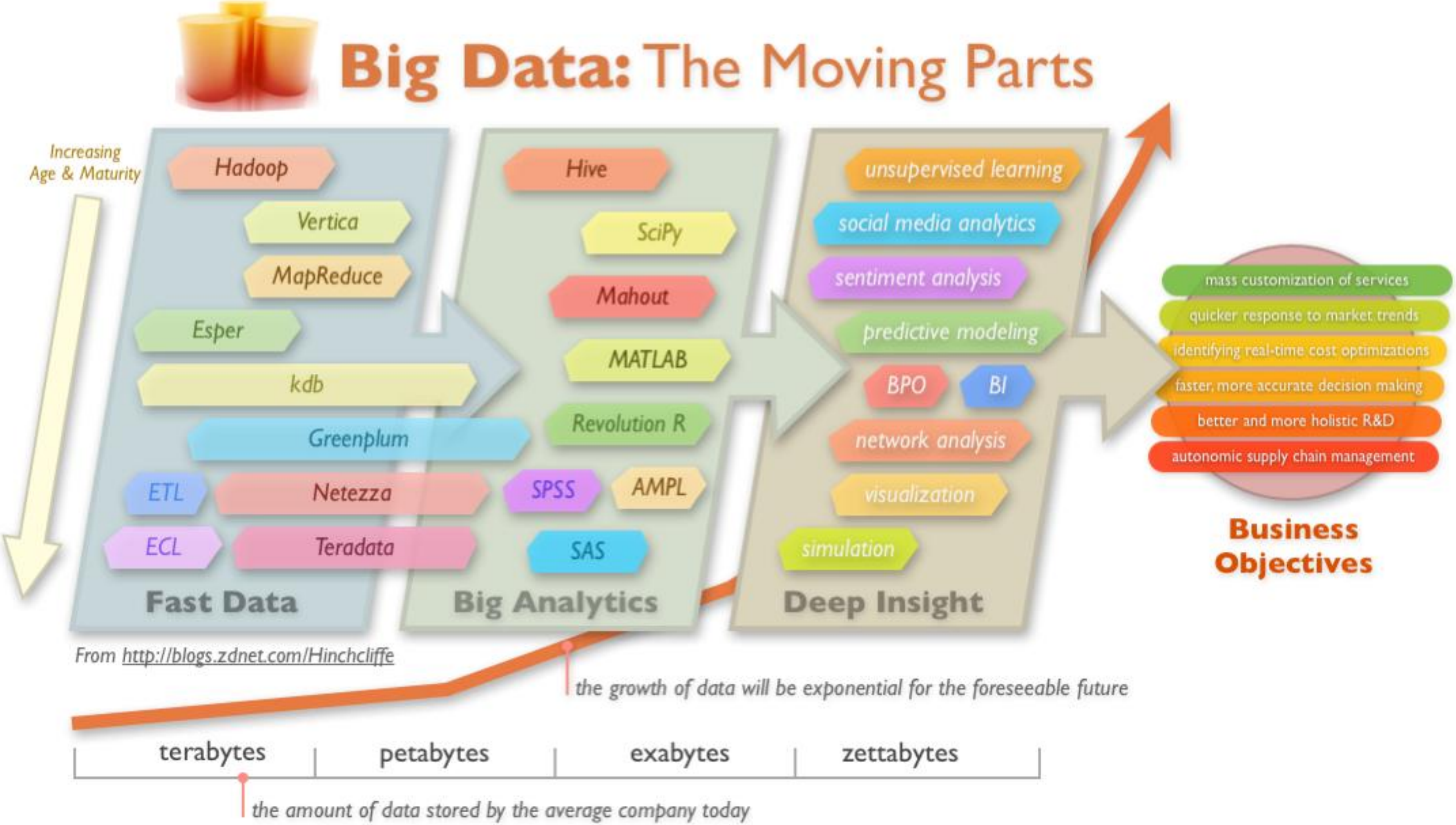
Structured Databases

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

