

Data Mining and BI

Topic One

Fundamental concepts

Tibebe B. (PhD)

Topics

- ▶ Fundamental concepts and the need for business intelligence, data mining , big data analysis
- ▶ BI/DM/BDA applications, DA models and frameworks
- ▶ Data and data warehousing
- ▶ Data mining/ analytics techniques
 - ▶ Association rule mining,
 - ▶ Classification
 - ▶ Cluster analysis
 - ▶ Others: sentiment mining, Big data and BI technologies, applications,
- ▶ Current trends in Data Analytics, data mining and BI

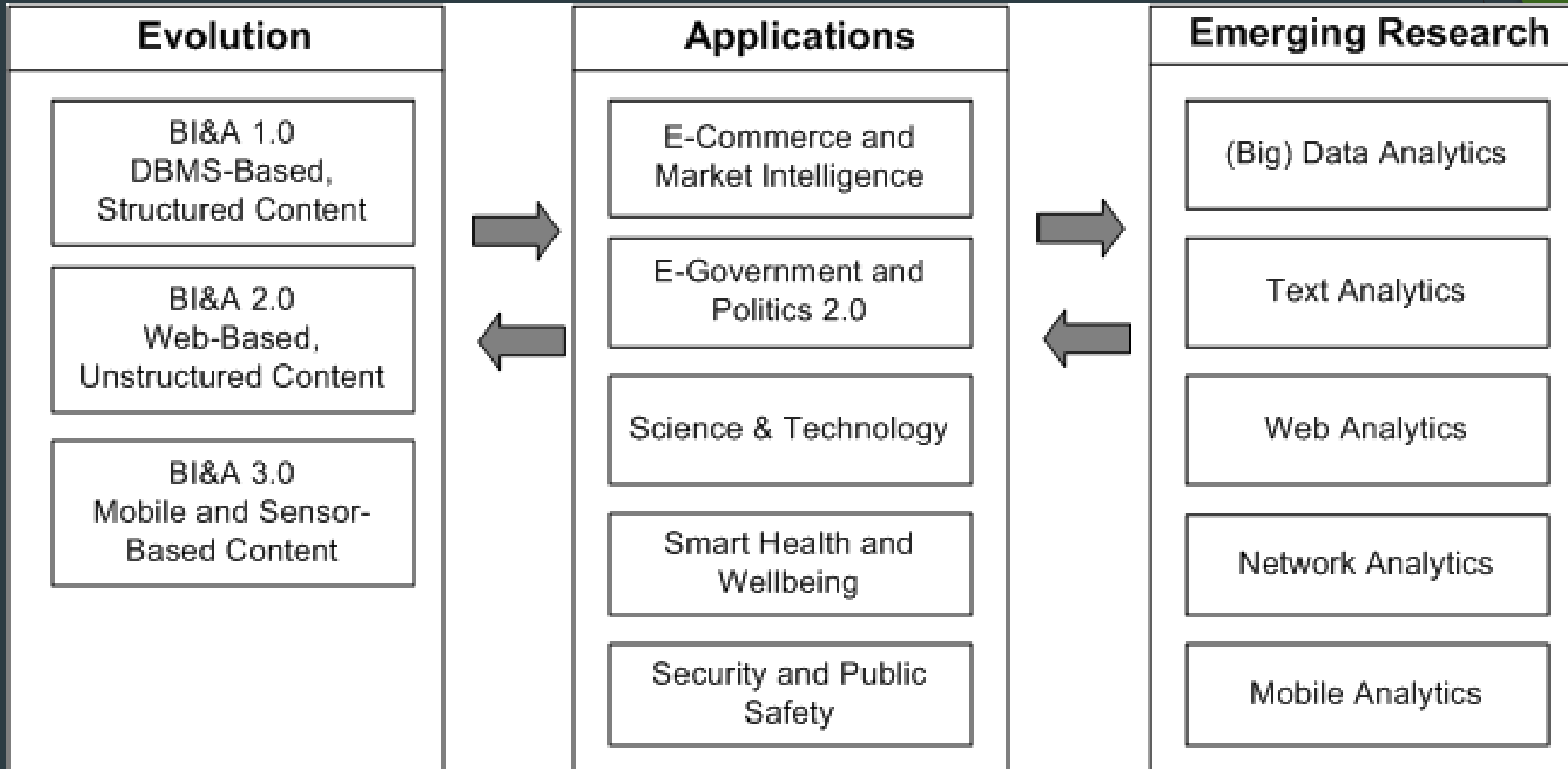
BI /DS/BD A/BA/Data Mining/Web
mining

Basic concepts

Business and its analytics

- ▶ According to the 2009 IBM Global CIO study, BI and analytics is the number one priority for chief information officers (CIOs)
- ▶ *Key terms*
 - ▶ *Business*
 - ▶ *Analytics*
- ▶ *Business intelligence and analytics (BI&A) has emerged as an important area of study for both practitioners and researchers, reflecting the magnitude and impact of data-related problems to be solved in contemporary business organizations*

