Big Data and Data Analytics

Dr. M. Durairaj
Associate Professor in Computer Science
School of Computer Science, Engineering and Applications,
Bharathidasan University

CONTENTS

- **Module 1:** Big Data
- **Module 2:** Business Intelligence/Analytics
- **Module 3:** Visualization
- **Module 4:** Data Mining

MODULE 1

What is Big Data?

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 …….

- **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? continued

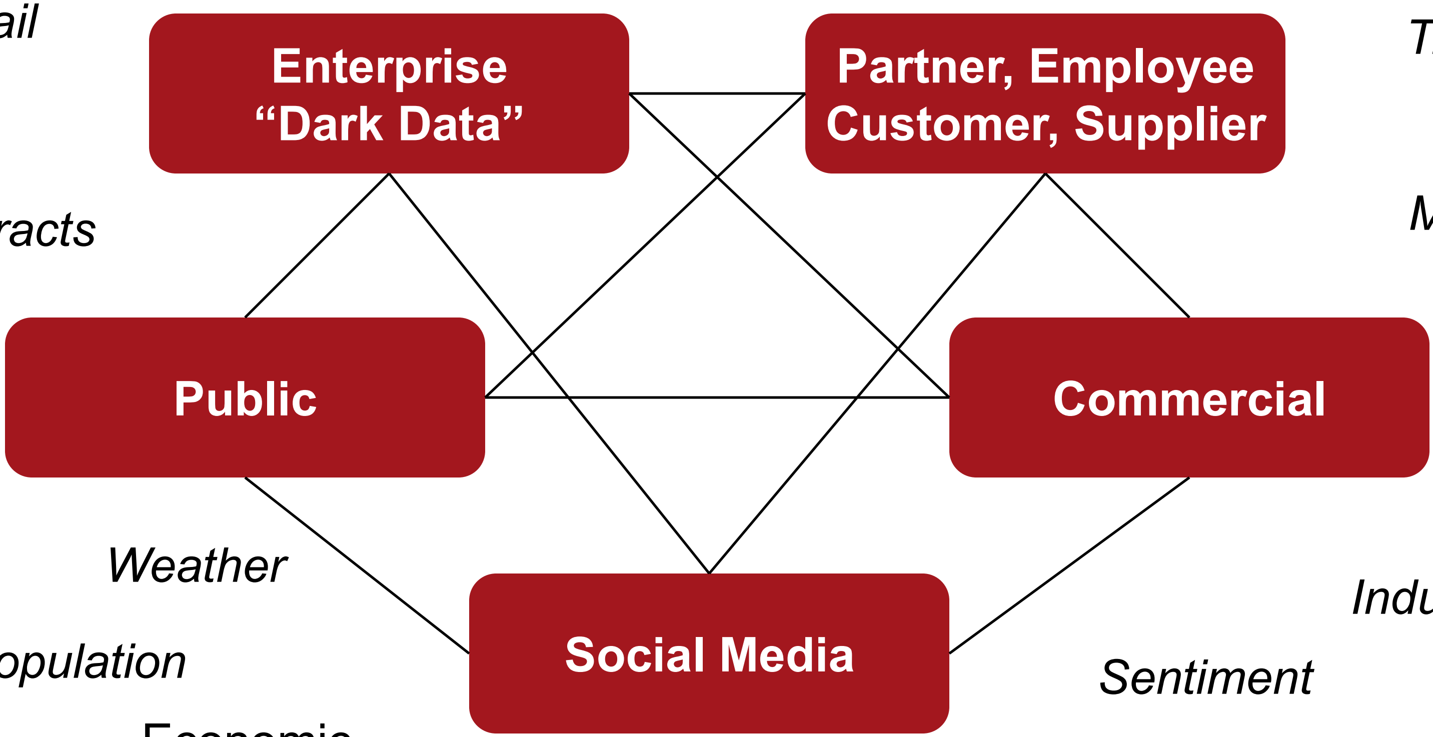
- Companies leverage data to adapt products and services to:
  - Meet customer needs
  - Optimize operations
  - Optimize infrastructure
  - Find new sources of revenue
  - Can reveal more patterns and anomalies

- 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

Where does Big Data come from?

Email

Transactions



Enterprise
"Dark Data"

Partner, Employee
Customer, Supplier

Public

Commercial

Social Media

Contracts

Monitoring

Credit

Sensor

Weather

Industry

Population

Sentiment

Economic

Network

Types of Data

The 4 V's of Big Data

Volume

The amount of data

Variety

The types of data

Velocity

The frequency of data

Veracity

The quality of data

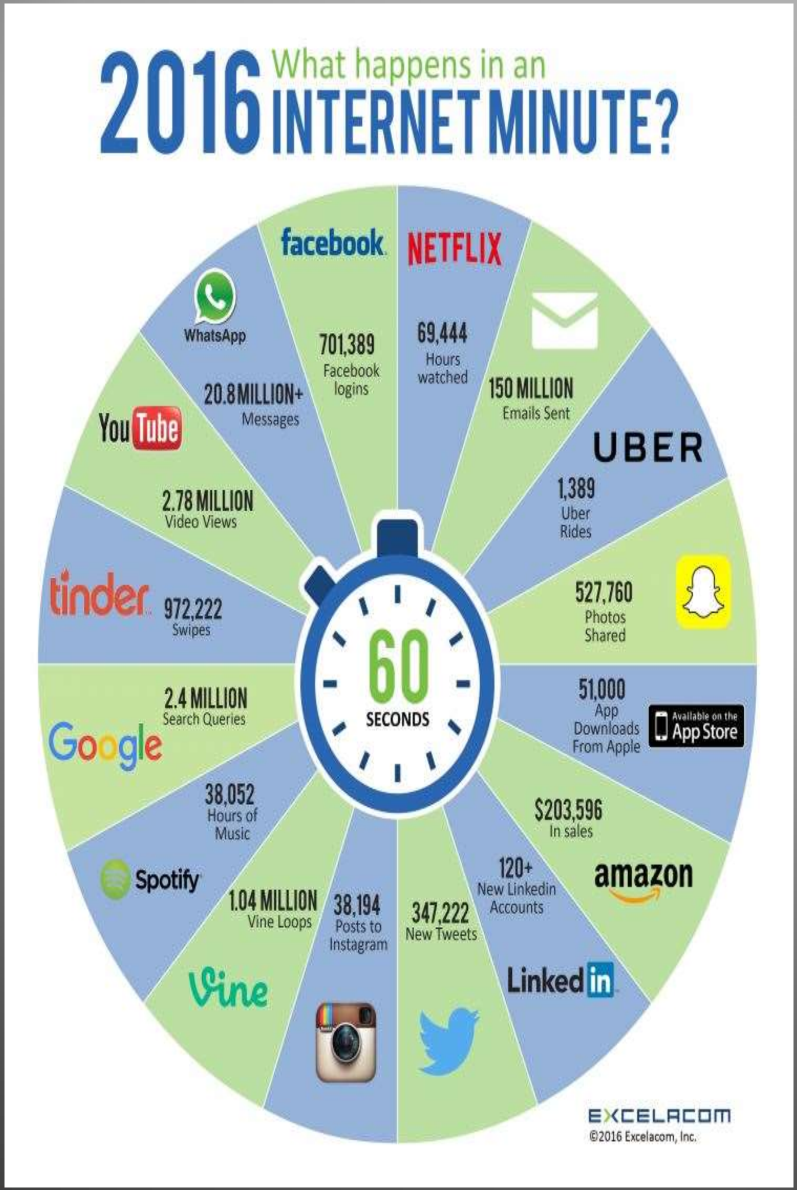
Volume: scale of data

Unit	Value	Size
bit (b)	0 or 1	1/8 of a byte
byte (B)	8 bits	1 byte
kilobyte (KB)	1000^1 bytes	1,000 bytes
megabyte (MB)	1000^2 bytes	1,000,000 bytes
gigabyte (GB)	1000^3 bytes	1,000,000.000 bytes
terabyte (TB)	1000^4 bytes	1,000,000,000,000 bytes
petabyte (PB)	1000^5 bytes	1,000,000,000,000,000 bytes
exabyte (EB)	1000^6 bytes	1,000,000,000,000,000,000 bytes
zettabyte (ZB)	1000^7 bytes	1,000,000,000,000,000,000,000 bytes
yottabyte (YB)	1000^8 bytes	1,000,000,000,000,000,000,000,000 bytes

Volume: scale of data

- 90% of today’s data has been created in just the last 2 years
- Every day we create 2.5 quintillion bytes of data or enough to fill 10 million Blu-ray discs
- 40 zettabytes (4o trillion gigabytes) of data will be created by 2020, an increase of 300 times from 2005, and the equivalent of 5,200 gigabytes of data for every man, woman and child on Earth
- Most companies in the US have over 100 terabytes (100,000 gigabytes) of data stored

Variety: different forms of data

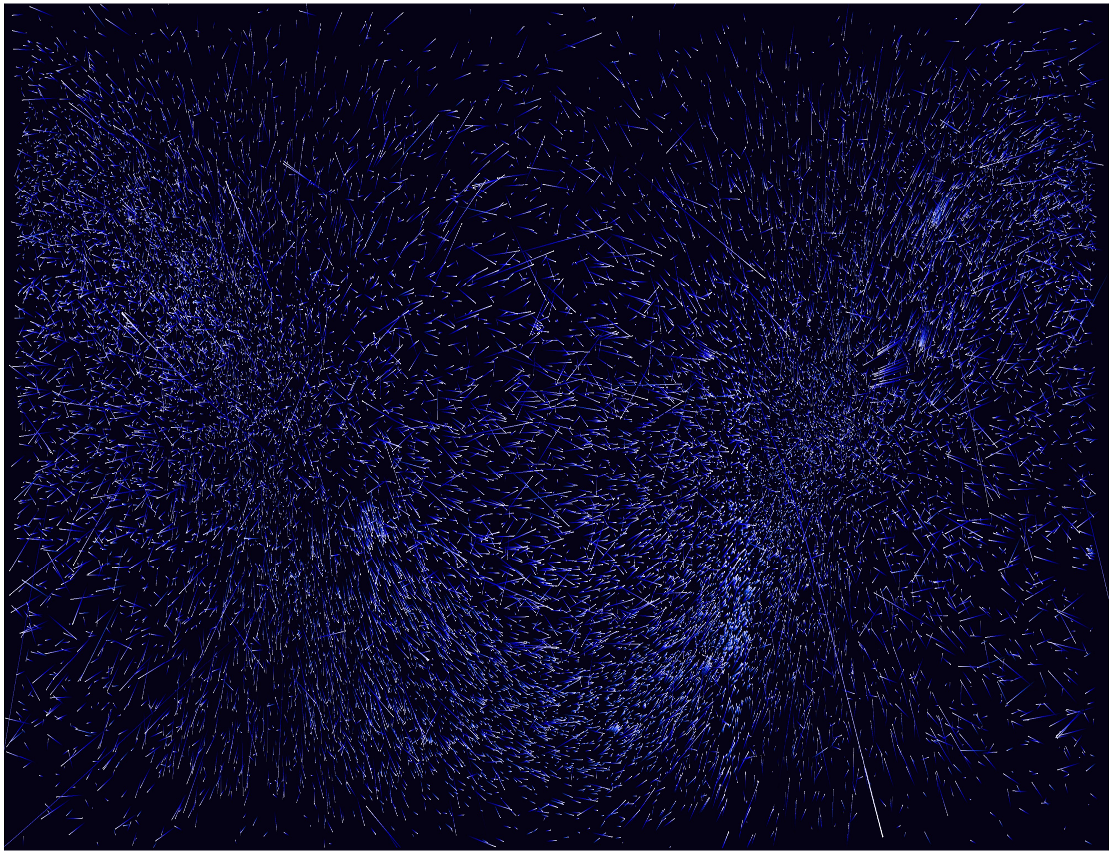


NIH U.S. National Library of Medicine
National Network of Libraries of Medicine
Pacific Northwest Region

Velocity: analysis of streaming data

Veracity: trustworthiness of data

- Origin
- Authenticity
- Trustworthiness
- Completeness
- Integrity



Value

Volume

The amount of data

Variety

The types of data

The 4 V's of Big Data

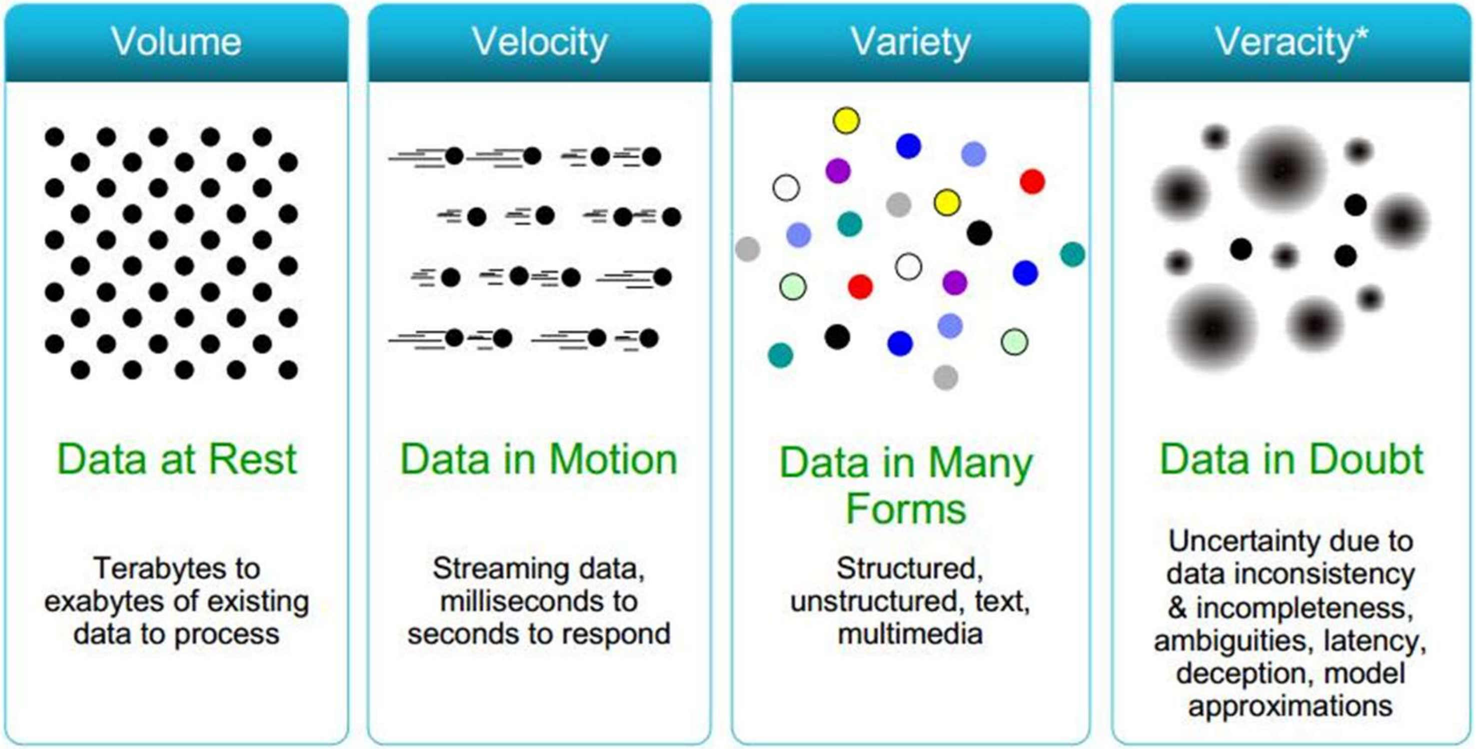
Velocity

The frequency of data

Veracity

The quality of data

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

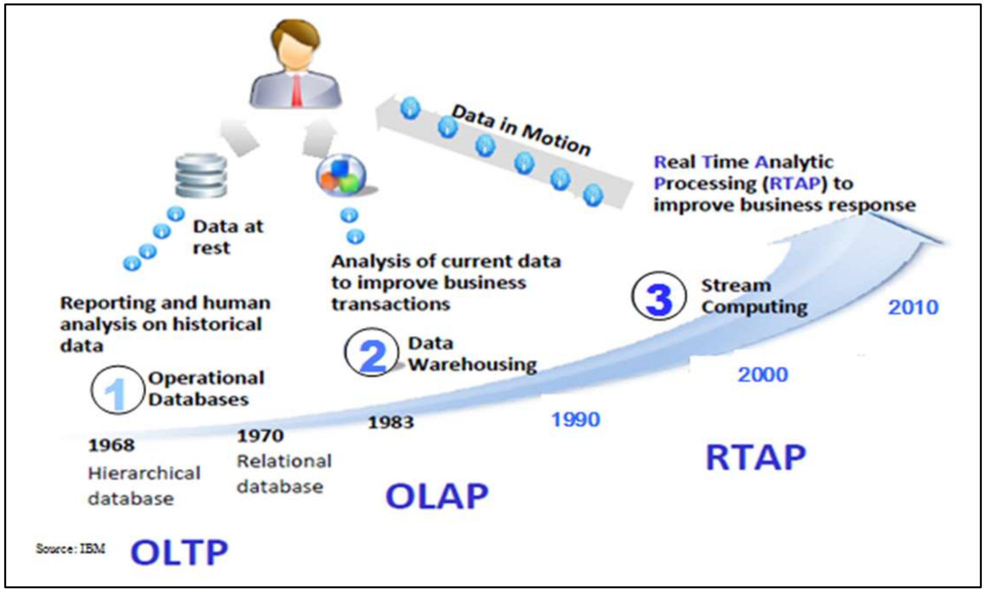
Structured, unstructured, text, multimedia

Veracity*

Data in Doubt

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

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
2 Data Warehousing
1 Operational Databases
2010
2000
1990
1983
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)