Technologies

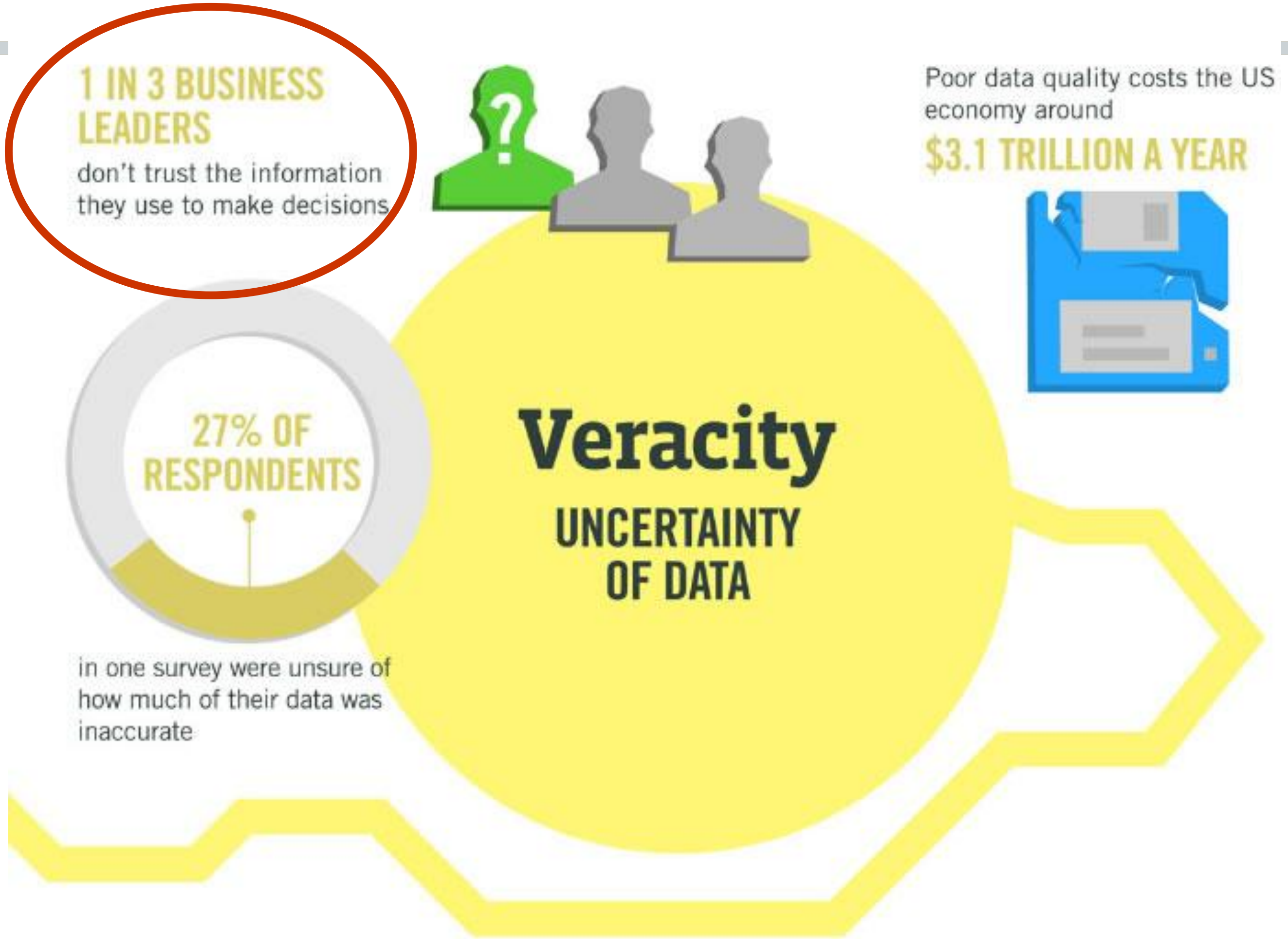
- hadoop
- hadoop Map Reduce
- mahout
- APACHE HBASE
- Cassandra