BI&A Overview: Evolution, Applications, and Emerging Research



Business intelligence (BI)

▶ Business

- ▶ Any meaningful set of activists in achieving a goal/ objective

▶ Intelligence

- ▶ Sensing content/environment and act/decide accordingly

- ▶ BI- is a set of theories, methodologies, architectures, and technologies that **transform raw data into meaningful and useful information for business purposes.**

- ▶ BI can handle enormous amounts of unstructured data to help identify, develop and otherwise create new opportunities.

- ▶ BI, in simple words, **makes interpreting voluminous data friendly.** Making use of new opportunities and implementing an effective strategy can provide a competitive market advantage and long-term stability.

A Framework for Business Intelligence (BI)

- ▶ **Business intelligence (BI)** is a conceptual framework for decision support.
- ▶ BI's Architecture and major Components
 - ▶ Data Analytics
 - ▶ Data/Data Warehouse
 - ▶ Data Analytics/data mining/BDA
 - ▶ Business
 - ▶ Performance and Strategy
 - ▶ User Interface

Cont...

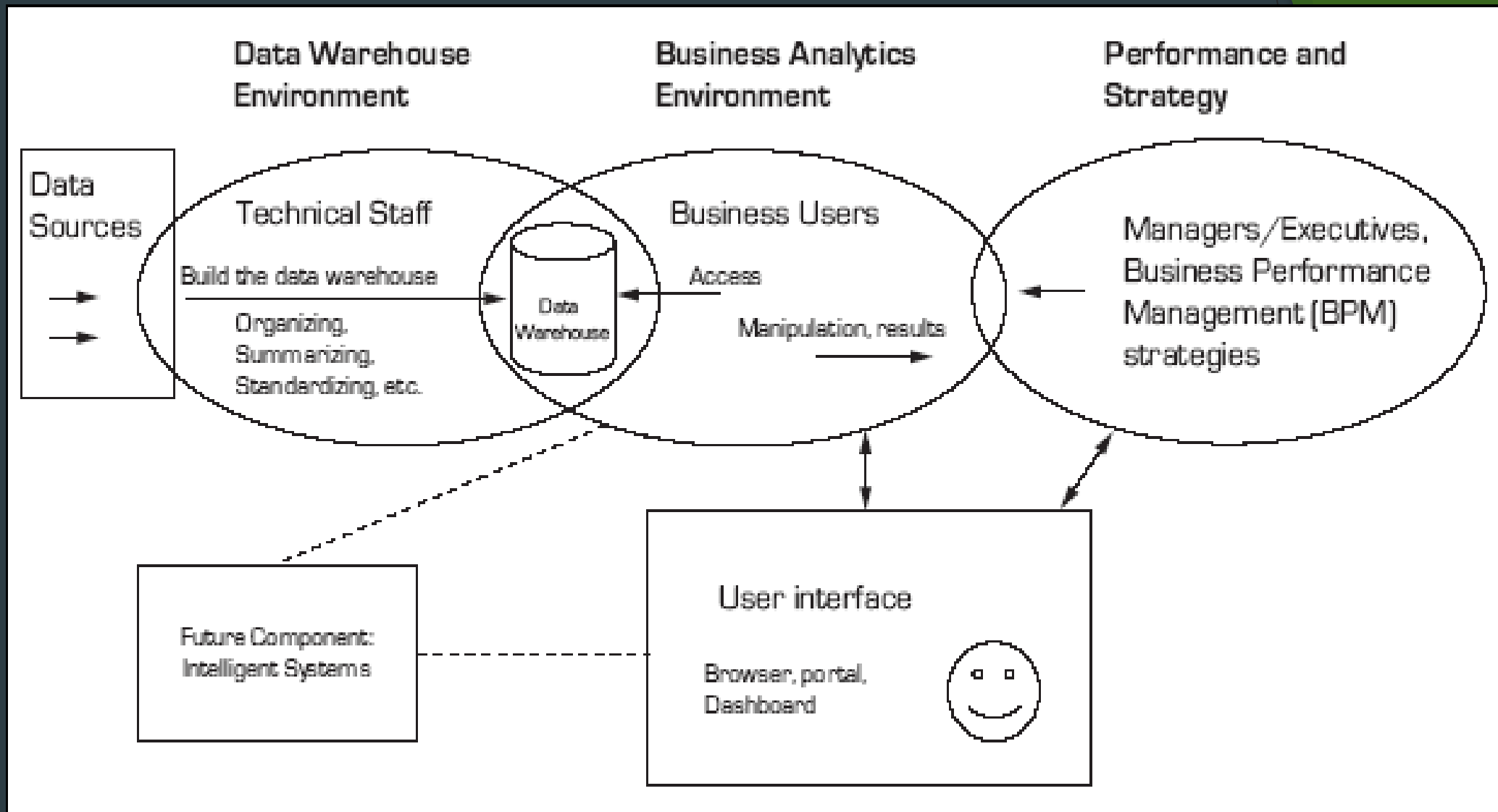
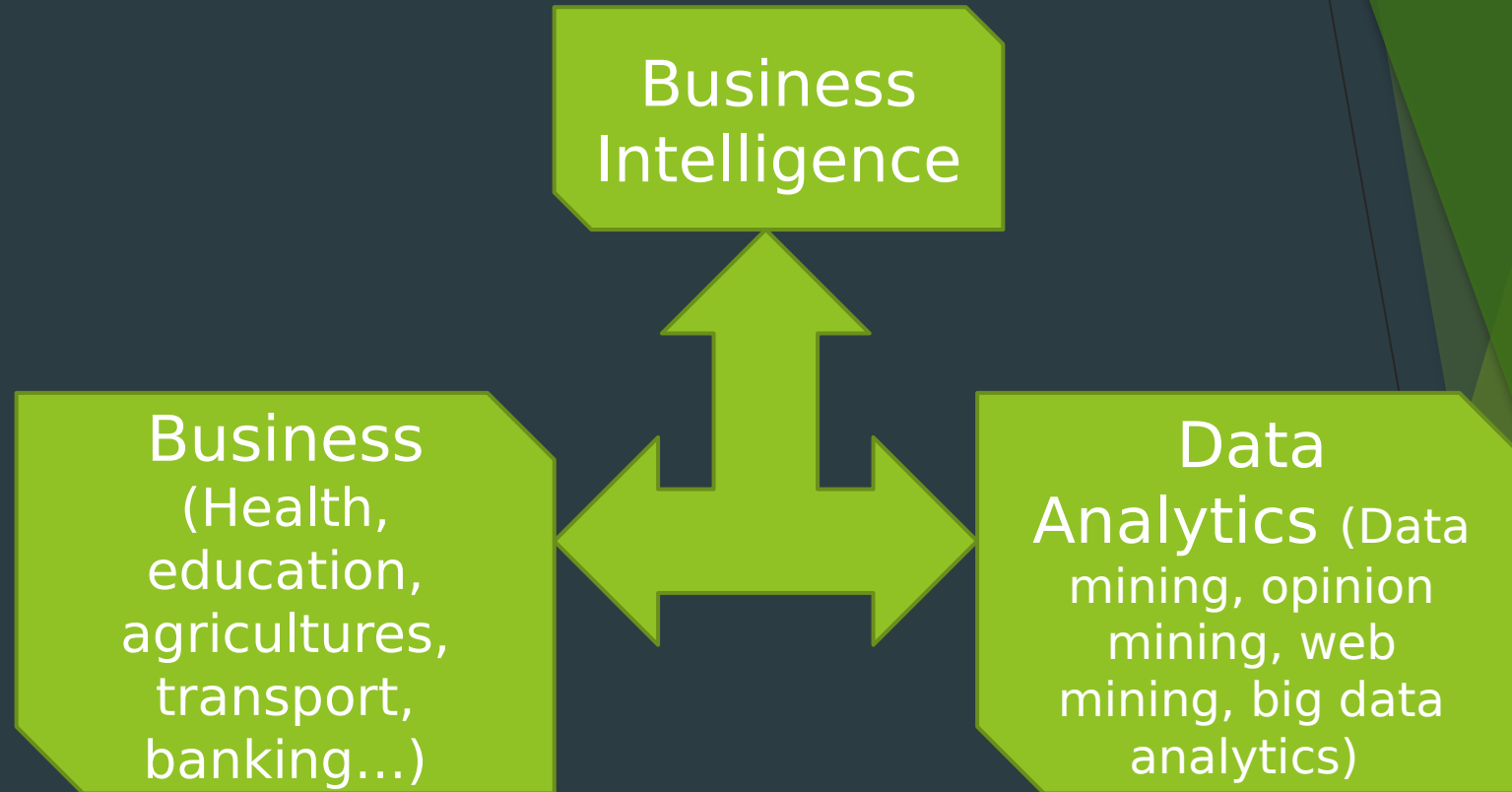


FIGURE 1.3 A High Level Architecture for BI

Conceptual Model



Data science

- ▶ The terms “data science” and “data mining” often are used interchangeably.
- ▶ At a high level, **data science** is a set of fundamental principles that guide the extraction of knowledge from data.
 - ▶ **Data mining** is the extraction of knowledge from data, via technologies that incorporate these principles.
- ▶ As a term, “data science” often is applied more broadly than the traditional use of “data mining”, but data mining techniques provide some of the clearest illustrations of the principles of data science.