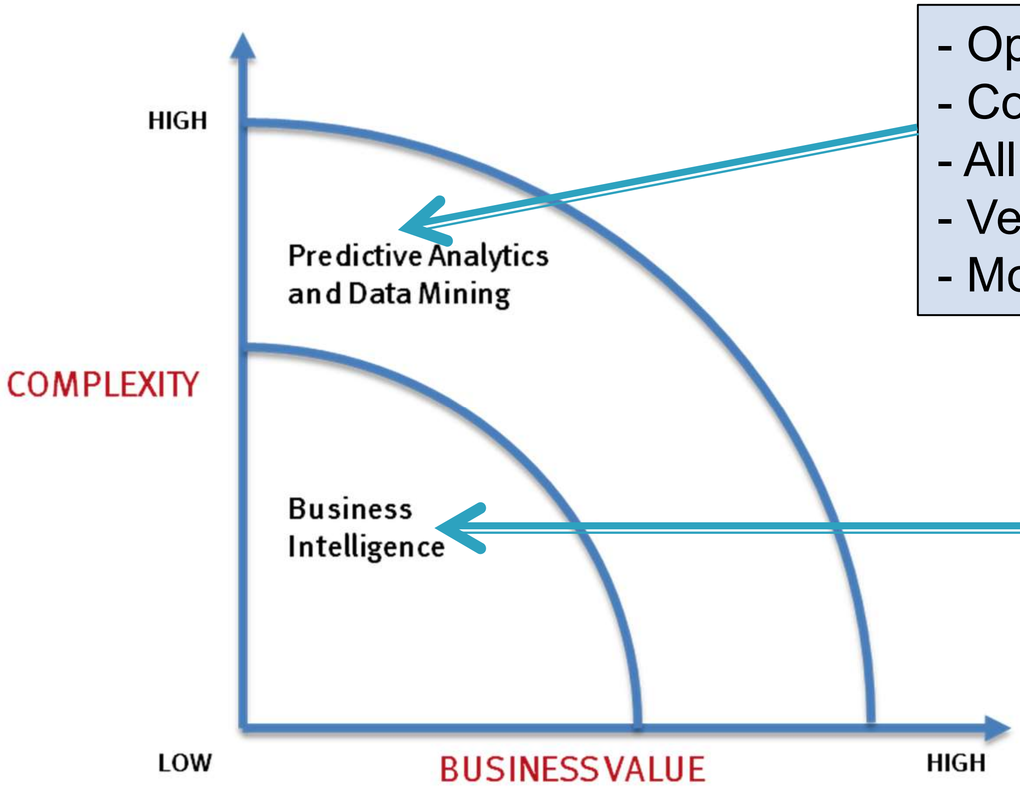
The Model Has Changed…

- **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

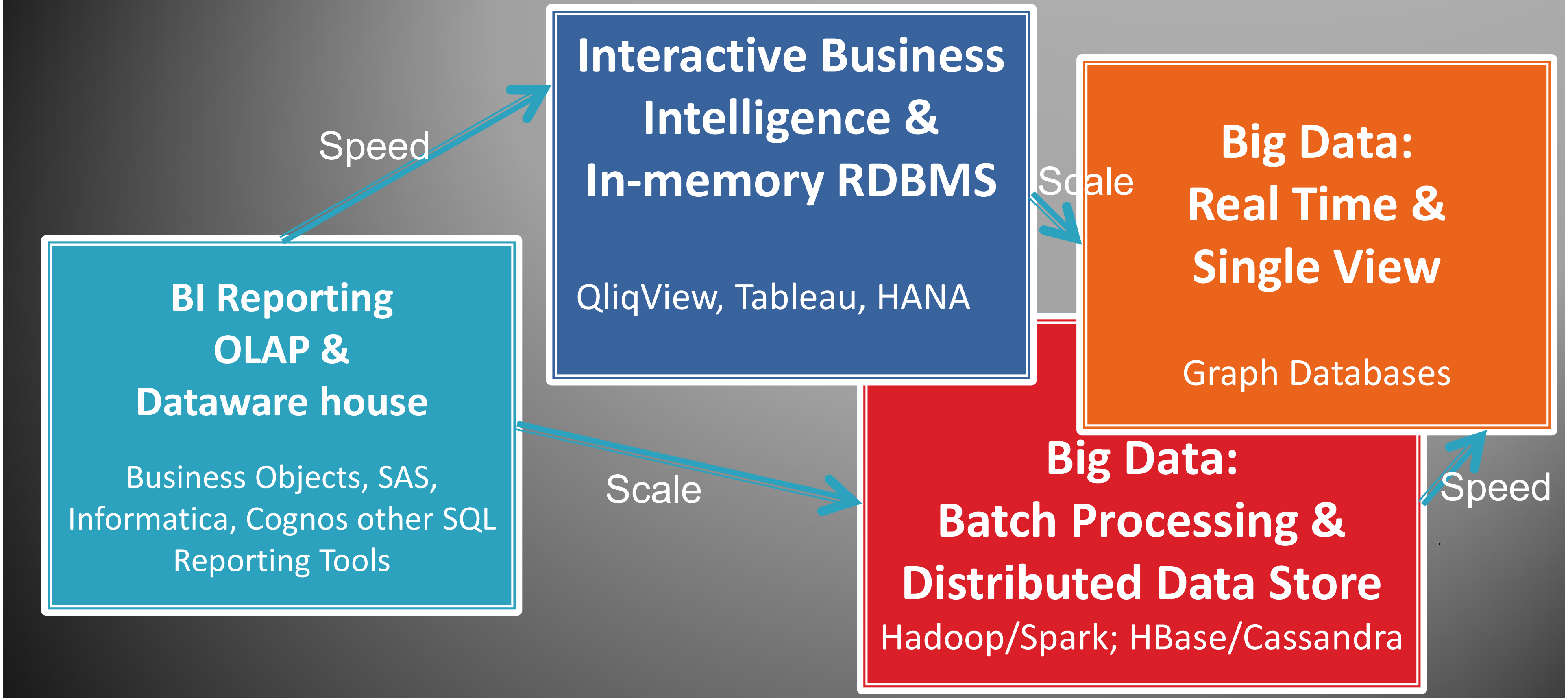
What’s driving Big Data



- 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**

Types of Data

- When collecting or gathering data we collect data from individuals cases on particular variables.

- A *variable* is a unit of data collection whose value can vary.

- Variables can be defined into *types* according to the level of mathematical scaling that can be carried out on the data.

- There are four types of data or levels of measurement:

1. Categorical (Nominal)	**2. Ordinal**
3. Interval	**4. Ratio**

Categorical (Nominal) data

- **Nominal or categorical** data is data that comprises of categories that *cannot* be rank ordered – each category is just different.

- The categories available **cannot be placed in any order** and no judgement can be made about the relative size or distance from one category to another.

  - Categories bear no quantitative relationship to one another
  - Examples:
    - customer's location (America, Europe, Asia)
    - employee classification (manager, supervisor, associate)

- What does this mean**? No mathematical operations can be performed on the data relative to each other**.

- Therefore, nominal data reflect **qualitative differences** rather than quantitative ones.

Nominal data

Examples:

What is your gender? *(please tick)*

Male □
Female □

Did you enjoy the film? *(please tick)*

Yes □
No □

- Systems for measuring nominal data must ensure that each category is **mutually exclusive** and the system of measurement needs to be **exhaustive**.

- Exhaustive: the system of categories system should have enough categories for all the observations

- Variables that have only two responses i.e. Yes or No, are known as *dichotomies*.

Ordinal data

Example:

How satisfied are you with the level of service you have received? *(please tick)*

Very satisfied □
Somewhat satisfied □
Neutral □
Somewhat dissatisfied □
Very dissatisfied □

- Ordinal data is data that **comprises of categories that *can* be rank ordered.**
- Similarly with nominal data the distance between each category cannot be calculated but the **categories can be ranked above or below each other.**
  - No fixed units of measurement
  - Examples:
    - college football rankings
    - survey responses (poor, average, good, very good, excellent)
- What does this mean? Can **make statistical judgements** and perform limited maths.

Interval and ratio data

- Both interval and ratio data are examples of **scale data**.
- Scale data:
  - data is in numeric format ($50, $100, $150)
  - data that can be **measured on a continuous scale**
  - the **distance** between each can be observed and as a result **measured**
  - the data can be **placed in rank order**.

Interval data

- Ordinal data but with constant differences between observations
- Ratios are not meaningful
- Examples:

•**Time** – moves along a continuous measure or seconds, minutes and so on and is without a zero point of time.

• **Temperature** – moves along a continuous measure of degrees and is without a true zero.

•**SAT scores**

Ratio data

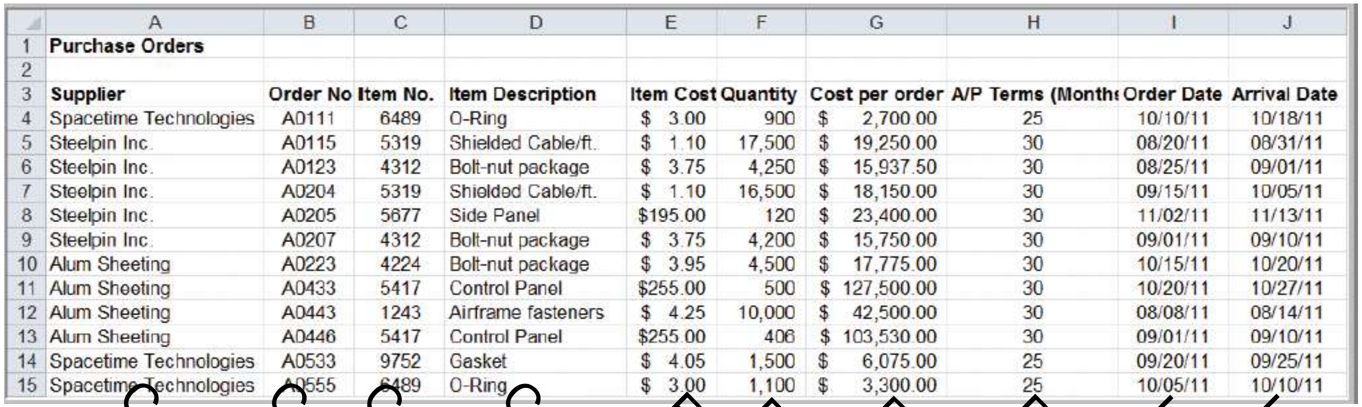
- Ratio data measured on a ***continuous* scale** and ***does* have a natural zero point.**

- Ratios are meaningful
- Examples:
  - monthly sales
  - delivery times
  - Weight
  - Height
  - Age

Data for Business Analytics

(continued)

Classifying Data Elements in a Purchasing Database



	A	B	C	D	E	F	G	H	I	J
1	**Purchase Orders**									
2										
3	**Supplier**	**Order No**	**Item No.**	**Item Description**	**Item Cost**	**Quantity**	**Cost per order**	**A/P Terms (Months**	**Order Date**	**Arrival Date**
4	Spacetime Technologies	A0111	6489	O-Ring	$ 3.00	900	$ 2,700.00	25	10/10/11	10/18/11
5	Steelpin Inc.	A0115	5319	Shielded Cable/ft.	$ 1.10	17,500	$ 19,250.00	30	08/20/11	08/31/11
6	Steelpin Inc.	A0123	4312	Bolt-nut package	$ 3.75	4,250	$ 15,937.50	30	08/25/11	09/01/11
7	Steelpin Inc.	A0204	5319	Shielded Cable/ft.	$ 1.10	16,500	$ 18,150.00	30	09/15/11	10/05/11
8	Steelpin Inc.	A0205	5677	Side Panel	$195.00	120	$ 23,400.00	30	11/02/11	11/13/11
9	Steelpin Inc.	A0207	4312	Bolt-nut package	$ 3.75	4,200	$ 15,750.00	30	09/01/11	09/10/11
10	Alum Sheeting	A0223	4224	Bolt-nut package	$ 3.95	4,500	$ 17,775.00	30	10/15/11	10/20/11
11	Alum Sheeting	A0433	5417	Control Panel	$255.00	500	$ 127,500.00	30	10/20/11	10/27/11
12	Alum Sheeting	A0443	1243	Airframe fasteners	$ 4.25	10,000	$ 42,500.00	30	08/08/11	08/14/11
13	Alum Sheeting	A0446	5417	Control Panel	$255.00	406	$ 103,530.00	30	09/01/11	09/10/11
14	Spacetime Technologies	A0533	9752	Gasket	$ 4.05	1,500	$ 6,075.00	25	09/20/11	09/25/11
15	Spacetime Technologies	A0555	6489	O-Ring	$ 3.00	1,100	$ 3,300.00	25	10/05/11	10/10/11

Figure 1.2

Categorical Categorical Categorical Categorical Ratio Ratio Ratio Ratio Interval Interval

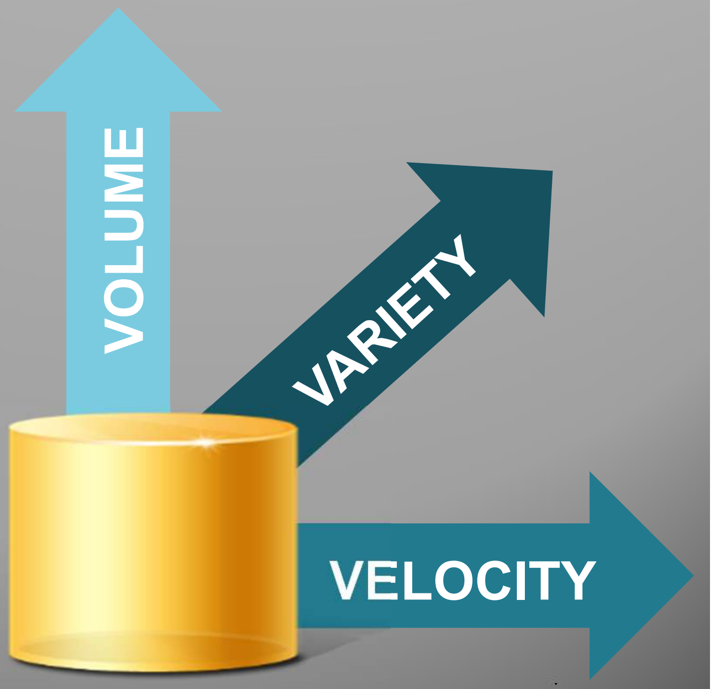
If there was field (column) for **Supplier Rating** (*Excellent, Good, Acceptable, Bad*), that data would be classified as **Ordinal**

Big Data Characteristics

Growing quantity of data
e.g. social media, behavioral, video

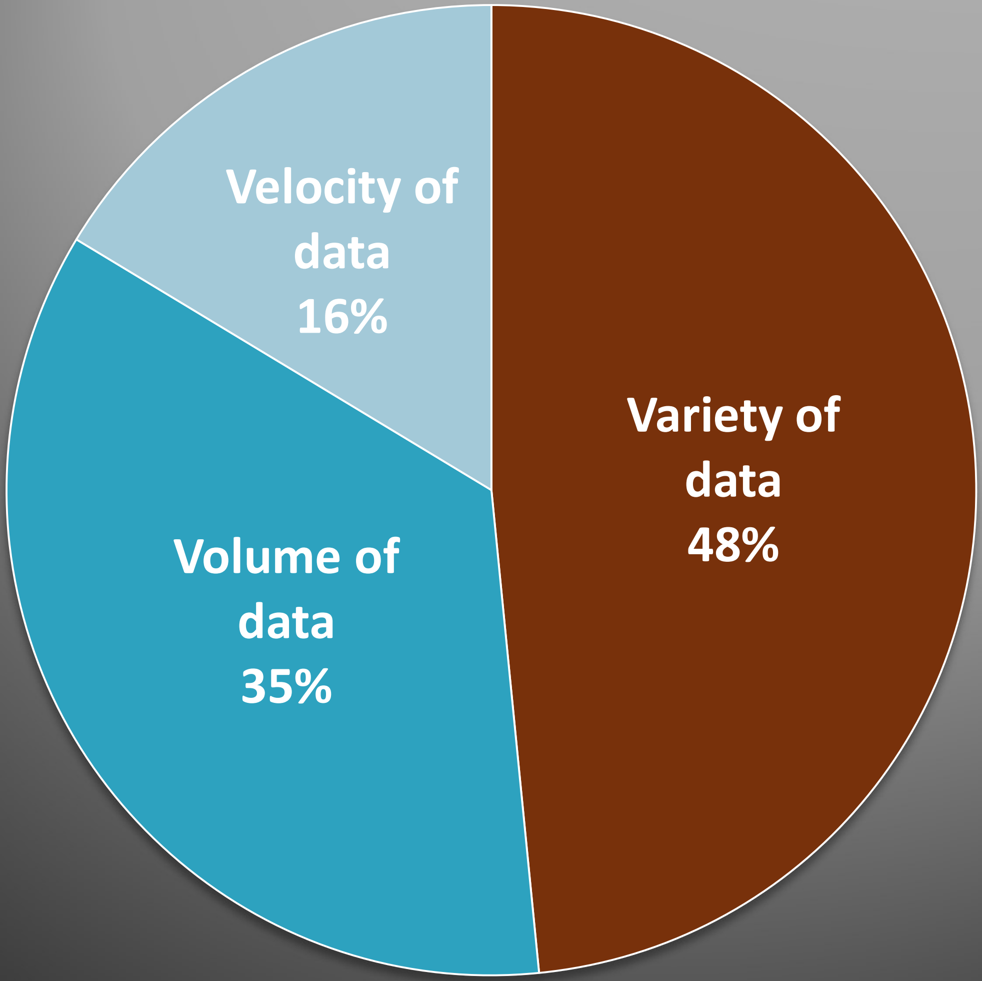
Quickening speed of data
e.g. smart meters, process monitoring