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

Uncertainty of Data



1 IN 3 BUSINESS LEADERS

don't trust the information they use to make decisions

27% OF RESPONDENTS

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

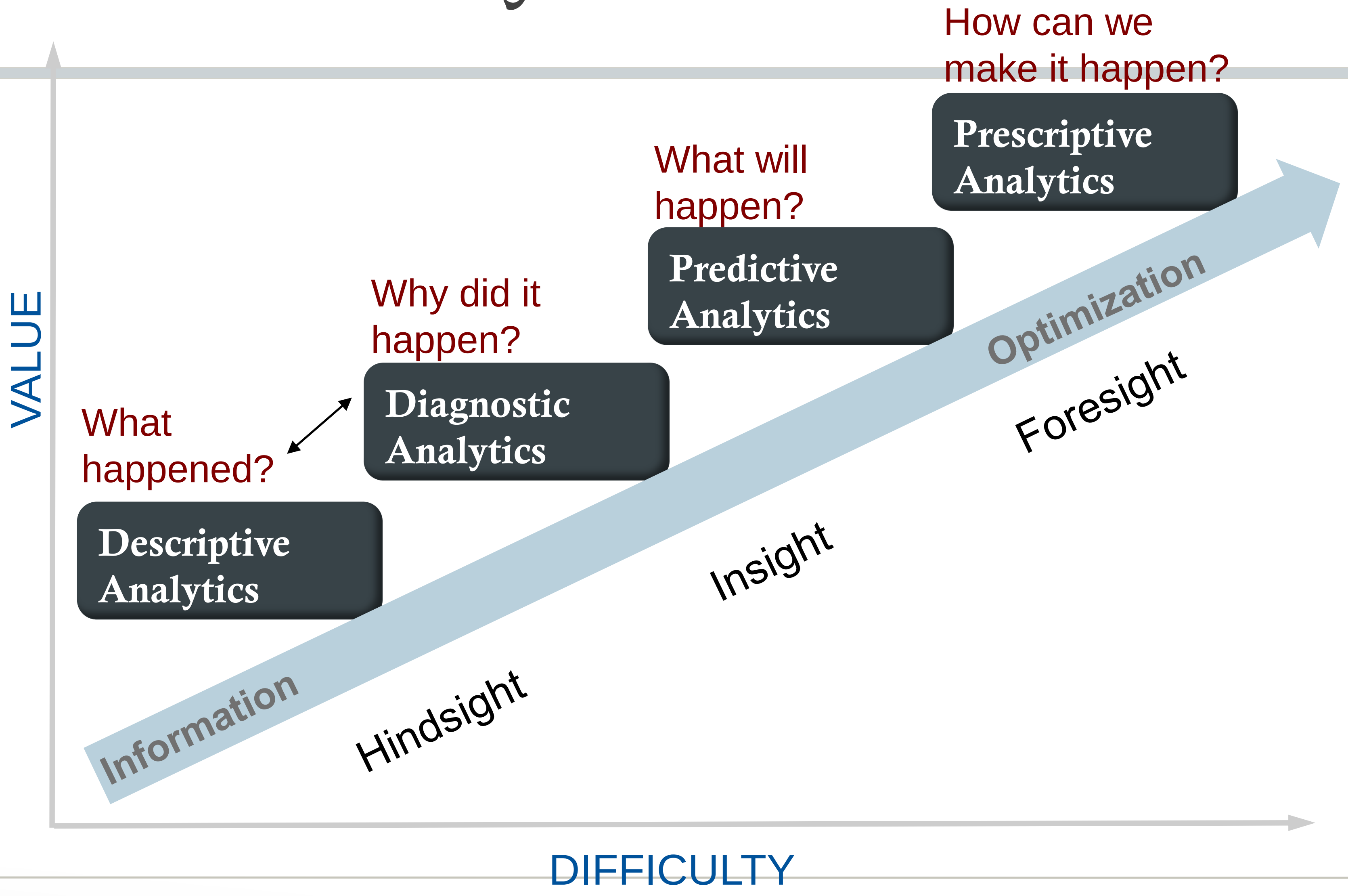
Veracity

UNCERTAINTY OF DATA

Poor data quality costs the US economy around

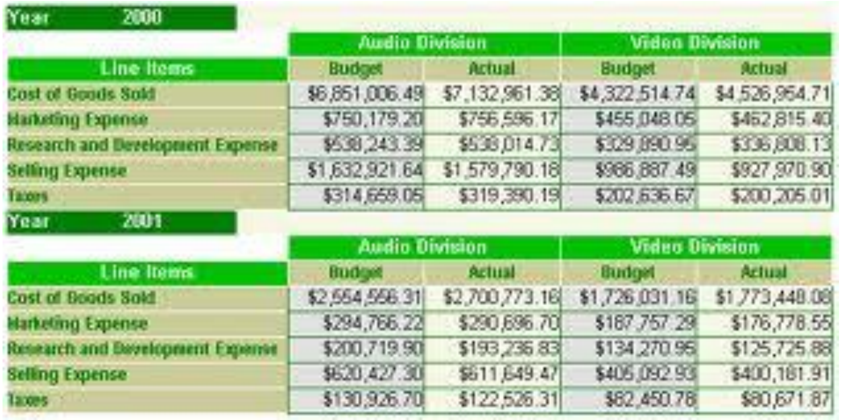
$3.1 TRILLION A YEAR

Analytics Models



Descriptive Analytics

- Descriptive analytics, such as reporting/OLAP, dashboards, and data visualization, have been widely used for some time.
- They are the core of traditional BI.