Increase in types of data
e.g. app data, unstructured data



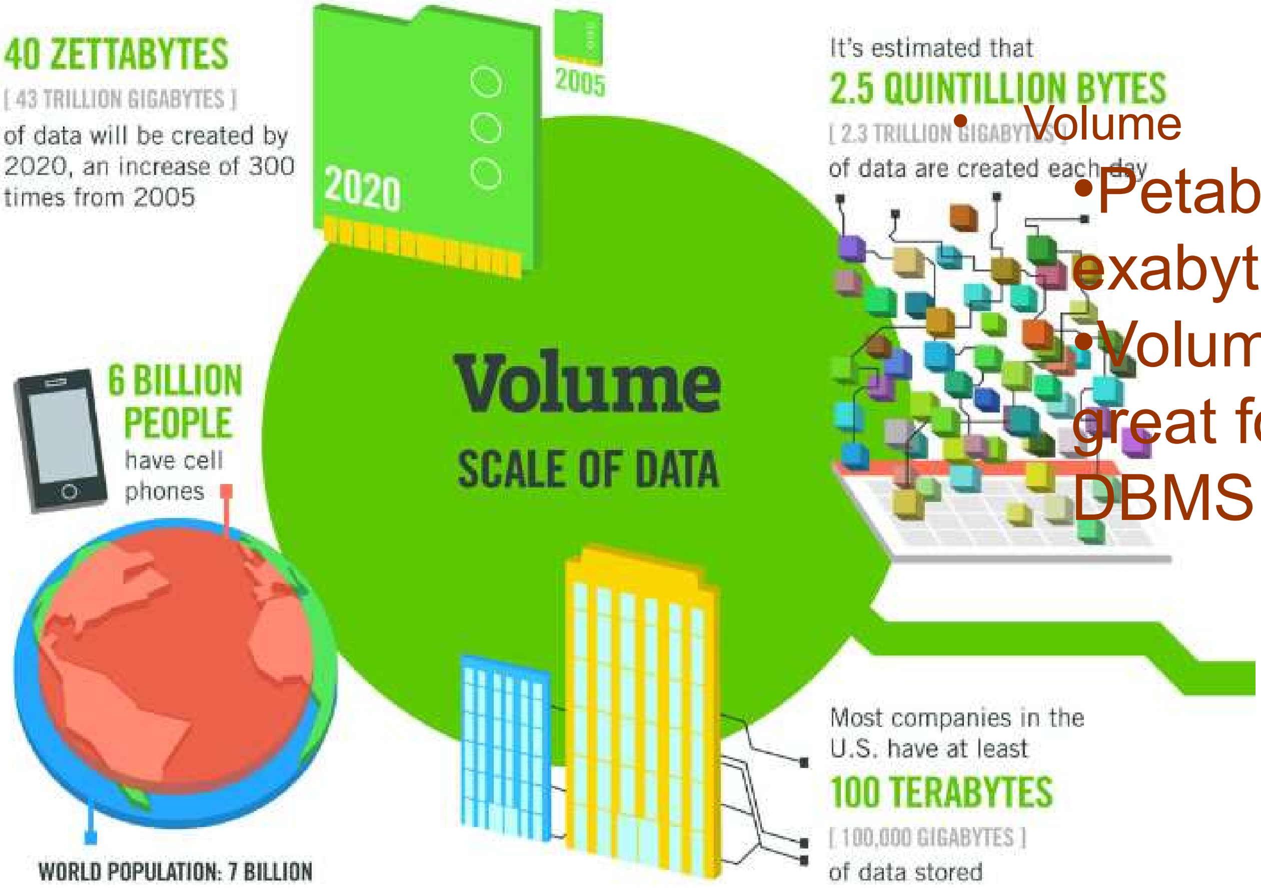
Gartner, Feb 2001

Which Big Data characteristic is the biggest issue for your organization?



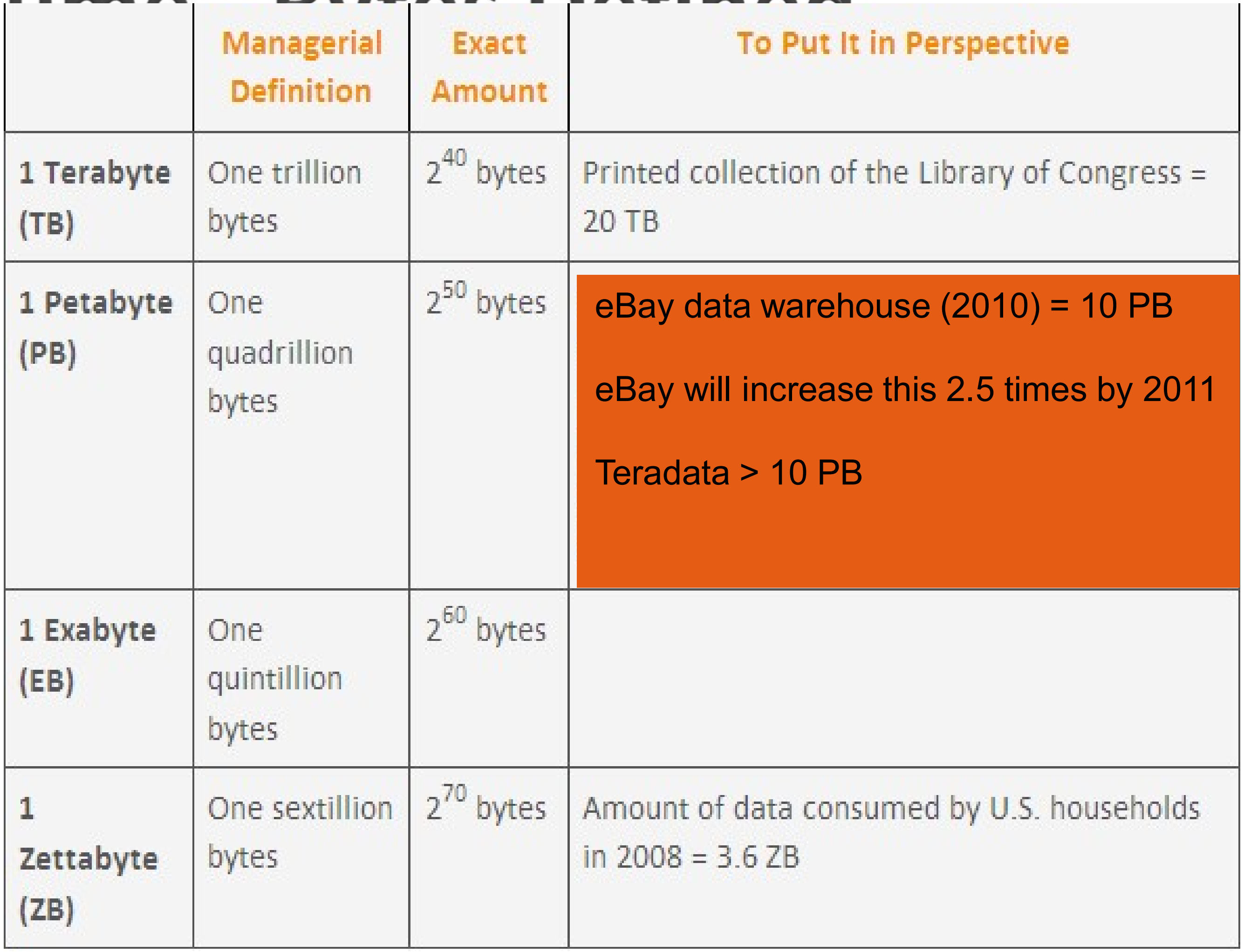
Source: Getting Value from Big Data, Gartner Webinar, May 2012

Volume


40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2020
2005
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

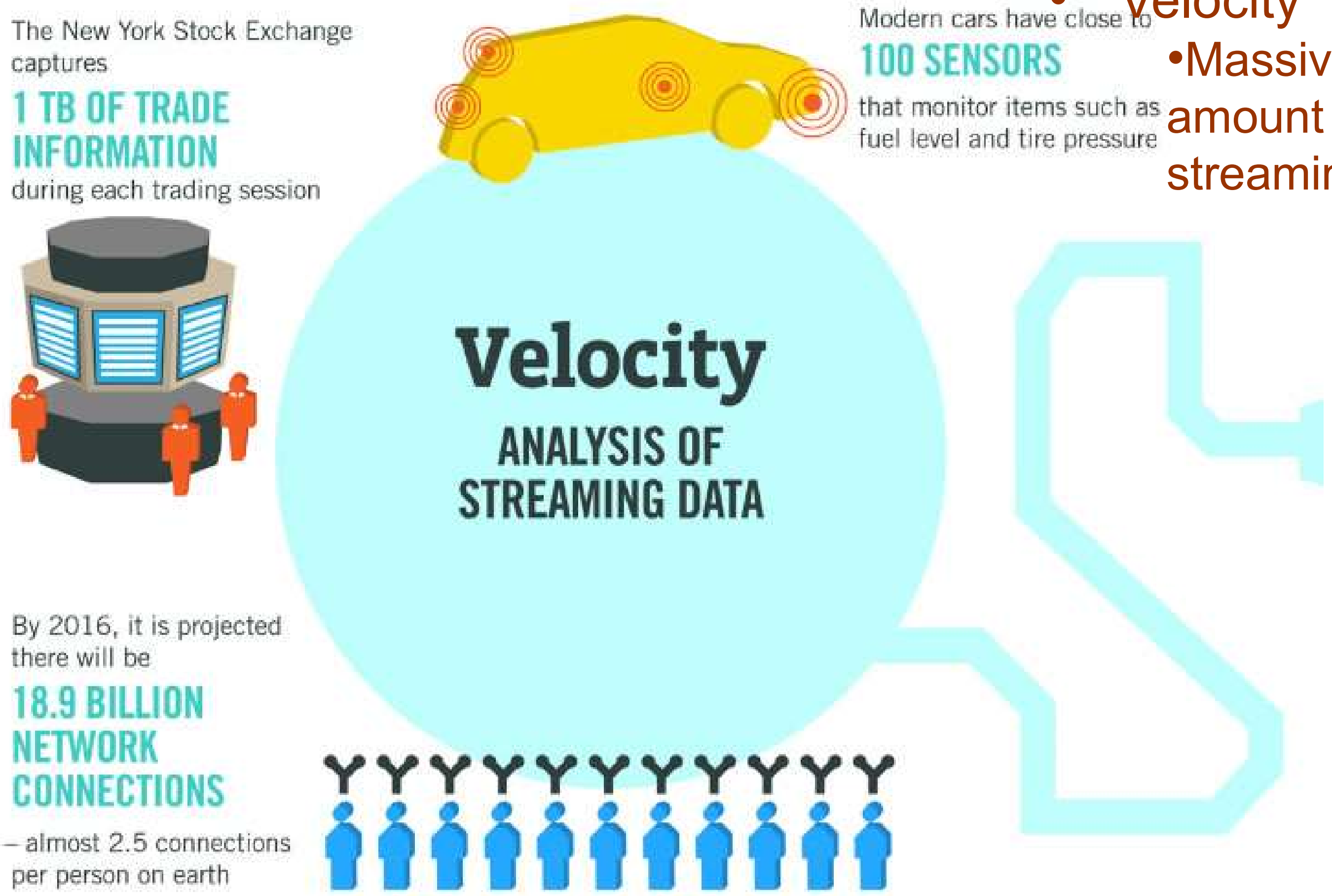


	Managerial Definition	Exact Amount	To Put It in Perspective
1 Terabyte (TB)	One trillion bytes	2^{40} bytes	Printed collection of the Library of Congress = 20 TB
1 Petabyte (PB)	One quadrillion bytes	2^{50} bytes	eBay data warehouse (2010) = 10 PB eBay will increase this 2.5 times by 2011 Teradata > 10 PB
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

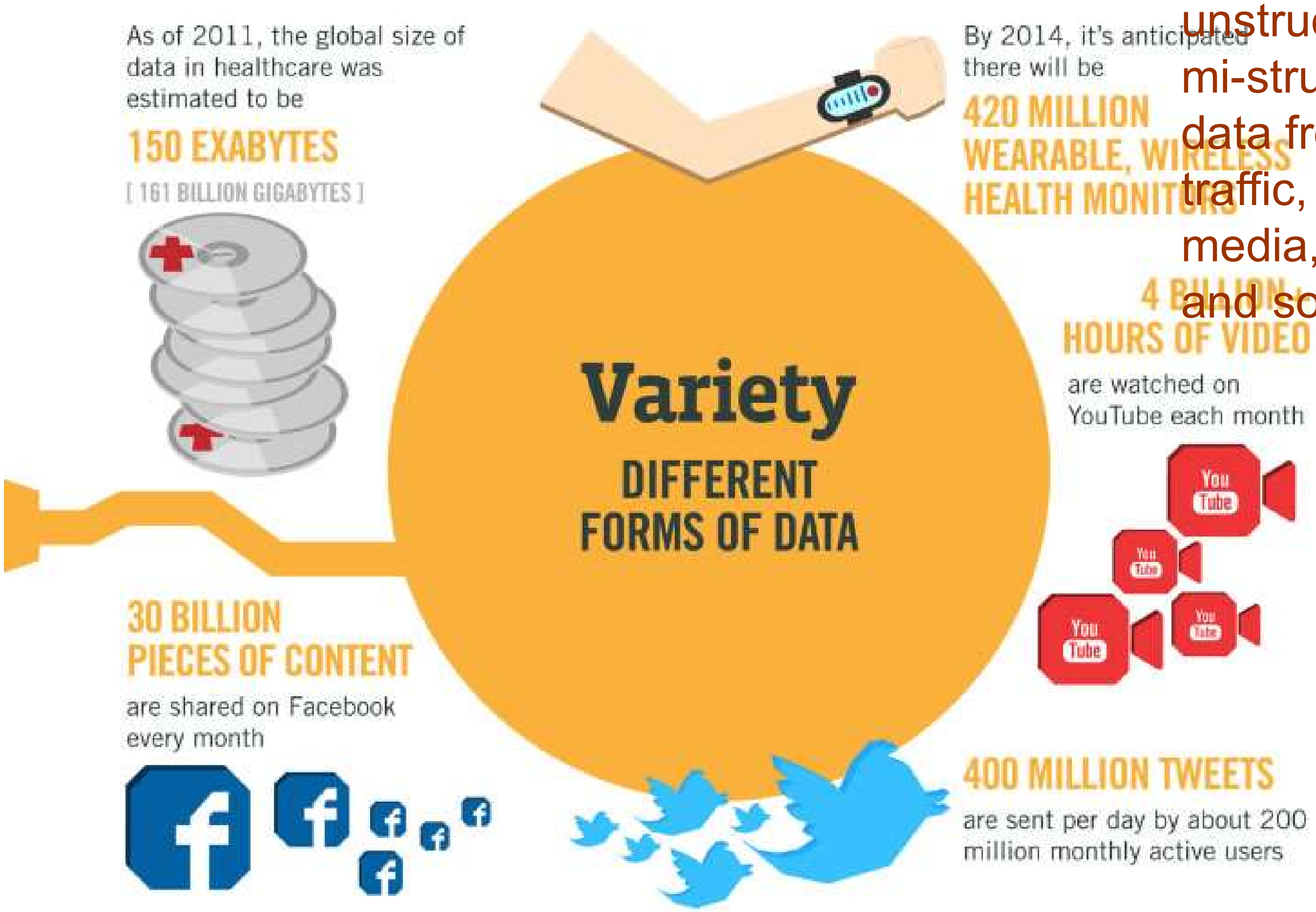
Velocity


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

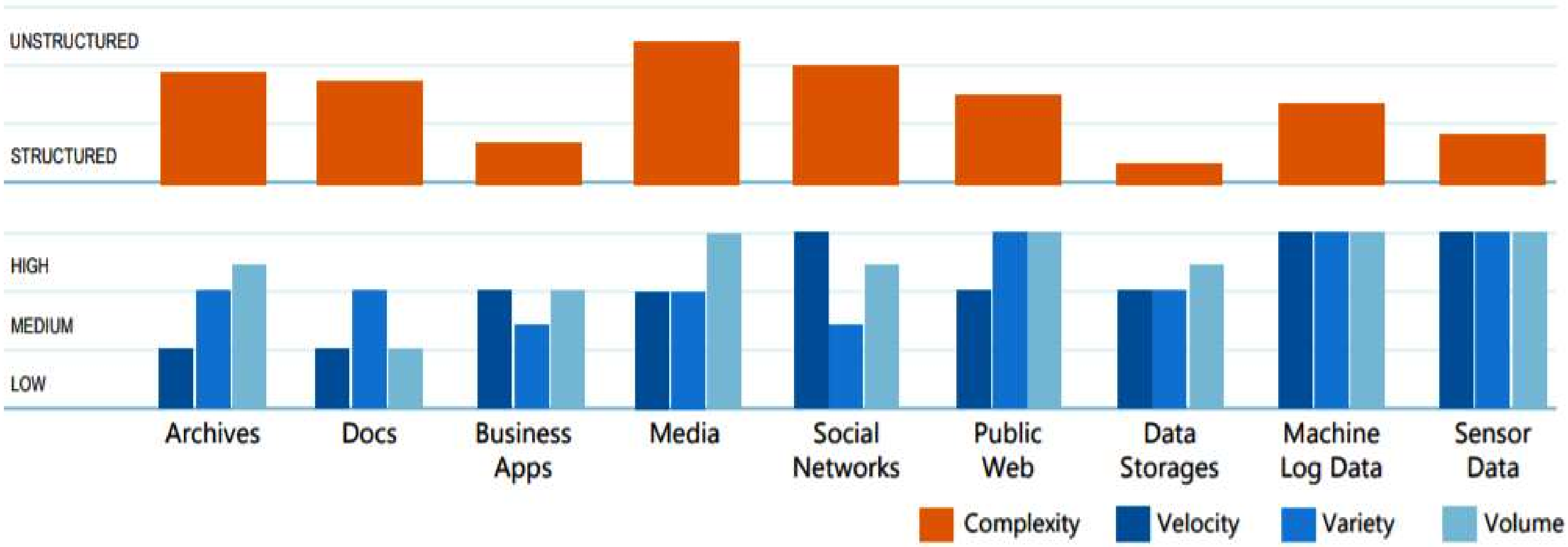
- Velocity
  - Massive amount of streaming data

Variety

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


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

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

What is Structured Data?

- Structured data usually resides in relational databases (RDBMS). Fields store length-delineated data phone numbers, Social Security numbers, or ZIP codes.
- Even text strings of variable length like names are contained in records, making it a simple matter to search.
- Data may be human- or machine-generated as long as the data is created within an RDBMS structure.
- This format is eminently searchable both with human generated queries and via algorithms using type of data and field names, such as alphabetical or numeric, currency or date.
- Common relational database applications with structured data include airline reservation systems, inventory control, sales transactions, and ATM activity.
- Structured Query Language (SQL) enables queries on this type of structured data within relational databases.

What is Unstructured Data?

- Unstructured data is essentially everything else.
- Unstructured data has internal structure but is not structured via pre-defined data models or schema.
- It may be textual or non-textual, and human- or machine-generated. It may also be stored within a non-relational database like NoSQL.
- Unstructured data, in contrast, refers to data that doesn’t fit neatly into the traditional row and column structure of relational databases.
- Examples of unstructured data include: emails, videos, audio files, web pages, and social media messages.
- In today’s world of Big Data, most of the data that is created is unstructured with some estimates of it being more than 95% of all data generated.

- As a result, enterprises are looking to this new generation of databases, known as NoSQL, to address unstructured data.
- MongoDB stands as a leader in this movement with over 10 million downloads and hundreds of thousands of deployments.
- As a document database with flexible schema, MongoDB was built specifically to handle unstructured data.
- MongoDB’s flexible data model allows for development without a predefined schema which resonates particularly when most of the data in your system is unstructured.

Un Structured data

- Unstructured data, in contrast, refers to data that doesn't fit neatly into the traditional row and column structure of relational databases.
- Examples of unstructured data include: emails, videos, audio files, web pages, and social media messages.
- In today's world of Big Data, most of the data that is created is unstructured with some estimates of it being more than 95% of all data generated.

- As a result, enterprises are looking to this new generation of databases, known as NoSQL, to address unstructured data.
- MongoDB stands as a leader in this movement with over 10 million downloads and hundreds of thousands of deployments.
- As a document database with flexible schema, MongoDB was built specifically to handle unstructured data.
- MongoDB's flexible data model allows for development without a predefined schema which resonates particularly when most of the data in your system is unstructured.

Contd..

- Typical human-generated unstructured data includes:
  - Text files: Word processing, spreadsheets, presentations, email, logs.
  - **Email**: Email has some internal structure thanks to its metadata, and we sometimes refer to it as semi-structured. However, its message field is unstructured and traditional analytics tools cannot parse it.
  - **Social Media:** Data from Facebook, Twitter, LinkedIn.
  - Website: YouTube, Instagram, photo sharing sites.
  - Mobile data: Text messages, locations.
  - Communications: Chat, IM, phone recordings, collaboration software.
  - Media: MP3, digital photos, audio and video files.
  - Business applications: MS Office documents, productivity applications.
  - Typical machine-generated unstructured data includes:

- Satellite imagery: Weather data, land forms, military movements.
- Scientific data: Oil and gas exploration, space exploration, seismic imagery, atmospheric data.
- Digital surveillance: Surveillance photos and video.
- Sensor data: Traffic, weather, oceanographic sensors.

What is Semi-structured data ?

- Semi-structured data maintains internal tags and markings that identify separate data elements, which enables information grouping and hierarchies.
- Both documents and databases can be semi-structured.
- This type of data only represents about 5-10% of the structured/semi-structured/unstructured data pie, but has critical business usage cases.

- Email is a very common example of a semi-structured data type.
- Although more advanced analysis tools are necessary for thread tracking, near-dedupe, and concept searching; email's native metadata enables classification and keyword searching without any additional tools.

- Email is a huge use case, but most semi-structured development centers on easing data transport issues.
- Sharing sensor data is a growing use case, as are Web-based data sharing and transport: electronic data interchange (EDI), many social media platforms, document markup languages, and NoSQL databases.

Examples of Semi-structured Data

Markup language XML

- This is a semi-structured document language. XML is a set of document encoding rules that defines a human- and machine-readable format.

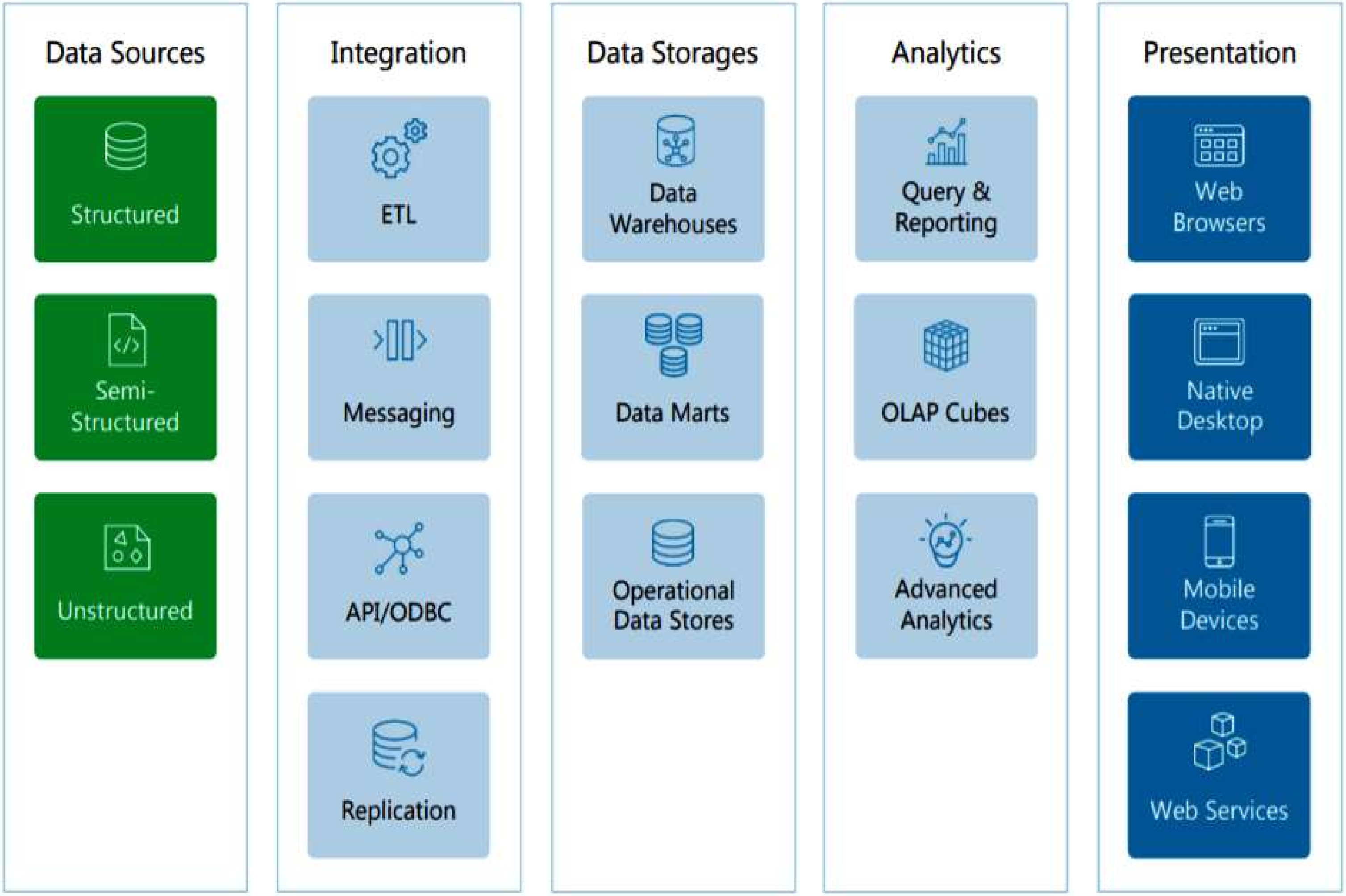
Open standard JSON (JavaScript Object Notation)

- JSON is another semi-structured data interchange format. Java is implicit in the name but other C-like programming languages recognize it.
- Its structure consists of name/value pairs (or object, hash table, etc.) and an ordered value list (or array, sequence, list). Since the structure is interchangeable among languages, JSON excels at transmitting data between web applications and servers.

NoSQL

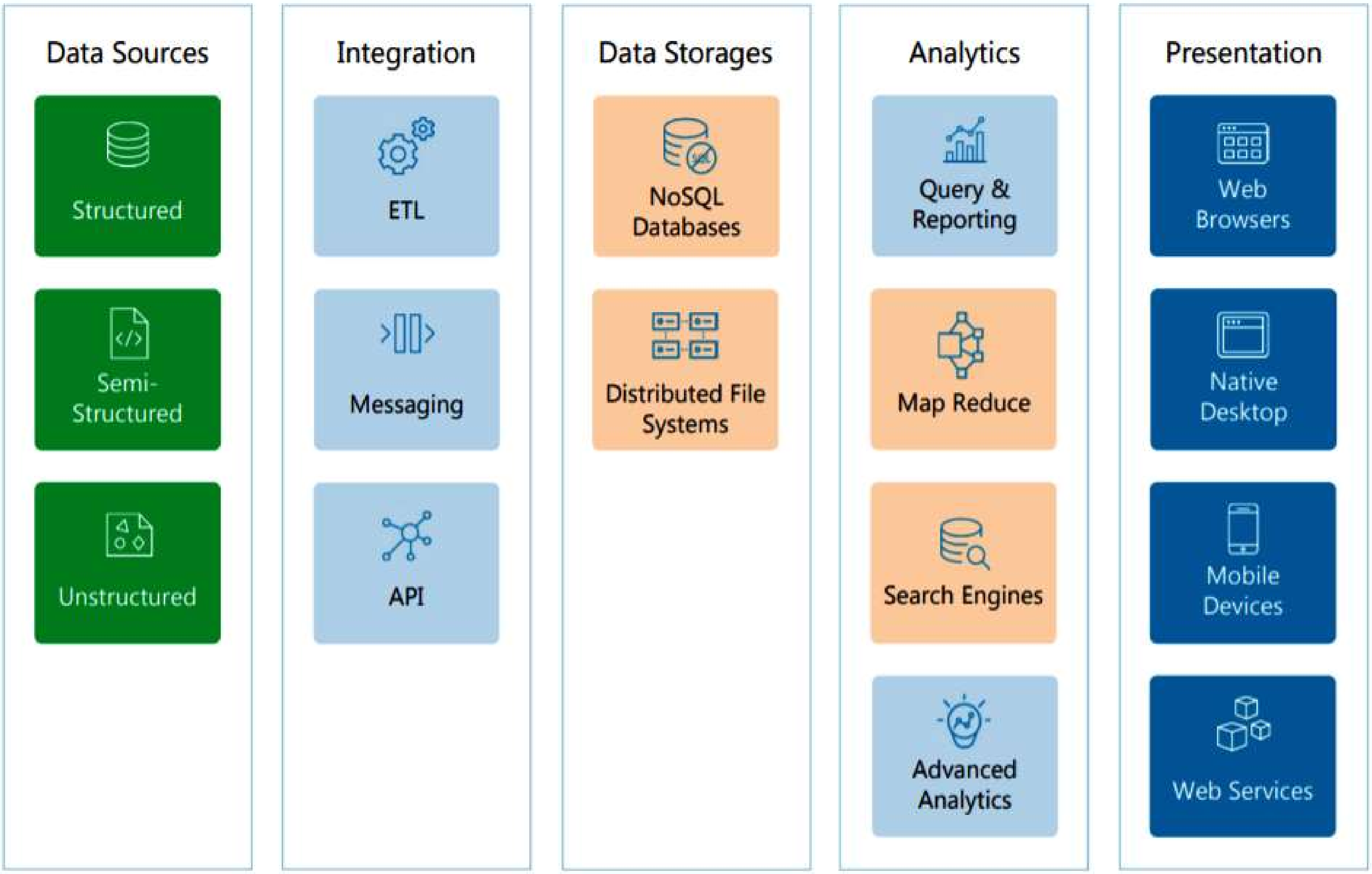
- Semi-structured data is also an important element of many NoSQL ("not only SQL") databases.
- NoSQL databases differ from relational databases because they do not separate the organization (schema) from the data.
- This makes NoSQL a better choice to store information that does not easily fit into the record and table format, such as text with varying lengths.

Relational Reference Architecture



Data Sources	Integration	Data Storages	Analytics	Presentation
Structured	ETL	Data Warehouses	Query & Reporting	Web Browsers
Semi-Structured	Messaging	Data Marts	OLAP Cubes	Native Desktop
Unstructured	API/ODBC	Operational Data Stores	Advanced Analytics	Mobile Devices
	Replication			Web Services

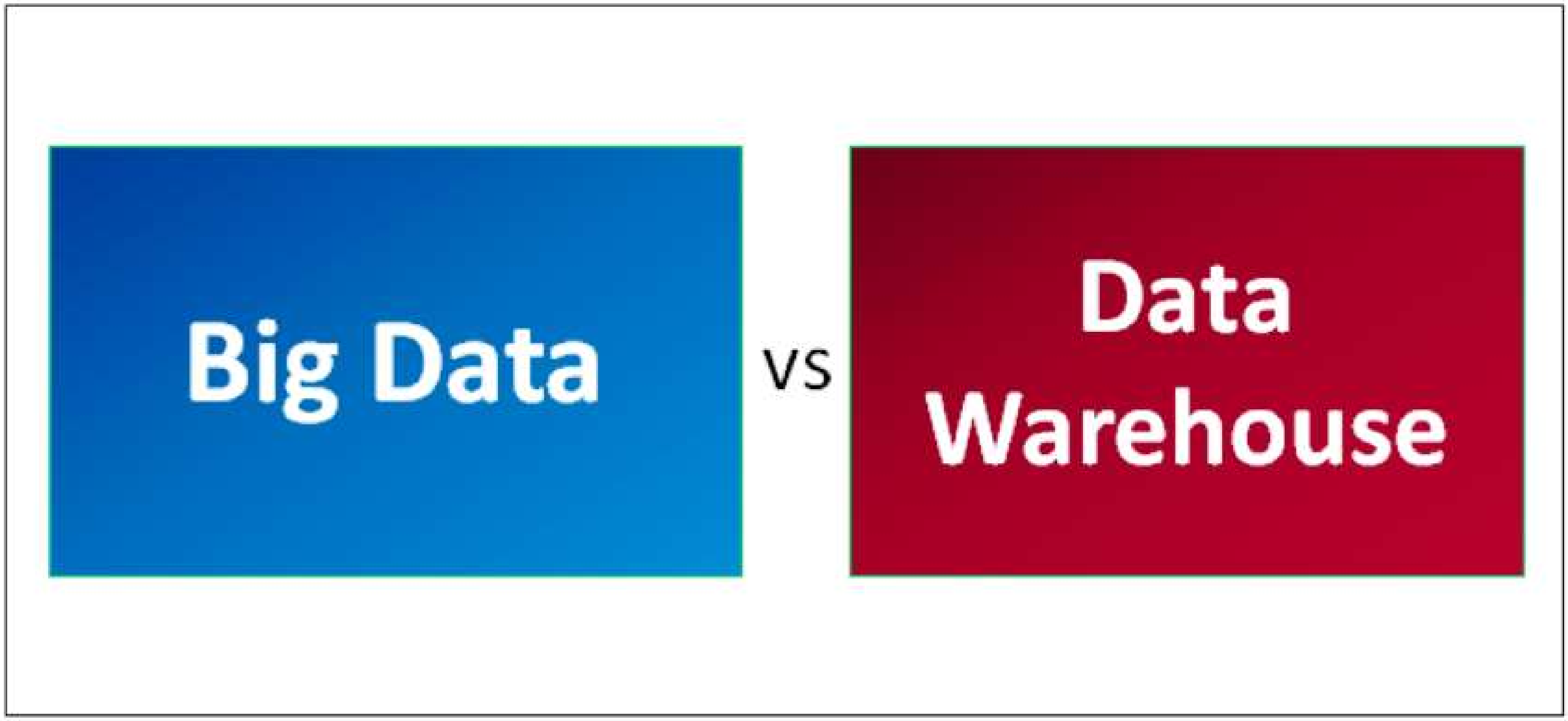
Non-Relational Reference Architecture



Data Sources	Integration	Data Storages	Analytics	Presentation
Structured	ETL	NoSQL Databases	Query & Reporting	Web Browsers
Semi-Structured	Messaging	Distributed File Systems	Map Reduce	Native Desktop
Unstructured	API		Search Engines	Mobile Devices
			Advanced Analytics	Web Services