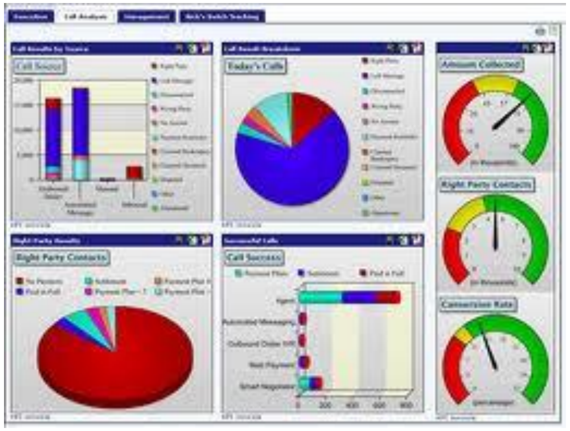


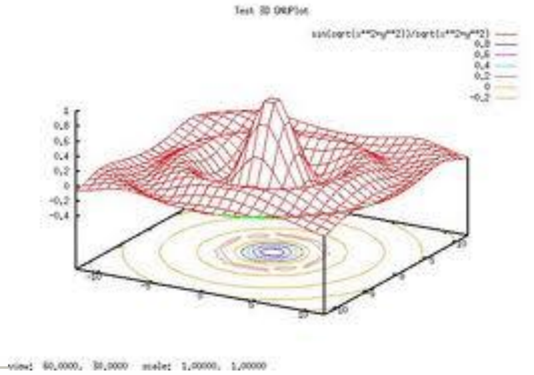
Year 2000

Line Items	Audio Division		Video Division	
	Budget	Actual	Budget	Actual
Cost of Goods Sold	$6,851,006.49	$7,132,961.38	$4,322,514.74	$4,526,954.71
Marketing Expense	$750,179.20	$756,596.17	$455,048.05	$462,815.40
Research and Development Expense	$538,243.39	$538,014.73	$329,890.95	$336,808.13
Selling Expense	$1,632,921.64	$1,579,790.18	$986,887.49	$927,970.90
Taxes	$314,659.05	$319,390.19	$202,636.67	$200,205.01

Year 2001

Line Items	Audio Division		Video Division	
	Budget	Actual	Budget	Actual
Cost of Goods Sold	$2,554,556.31	$2,700,773.16	$1,726,031.16	$1,773,448.08
Marketing Expense	$294,766.22	$290,696.70	$187,757.29	$176,778.55
Research and Development Expense	$200,719.90	$193,236.83	$134,270.95	$125,725.88
Selling Expense	$620,427.30	$611,649.47	$405,092.93	$400,161.91
Taxes	$130,926.70	$122,526.31	$82,450.78	$80,671.87



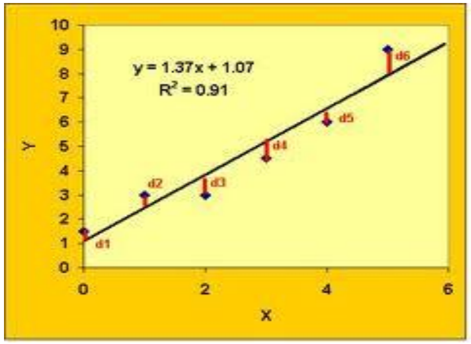

Test 3D GNUPlot
sin(sqrt(x**2+y**2))/sqrt(x**2+y**2)
0.8
0.6
0.4
0.2
0
-0.2
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
view: 60.0000, 30.0000 scale: 1.00000, 1.00000

What has occurred?

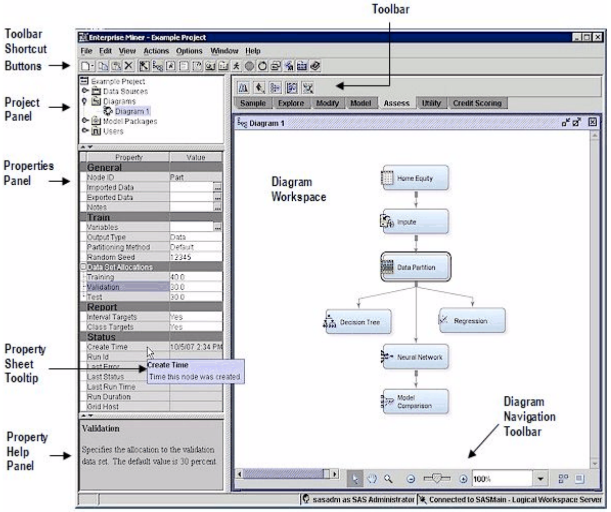
Descriptive analytics, such as data visualization, is important in helping users interpret the output from predictive and predictive analytics.