🟧 Key components introduced with non-relational movement

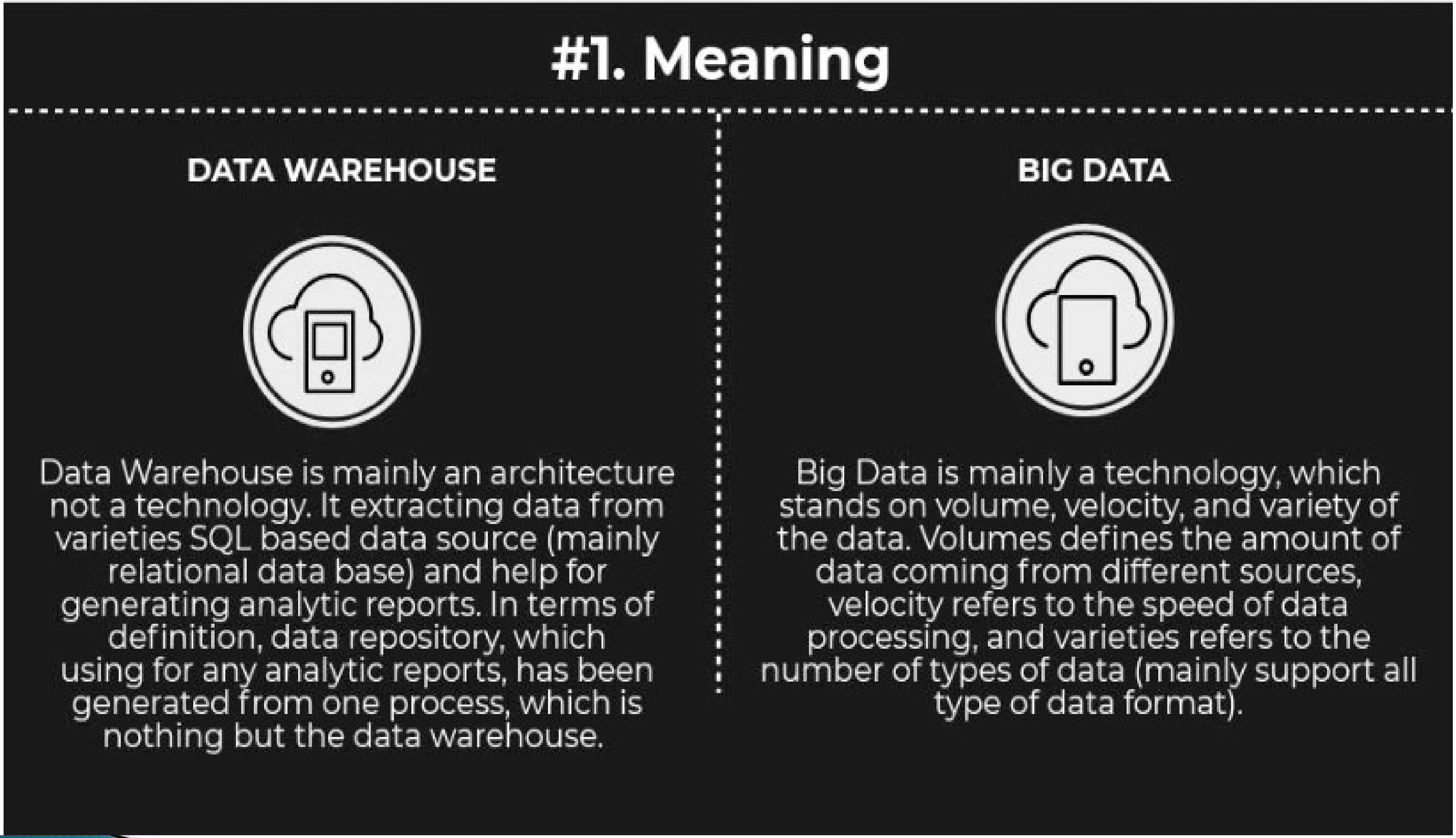
Data Environment



Difference between Big Data and Data Warehouse

- Data Warehousing is one of the common words for last 10-20 years,
- whereas Big Data is a hot trend for last 5-10 years.
- Both of them hold a lot of data, used for reporting, managed by an electronic storage device. So one common thought of maximum people that recent big data will replace old data warehousing very soon.
- But still, big data and data warehousing is not interchangeable as they used totally for a different purpose.

Data Warehouse vs Big Data



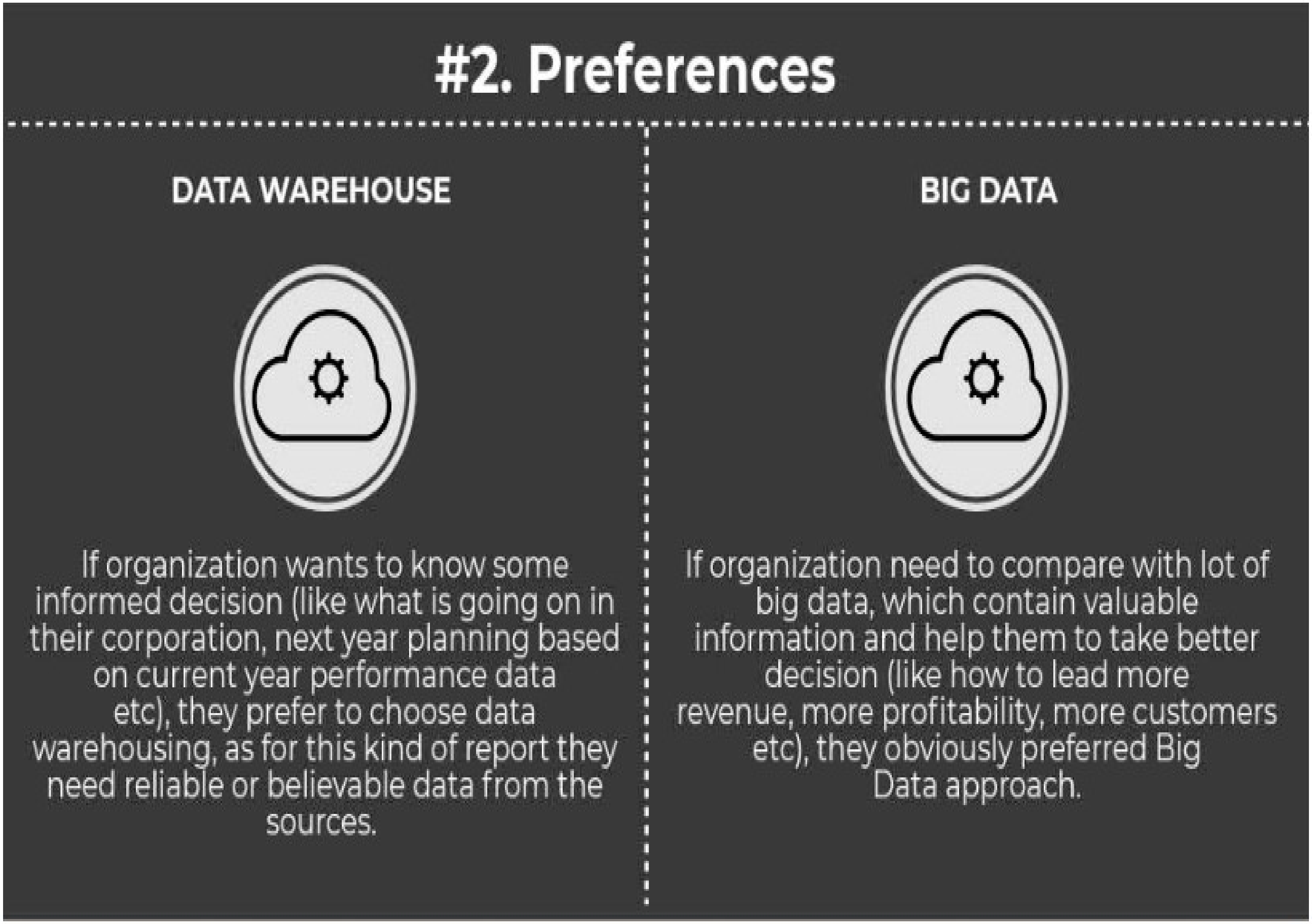
#1. Meaning

DATA WAREHOUSE

Data Warehouse is mainly an architecture not a technology. It extracting data from varieties SQL based data source (mainly relational data base) and help for generating analytic reports. In terms of definition, data repository, which using for any analytic reports, has been generated from one process, which is nothing but the data warehouse.

BIG DATA

Big Data is mainly a technology, which stands on volume, velocity, and variety of the data. Volumes defines the amount of data coming from different sources, velocity refers to the speed of data processing, and varieties refers to the number of types of data (mainly support all type of data format).



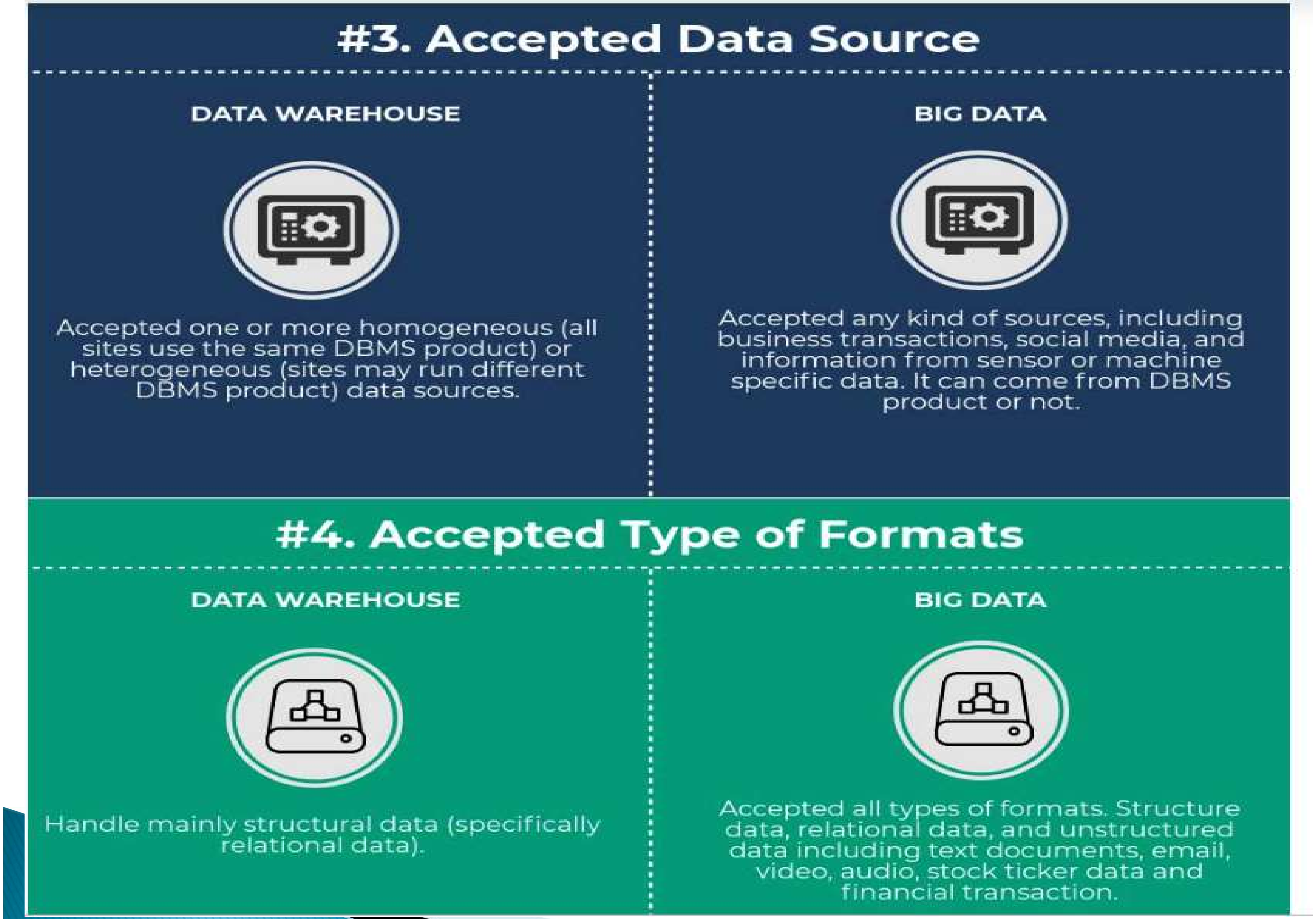
#2. Preferences

DATA WAREHOUSE

If organization wants to know some informed decision (like what is going on in their corporation, next year planning based on current year performance data etc), they prefer to choose data warehousing, as for this kind of report they need reliable or believable data from the sources.

BIG DATA

If organization need to compare with lot of big data, which contain valuable information and help them to take better decision (like how to lead more revenue, more profitability, more customers etc), they obviously preferred Big Data approach.

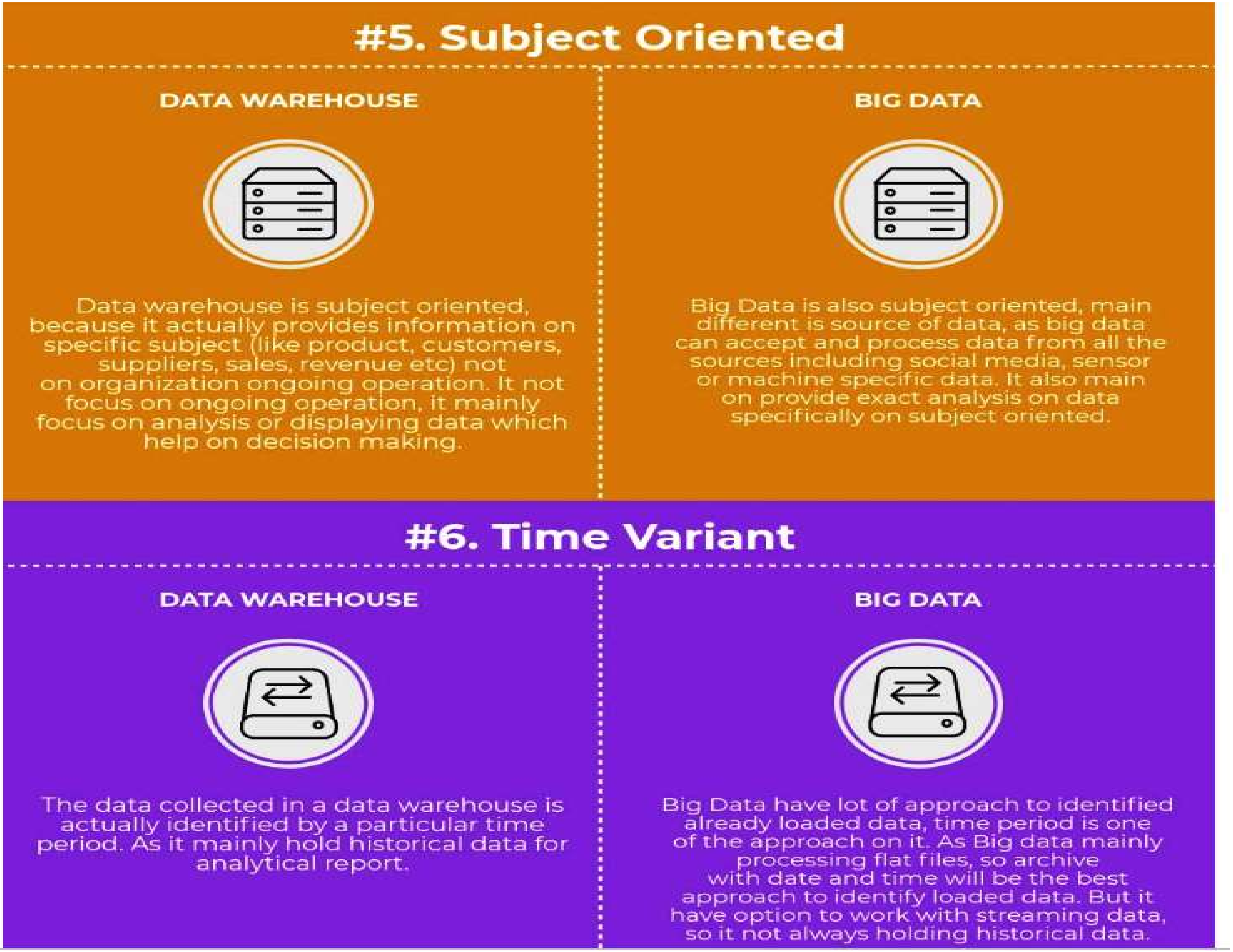


#3. Accepted Data Source

DATA WAREHOUSE	BIG DATA
Accepted one or more homogeneous (all sites use the same DBMS product) or heterogeneous (sites may run different DBMS product) data sources.	Accepted any kind of sources, including business transactions, social media, and information from sensor or machine specific data. It can come from DBMS product or not.

#4. Accepted Type of Formats

DATA WAREHOUSE	BIG DATA
Handle mainly structural data (specifically relational data).	Accepted all types of formats. Structure data, relational data, and unstructured data including text documents, email, video, audio, stock ticker data and financial transaction.



#5. Subject Oriented

DATA WAREHOUSE

Data warehouse is subject oriented, because it actually provides information on specific subject (like product, customers, suppliers, sales, revenue etc) not on organization ongoing operation. It not focus on ongoing operation, it mainly focus on analysis or displaying data which help on decision making.

BIG DATA

Big Data is also subject oriented, main different is source of data, as big data can accept and process data from all the sources including social media, sensor or machine specific data. It also main on provide exact analysis on data specifically on subject oriented.

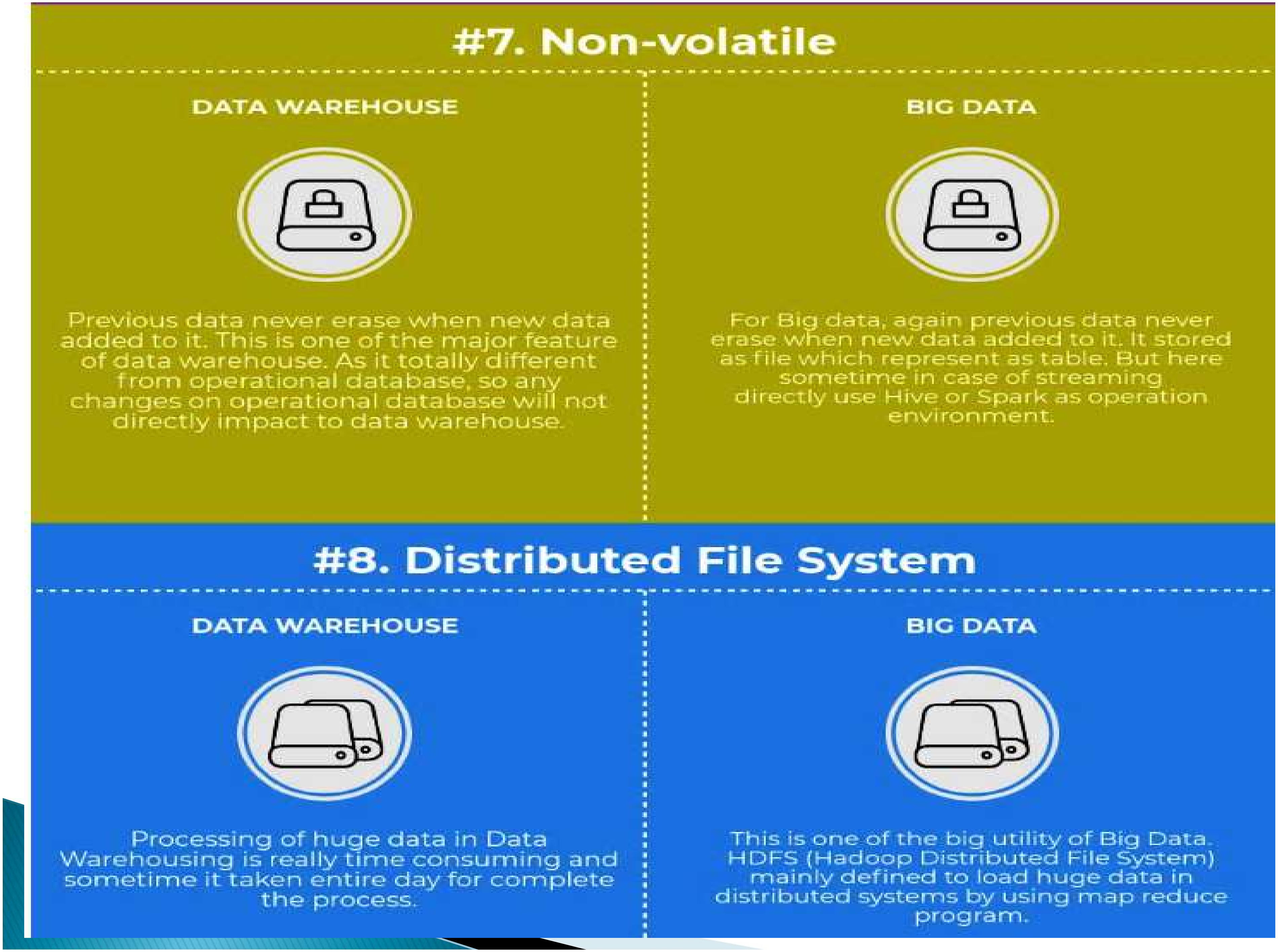
#6. Time Variant

DATA WAREHOUSE

The data collected in a data warehouse is actually identified by a particular time period. As it mainly hold historical data for analytical report.

BIG DATA

Big Data have lot of approach to identified already loaded data, time period is one of the approach on it. As Big data mainly processing flat files, so archive with date and time will be the best approach to identify loaded data. But it have option to work with streaming data, so it not always holding historical data.



#7. Non-volatile

DATA WAREHOUSE

Previous data never erase when new data added to it. This is one of the major feature of data warehouse. As it totally different from operational database, so any changes on operational database will not directly impact to data warehouse.

BIG DATA

For Big data, again previous data never erase when new data added to it. It stored as file which represent as table. But here sometime in case of streaming directly use Hive or Spark as operation environment.

#8. Distributed File System

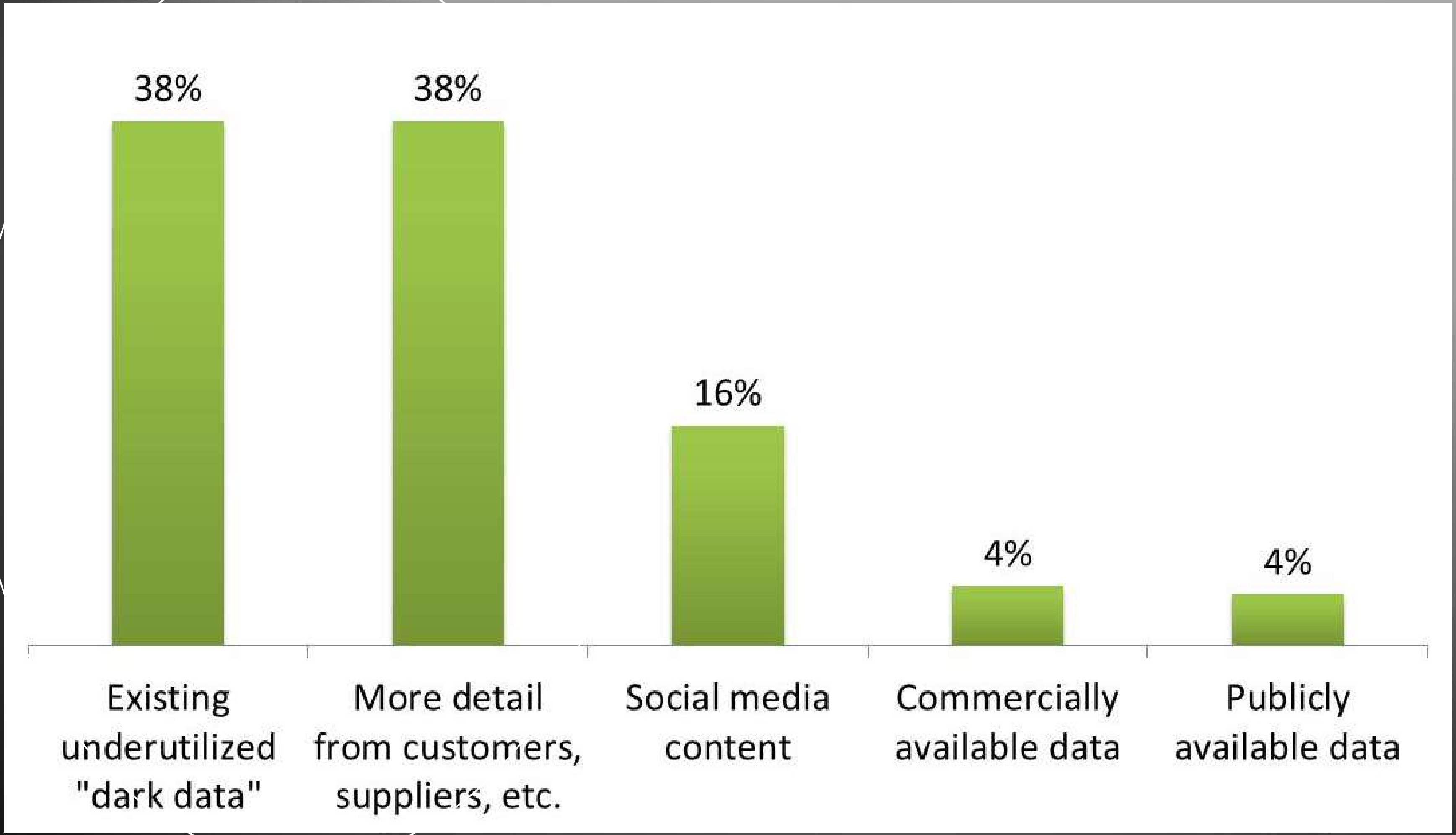
DATA WAREHOUSE

Processing of huge data in Data Warehousing is really time consuming and sometime it taken entire day for complete the process.

BIG DATA

This is one of the big utility of Big Data. HDFS (Hadoop Distributed File System) mainly defined to load huge data in distributed systems by using map reduce program.

Which source of data represents the most immediate opportunity?



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

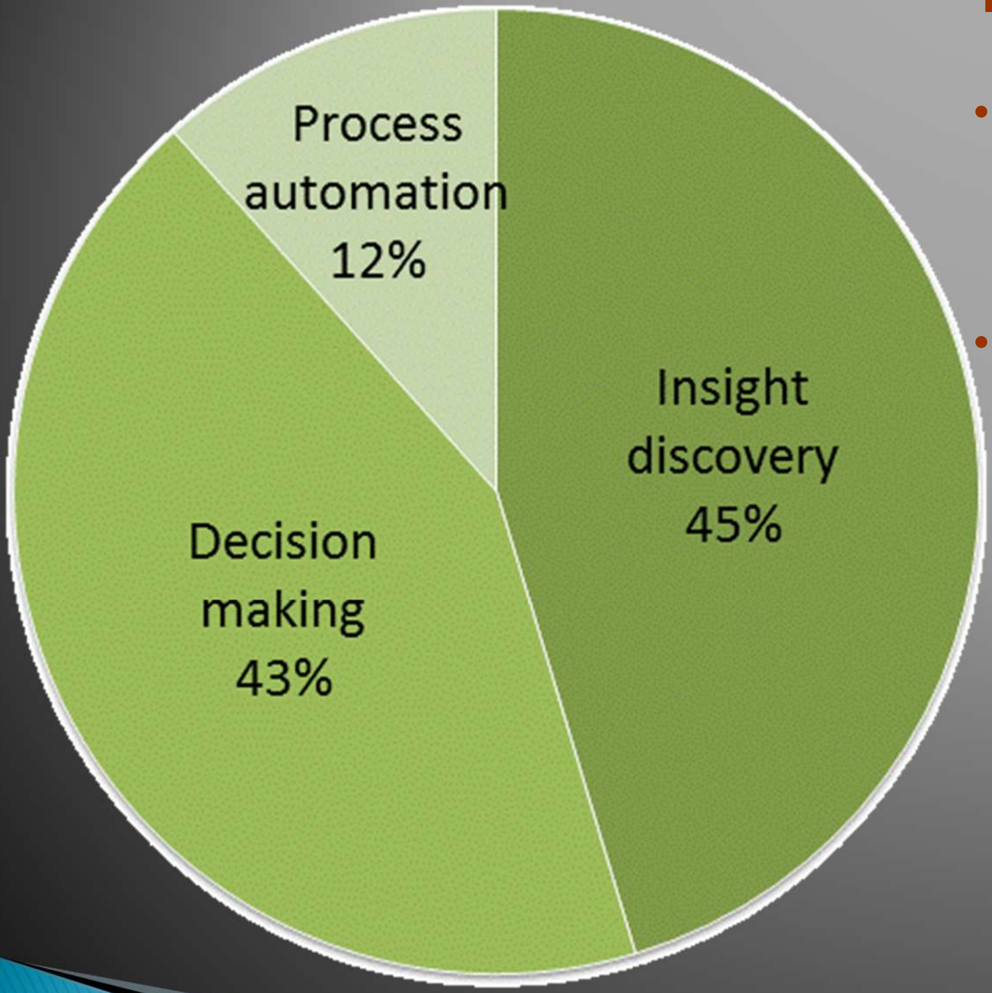
Big Data Opportunities

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

Which is the biggest opportunity for Big Data in your organization?

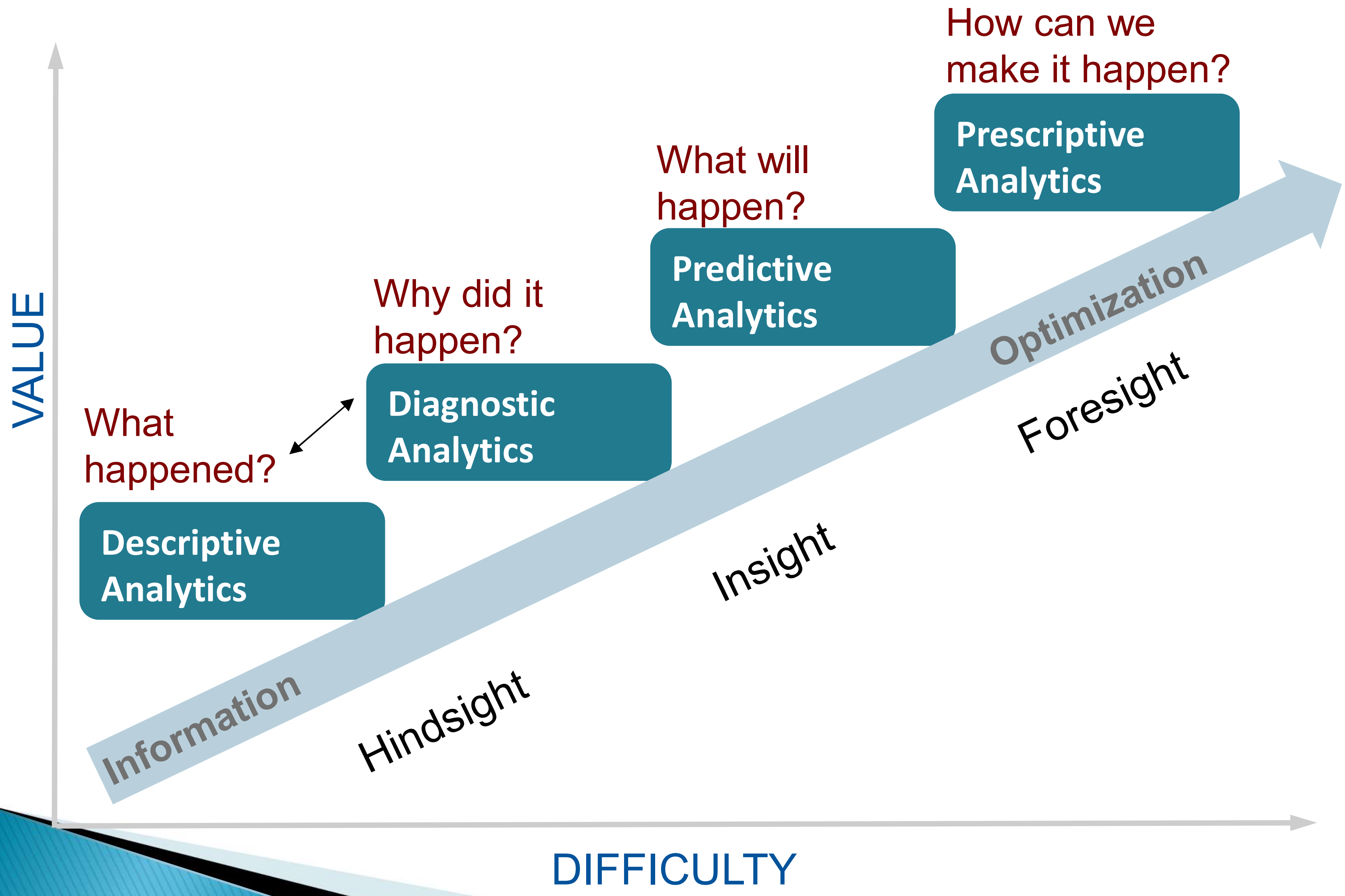


Through 2017:

- 85% of Fortune 500 organizations will be unable to exploit big data for competitive advantage.
- Business analytics needs will drive 70% of investments in the expansion and modernization of information infrastructure.

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

Analytics Models



How can we make it happen?

Prescriptive Analytics

What will happen?

Predictive Analytics

Why did it happen?

Diagnostic Analytics

What happened?

Descriptive Analytics

VALUE

Information

Optimization

Hindsight

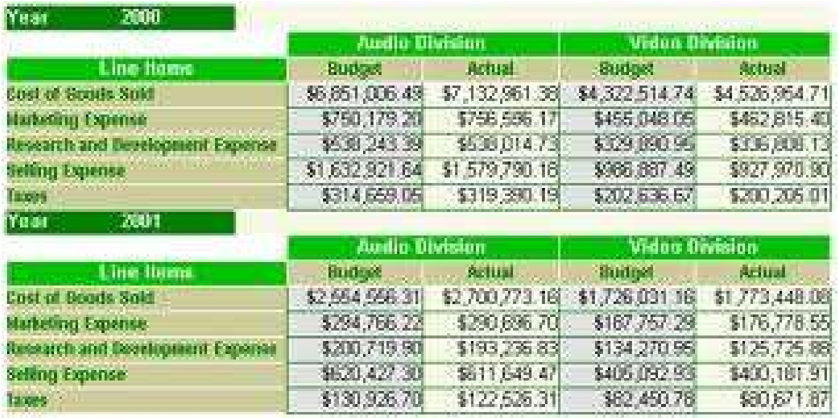
Insight

Foresight

DIFFICULTY

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.

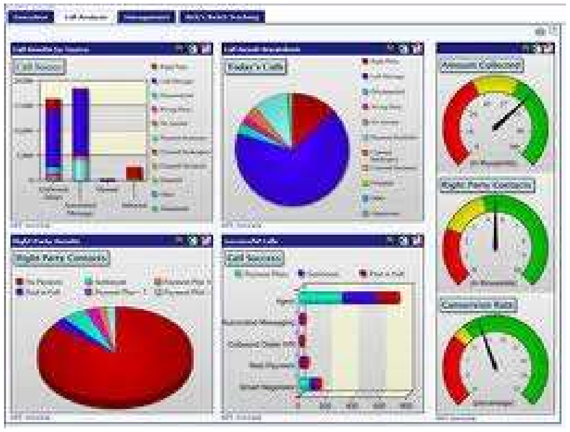


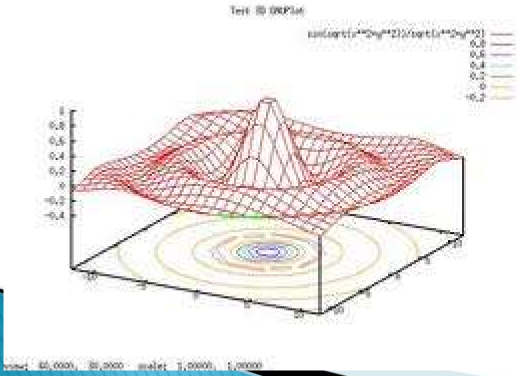
Year 2000

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	$750,179.28	$756,596.17	$455,048.05	$462,815.40
Research and Development Expense	$538,243.39	$538,014.73	$329,890.96	$336,808.13
Selling Expense	$1,632,921.64	$1,579,790.18	$986,887.49	$927,970.90
Taxes	$314,659.06	$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	$2,564,956.31	$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	$400,181.91
Taxes	$130,926.70	$122,526.31	$82,450.78	$80,671.87



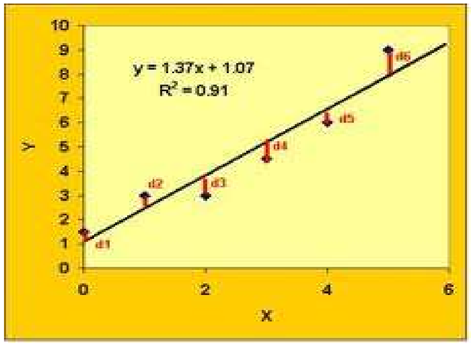


What has occurred?