Predictive Analytics

- Algorithms for predictive analytics, such as regression analysis, machine learning, and neural networks, have also been around for some time.



What will occur?



Property	Value
General	
Node ID	Part
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	40.0
Validation	30.0
Test	30.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	10/5/07 2:34 PM
Run Id	
Last Error	
Last Status	
Last Run Time	
Run Duration	
Grid Host	

Create Time
Time this node was created

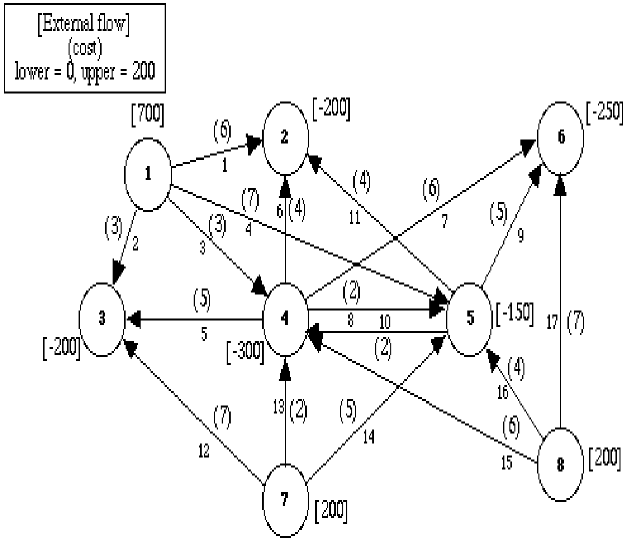
Validation

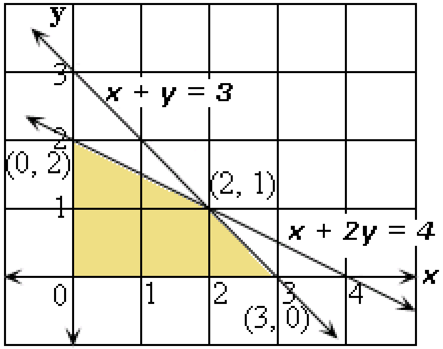
Specifies the allocation to the validation data set. The default value is 30 percent.

- Marketing is the target for many predictive analytics applications.
- Descriptive analytics, such as data visualization, is important in helping users interpret the output from predictive and prescriptive analytics.

Prescriptive Analytics

- Prescriptive analytics are often referred to as advanced analytics.
- Often for the allocation of scarce resources
- Optimization





What should occur?

Prescriptive analytics can benefit healthcare strategic planning by using analytics to leverage operational and usage data combined with data of external factors such as economic data, population demographic trends and population health trends, to more accurately plan for future capital investments such as new facilities and equipment utilization as well as understand the trade-offs between adding additional beds and expanding an existing facility versus building a new one.

What You Will Learn…

- We focus on ***Hadoop/MapReduce technology***
- **Learn the platform (how it is designed and works)**
  - How big data are managed in a scalable, efficient way
- **Learn writing Hadoop jobs in different languages**
  - Programming Languages: Java, C, Python
  - High-Level Languages: Apache Pig, Hive
- **Learn advanced analytics tools on top of Hadoop**
  - RHadoop: Statistical tools for managing big data
  - Mahout: Data mining and machine learning tools over big data
- **Learn state-of-art technology from recent research papers**
  - Optimizations, indexing techniques, and other extensions to Hadoop

Course Output: What You Will Learn…

- We focus on ***Hadoop/MapReduce technology***
- **Learn the platform (how it is designed and works)**
  - How big data are managed in a scalable, efficient way
- **Learn writing Hadoop jobs in different languages**
  - Programming Languages: Java, C, Python
  - High-Level Languages: Apache Pig, Hive
- **Learn advanced analytics tools on top of Hadoop**
  - RHadoop: Statistical tools for managing big data
  - Mahout: Analytics and data mining tools over big data
- **Learn state-of-art technology from recent research papers**
  - Optimizations, indexing techniques, and other extensions to Hadoop