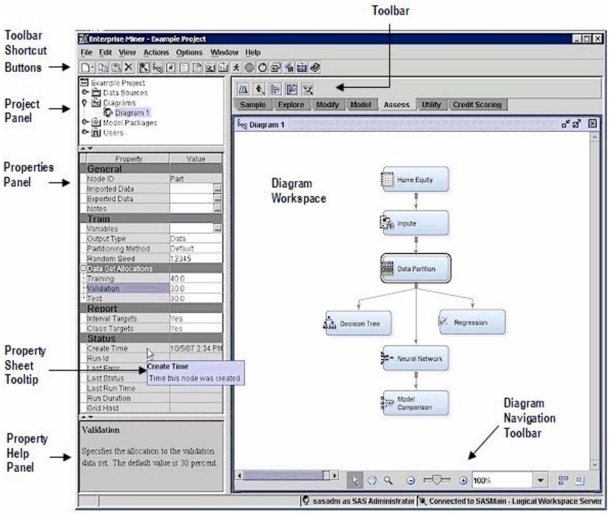
Descriptive analytics, such as data visualization, is important in helping users interpret the output from predictive and predictive analytics.

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.

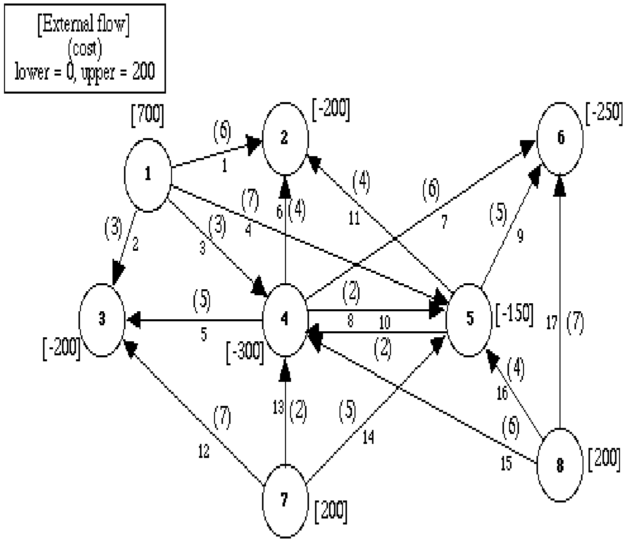
Validation

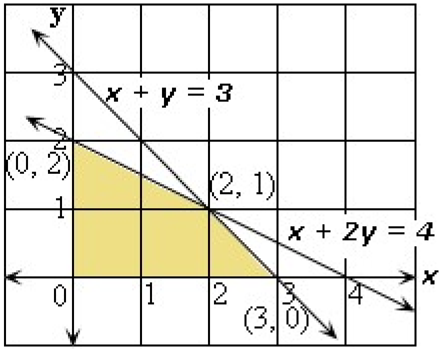
Specifies the allocation to the validation data set. The default value is 30 percent.

- 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.

Organizational Transformation

- Analytics are a competitive requirement

- For BI-based organizations, the use of BI/analytics is a ***requirement*** for successfully competing in the marketplace.
- TDWI report on Big Data Analytics found that 85% of respondents indicated that their firms would be using advanced analytics within three years

- IBM/*MIT Sloan Management Review* research study found that top performing companies in their industry are much more likely to use analytics rather than intuition across the widest range of possible decisions.

Complex Systems Require Analytics

- Tackle **complex problems** and provide **individualized solutions**
- **Products and services are organized** around the needs of **individual customers**
- **Dollar value** of interactions with **each customer** is **high**
- There is **high level of interaction** with **each customer**
- Examples: IBM, World Bank, Halliburton

Volume Operations Require Analytics

- Serves **high-volume markets** through standardized products and services
- Each customer interaction has a **low dollar value**
- **Customer interactions are generally conducted through technology** rather than person-to-person
- Are likely to be analytics-based
- Examples: Amazon.com, eBay, Hertz

The Nature of the Industry

- Online retailers like Amazon.com and Overstock.com are high volume operations who rely on analytics to compete.
- When you enter their sites a cookie is placed on your PC and all clicks are recorded.
- Based on your clicks and any search terms, recommendation engines decide what products to display.
- After you purchase an item, they have additional information that is used in marketing campaigns.
- Customer segmentation analysis is used in deciding what promotions to send you.
- How profitable you are influences how the customer care center treats you.
- A pricing team helps set prices and decides what prices are needed to clear out merchandise.
- Forecasting models are used to decide how many items to order for inventory.
- Dashboards monitor all aspects of organizational performance

Knowledge Requirements for Advanced Analytics

Business Domain

Data

Modeling

- Choosing the **right data** to include in models is important.
- Important to have some thoughts as to **what variables might be related**.
- **Domain knowledge is necessary to understand how they can be used**. Role of Business Analyst is crucial
- Consider the story of the relationship between beer and diapers in the market basket of young males in convenience stores.
  - You still have to decide (or experiment to discover) whether it is better to put them together or spread them across the store (in the hope that other things will be bought while walking the isles).

The findings were that men between 30- 40 years in age, shopping between 5pm and 7pm on Fridays, who purchased diapers were most likely to also have beer in their carts. This motivated the grocery store to move the beer isle closer to the diaper isle and instantaneously, a 35% increase in sales of both!

MODULE 3

Visualization

Visualization: Acquisition of Insight

- Many people and institutions possess data that may ‘hide‘ fundamental relations
  - Realtors
  - Bankers
  - Air Traffic Controller
  - Fraud investigators
  - Engineers
- They want to be able to view some graphica representation of that data, maybe interact with it, anc then be able to say…….ahha!

Example: Fraud Detection

- The Serious Fraud Office (SFO) suspected mortgage fraud
- The SFO provided 12 filing cabinets of data
- After 12 person years a suspect was identified
- The suspect was arrested, tried and convicted

Example: Fraud Detection continued

- The data was supplied in electronic form
- A visualization tool (Netmap) was used to examine the data
- After 4 person weeks the same suspect was identified
- **A master criminal behind the fraud was also identified**

Is Information Visualization Useful?

Drugs and Chips

Texas Instruments

Manufactures microprocessors on silicon wafers that are routed through 400 steps in many weeks. This process is monitored, gathering 140,000 pieces of information about each wafer. Somewhere in that heap of data can be warnings about things going wrong. Detect a bug early before bad chips are made. TI uses visualization tools to make the detection process easier

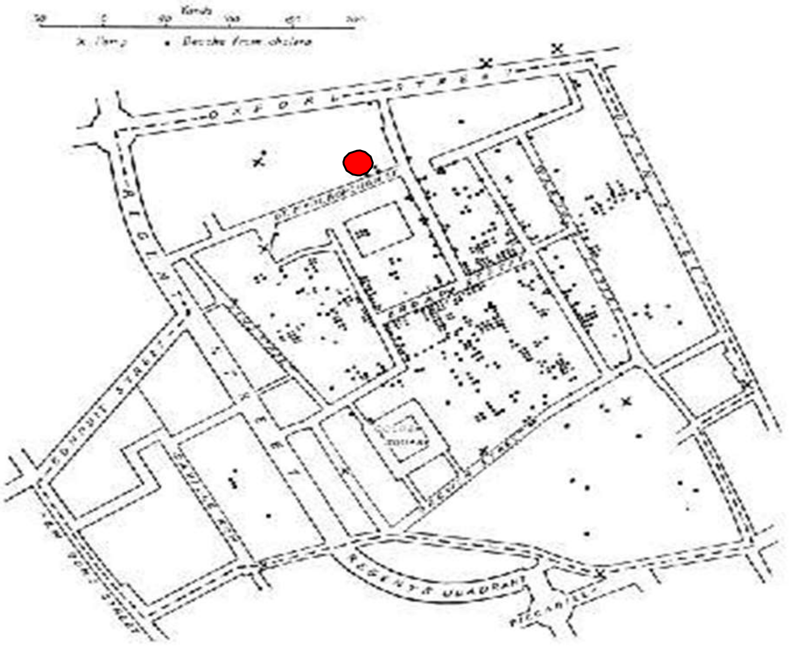
Eli Lilly

Has 1500 scientists using an advanced information visualization tool (Spotfire) for decision making. “With its ability to represent multiple sources of information and interactively change your view, it’s helpful for homing in on specific molecules and deciding whether we should be doing further testing on them”

Sheldon Ort of Eli Lilly, speaking to Fortune

The Cholera Epidemic, London 1845

Dr. John Snow, medical officer for London, investigated the cholera epidemic of 1845 in Soho. He mapped the deaths and noted that the deaths, indicated by points, tended to occur near the Broad Street pump. Closure of the pump coincided with a reduction in cholera.



Challenger Disaster

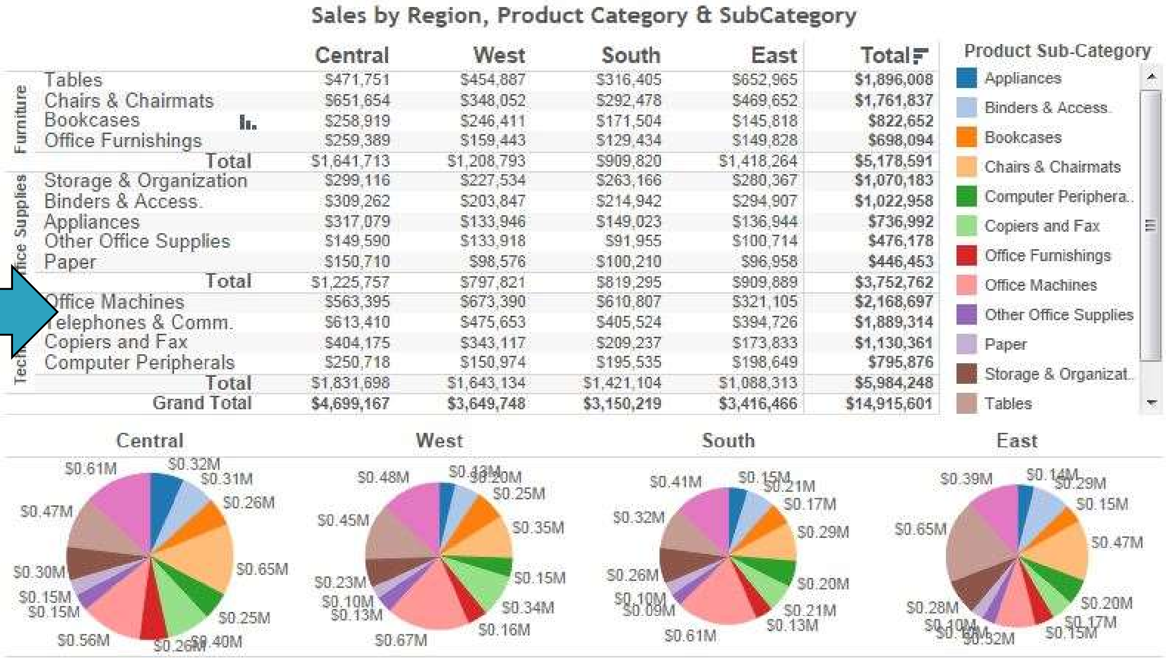
- On 28th January 1986 the space shuttle Challenger exploded, and seven astronauts died, because two rubber O-Rings leaked.
- The previous day, engineers who designed the rocket opposed the launch, concerned that the O-Rings would not seal at the forecast temperature (25 to 29oF).
- After much discussion, the decision was taken to go ahead.
- Cause of the accident:
- An inability to assess the link between cool temperature and O-Ring damage on earlier flights.
- **Many charts poorly presented**

Visualization

- Refers to the innovative use of images and interactive technology to explore large, high- density datasets
- Help users see patterns and relationships that would be difficult to see in text lists
  - Rich graphs, charts
  - Dashboards
  - Maps
- Increasingly is being used to identify insights into both structured and unstructured data for such areas as
  - operational efficiencies
  - profitability
  - strategic planning

Video Tableau

```
http://www.bu.edu/students/life/
http://www.bu.edu/students/life/housing/
http://www.bu.edu/students/life/dining-vending/
http://www.bu.edu/students/life/phone/
http://www.bu.edu/students/life/safety/
http://www.bu.edu/students/life/transportation/
http://www.bu.edu/students/life/activities/
http://www.bu.edu/students/life/bu-global-orientation-registratior
http://www.bu.edu/students/academics/
http://www.bu.edu/students/academics/link/
http://www.bu.edu/students/academics/admissions/
http://www.bu.edu/students/academics/registration/
http://www.bu.edu/students/academics/advising/
http://www.bu.edu/students/academics/grades/
http://www.bu.edu/students/academics/services/
http://www.bu.edu/students/academics/support/
http://www.bu.edu/students/health/
http://www.bu.edu/students/health/services/
http://www.bu.edu/students/health/counseling/
http://www.bu.edu/students/health/facilities/
http://www.bu.edu/students/health/clubsports/
http://www.bu.edu/students/health/varsitysports/
```

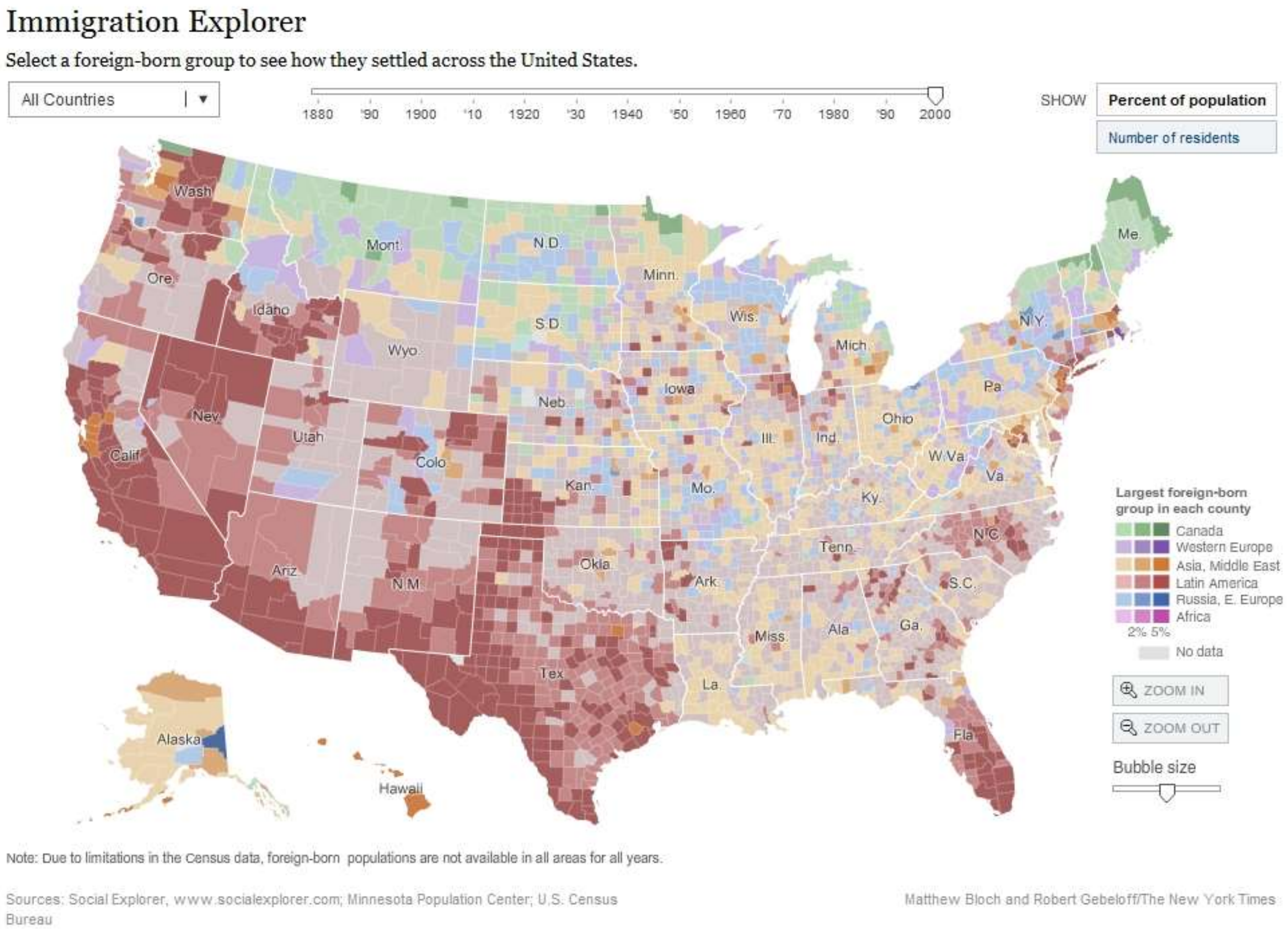


Sales by Region, Product Category & SubCategory

		Central	West	South	East	Total
Furniture	Tables	$471,751	$454,887	$316,405	$652,965	$1,896,008
	Chairs & Chairmats	$651,654	$348,052	$292,478	$469,652	$1,761,837
	Bookcases	$258,919	$246,411	$171,504	$145,818	$822,652
	Office Furnishings	$259,389	$159,443	$129,434	$149,828	$698,094
	Total	$1,641,713	$1,208,793	$909,820	$1,418,264	$5,178,591
fice Supplies	Storage & Organization	$299,116	$227,534	$263,166	$280,367	$1,070,183
	Binders & Access.	$309,262	$203,847	$214,942	$294,907	$1,022,958
	Appliances	$317,079	$133,946	$149,023	$136,944	$736,992
	Other Office Supplies	$149,590	$133,918	$91,955	$100,714	$476,178
	Paper	$150,710	$98,576	$100,210	$96,958	$446,453
	Total	$1,225,757	$797,821	$819,295	$909,889	$3,752,762
Tech	Office Machines	$563,395	$673,390	$610,807	$321,105	$2,168,697
	elephones & Comm.	$613,410	$475,653	$405,524	$394,726	$1,889,314
	Copiers and Fax	$404,175	$343,117	$209,237	$173,833	$1,130,361
	Computer Peripherals	$250,718	$150,974	$195,535	$198,649	$795,876
	Total	$1,831,698	$1,643,134	$1,421,104	$1,088,313	$5,984,248
	Grand Total	$4,699,167	$3,649,748	$3,150,219	$3,416,466	$14,915,601

Examples

- Geo data mapping



Immigration Explorer

Select a foreign-born group to see how they settled across the United States.

Note: Due to limitations in the Census data, foreign-born populations are not available in all areas for all years.

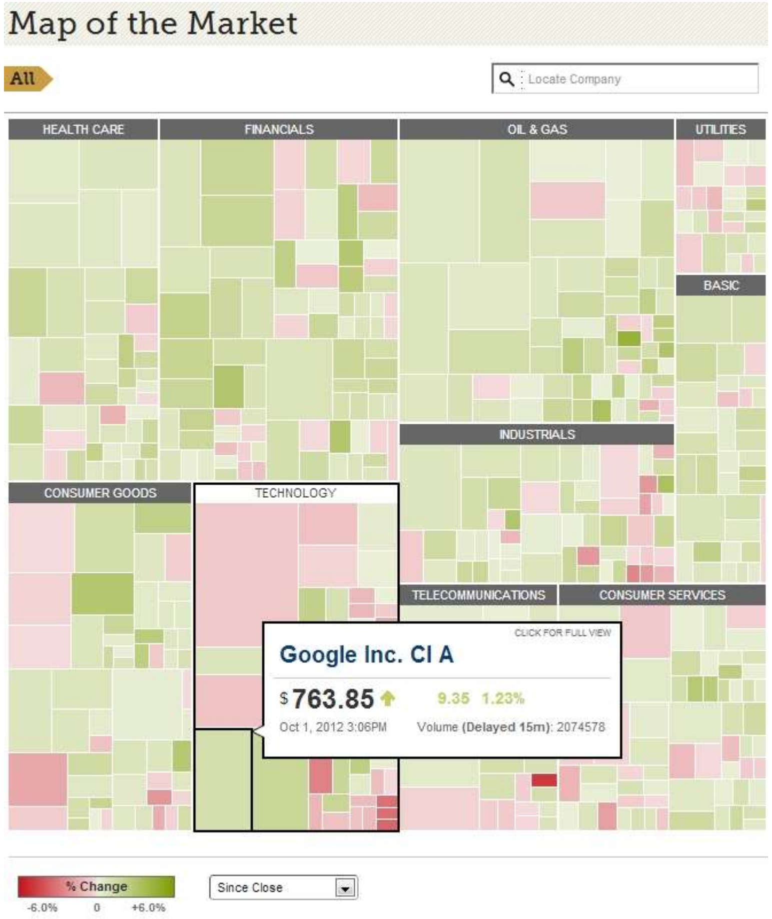
Sources: Social Explorer, www.socialexplorer.com; Minnesota Population Center; U.S. Census Bureau

Matthew Bloch and Robert Gebeloff/The New York Times

- Demo

Examples

- Treemap



Google Inc. Cl A

$763.85 9.35 1.23%

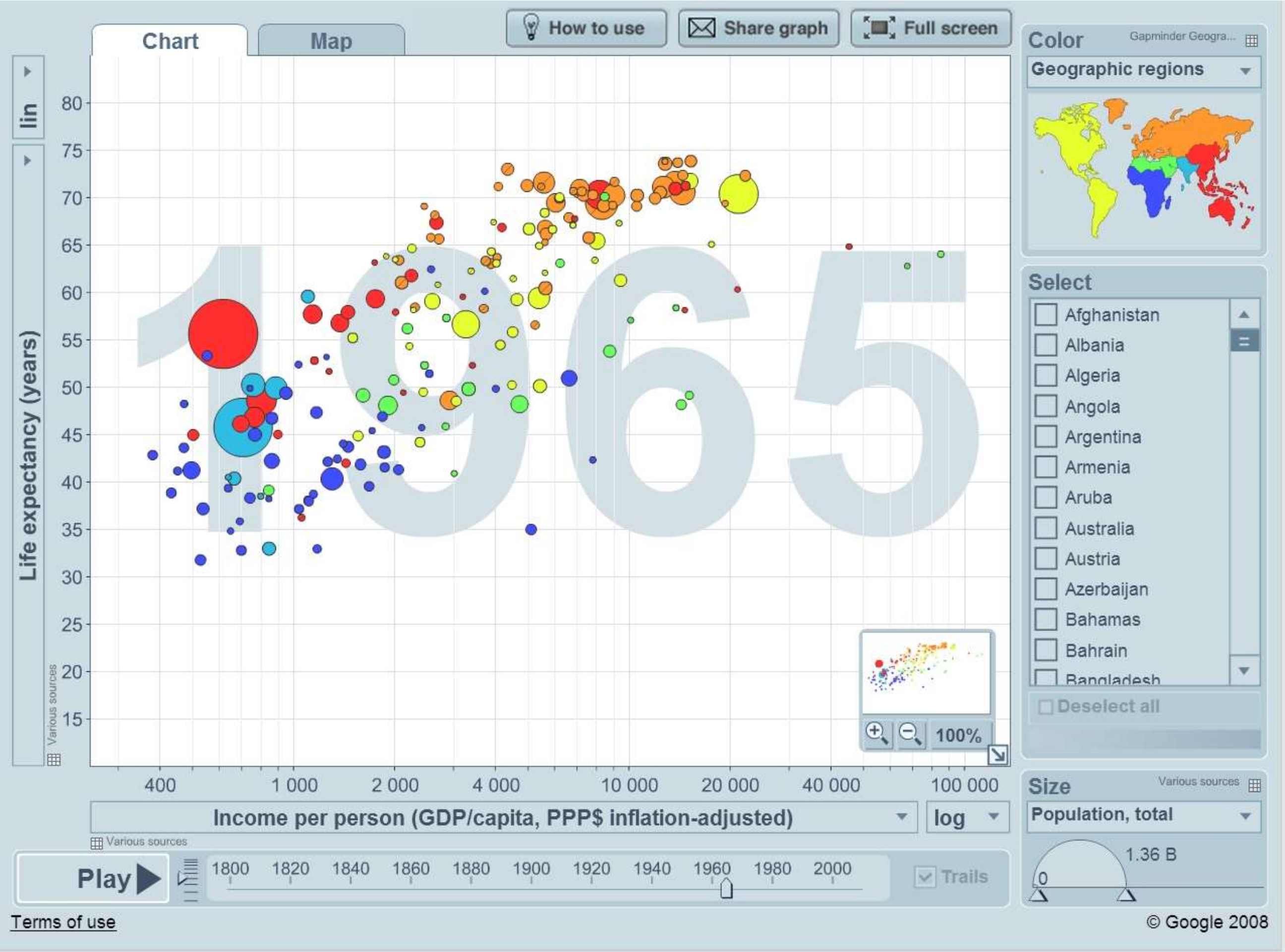
Oct 1, 2012 3:06PM Volume (Delayed 15m): 2074578

- Demo

Examples

- Population

  “Trendalyzer”



Income per person (GDP/capita, PPP$ inflation-adjusted) log

Color: Geographic regions (Gapminder Geogra...)

Select
- □ Afghanistan
- □ Albania
- □ Algeria
- □ Angola
- □ Argentina
- □ Armenia
- □ Aruba
- □ Australia
- □ Austria
- □ Azerbaijan
- □ Bahamas
- □ Bahrain
- □ Bangladesh

□ Deselect all

Size: Population, total (Various sources)

- Demo

- Video on the use of visualization to extract knowledge from data.
    - Watch Gary Flake on extreme visualization

MODULE 4

Data Mining

What is Data Mining?

- The process of semi-automatically analyzing large databases to find useful patterns (Silberschatz)
- Areas of Use
  - Sales/ Marketing
    - Diversify target market
    - Identify clients needs to increase response rates
  - Risk Assessment
    - Identify Customers that pose high credit risk
  - Fraud Detection
    - Identify people misusing the system. E.g. People who have two Social Security Numbers
    - Credit Card Fraud Detection
  - Detect significant deviations from normal behavior:
    - Network Intrusion Detection
  - Customer Care
    - Identify customers likely to change providers
    - Identify customer needs
  - Medicine
    - Match patients with similar problems → cure

Data Mining Techniques...

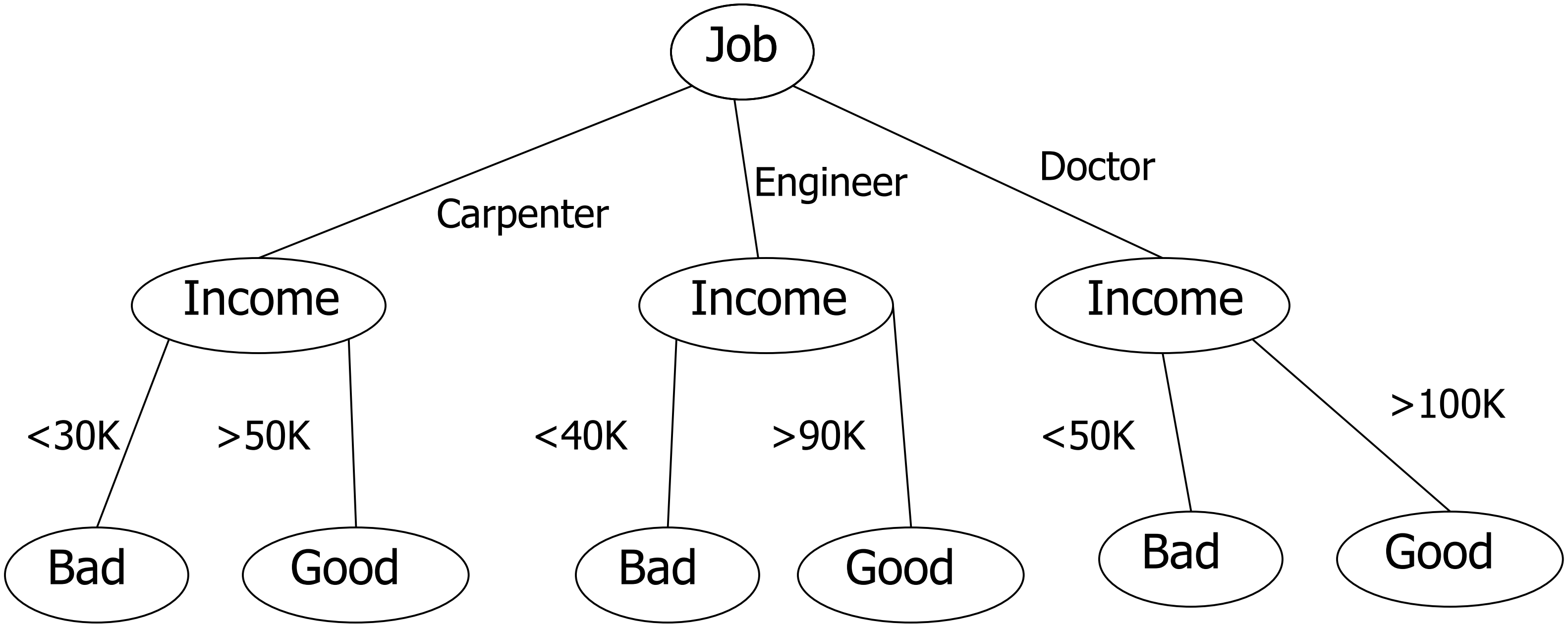
- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]

Classification

- Classification is the process of predicting the class of a new item.
- Categorize the new item and identify to which class it belongs
- Example: A bank wants to classify its Home Loan Customers into groups according to their response to bank advertisements. The bank might use the classifications “Responds Rarely, Responds Sometimes, Responds Frequently”.
- The bank will then attempt to find rules about the customers that respond Frequently and Sometimes.
- The rules could be used to predict needs of potential customers.

Technique for Classification

- Decision-Tree Classifiers



Predicting credit risk of a person with the jobs specified.

Classification: Application 1

- Direct Marketing
  - Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.
  - Approach:
    - Use the data for a similar product introduced before.
    - We know which customers decided to buy and which decided otherwise. This *{buy, don't buy}* decision forms the *class attribute*.
    - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
    - Type of business, where they stay, how much they earn, etc.
    - Use this information as input attributes to learn a classifier model.

Classification: Application 2

- Fraud Detection
  - Goal: Predict fraudulent cases in credit card transactions.
  - Approach:
    - Use credit card transactions and the information on its account-holder as attributes.
      - When does a customer buy, what does he buy, how often he pays on time, etc.
    - Label past transactions as fraud or fair transactions. This forms the class attribute.
    - Learn a model for the class of the transactions.
    - Use this model to detect fraud by observing credit card transactions on an account.

Classification: Application 3

- Customer Attrition/Churn:
  - Goal: To predict whether a customer is likely to be lost to a competitor.
  - Approach:
    - Use detailed record of transactions with each of the past and present customers, to find attributes.
      - How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
    - Label the customers as loyal or disloyal.
    - Find a model for loyalty.

Clustering

- Clustering algorithms find groups of items that are similar. … It divides a data set so that records with similar content are in the same group, and groups are as different as possible from each other.

- Example: Insurance company could use clustering to group clients by their age, location and types of insurance purchased.

- The categories are unspecified and this is referred to as ‘unsupervised learning’

Clustering continued

- Group data into clusters
  - Similar data is grouped in the same cluster
  - Dissimilar data is grouped in the a different cluster

- How is this achieved ?
  - Hierarchical
    - Group data into t-trees
  - K-Nearest Neighbor
    - A classification method that classifies a point by calculating the distances between the point and points in the training data set. Then it assigns the point to the class that is most common among its k-nearest neighbors (where k is an integer)

Clustering: Application 1

- Document Clustering:
  - Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
  - Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
  - Gain: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.

Clustering: Application 2

- Market Segmentation:
  - Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
  - Approach:
    - Collect different attributes of customers based on their geographical and lifestyle related information.
    - Find clusters of similar customers.
    - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Association Rule: Definition

- Given a set of records each of which contain some number of items from a given collection;
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.
- Example: When a customer buys a hammer, then 90% of the time they will buy nails.

TID	*Items*
1	**Bread, Coke, Milk**
2	**Beer, Bread**
3	**Beer, Coke, Diaper, Milk**
4	**Beer, Bread, Diaper, Milk**
5	**Coke, Diaper, Milk**

Rules Discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

Association Rule Discovery: Application 1

- Marketing and Sales Promotion:
  - Let the rule discovered be

    {Bagels, … } --> {Potato Chips}
  - **Potato Chips as consequent** => Can be used to determine what should be done to boost its sales.
  - **Bagels in the antecedent** => Can be used to see which products would be affected if the store discontinues selling bagels.
  - **Bagels in antecedent *and* Potato chips in consequent** => Can be used to see what products should be sold with Bagels to promote sale of Potato chips!

Association Rule Discovery: Application 2

- Supermarket shelf management.
  - Goal: To identify items that are bought together by sufficiently many customers.
  - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
  - A classic rule --
    - If a customer buys diaper and milk, then he is very likely to buy beer.
    - So, don’t be surprised if you find six-packs stacked next to diapers!

Association Rule Discovery: Application 3

- Inventory Management:
  - Goal: A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicles equipped with right parts to reduce on number of visits to consumer households.
  - Approach: Process the data on tools and parts required in previous repairs at different consumer locations and discover the co-occurrence patterns.

Sequential Pattern Discovery

- Given is a set of objects, with each object associated with its own timeline of events, find rules that predict strong sequential dependencies among different events.
- Rules are formed by first discovering patterns. Event occurrences in the patterns are governed by timing constraints.

- In telecommunications alarm logs,
  - (Inverter_Problem Excessive_Line_Current)
  - (Rectifier_Alarm) --> (Fire_Alarm)
- In point-of-sale transaction sequences,
  - Computer Bookstore:

    (Intro_To_Visual_C) (C++_Primer) -->
    (Perl_for_dummies,Tcl_Tk)
  - Athletic Apparel Store:

    (Shoes) (Racket, Racketball) --> (Sports_Jacket)

Regression

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Greatly studied in statistics, neural network fields.
- Examples:
  - Predicting sales amounts of new product based on advertising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.

Using Databases to Improve Business Performance and Decision Making

- **Web mining**
  - **Discovery and analysis of useful patterns and information from Web**
    - Understand customer behavior
    - Evaluate effectiveness of Web site, and so on
  - **Web content mining**
    - Mines content of Web pages
  - **Web structure mining**
    - Analyzes links to and from Web page
  - **Web usage mining**
    - Mines user interaction data recorded by Web server

Using Databases to Improve Business Performance and Decision Making

- **Text mining**
  - **Extracts key elements from large unstructured data sets**
    - Stored e-mails
    - Call center transcripts
    - Legal cases
    - Patent descriptions
    - Service reports, and so on
  - **Sentiment analysis software**
    - Mines e-mails, blogs, social media to detect opinions

Big Data, Big Rewards

- Read the case study “**Big Data, Big Rewards**”