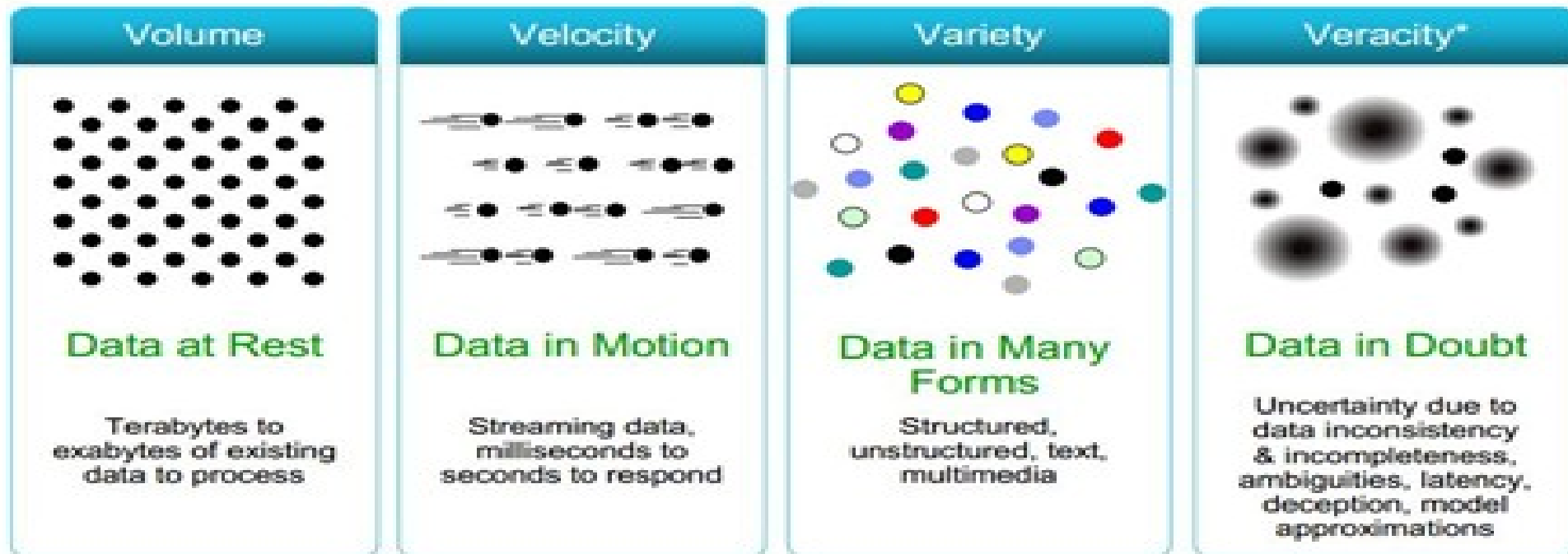
Big Data

- ▶ **Big data** is the **term** and **technology** 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.
- ▶ *“Big data is high-**volume**, high-**velocity** and high-**variety** information assets that demand **cost-effective**, **innovative** forms of information processing for **enhanced insight and decision making**.” -- Gartner*

Cont...

- There are many examples of "data", but what makes some of it "big"? The classic definition revolves around the **three Vs**.
- **Volume, velocity, and variety.**
 - **Volume:** There is just a lot of it being generated all the time. Things get interesting and "big", when you can't fit it all on one computer anymore. Why? There are many ideas here such as MapReduce, Hadoop, etc. that all revolve around being able to process data that goes from Terabytes, to Petabytes, to Exabytes.
 - **Velocity:** Data is being generated very quickly. Can you even store it all? If not, then what do you get rid of and what do you keep?
 - **Variety:** The data types you mention all take different shapes. What does it mean to store them so that you can play

Some make it 4 v's



Characteristics of big data

Volume



Data at scale

Terabytes to
petabytes of data

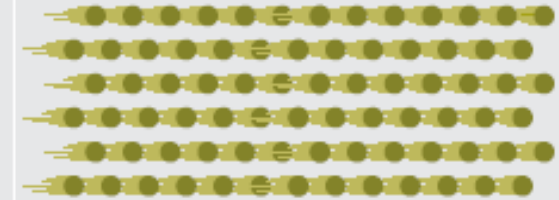
Variety



Data in many forms

Structured, unstructured,
text, multimedia

Velocity



Data in motion

Analysis of streaming data
to enable decisions within
fractions of a second

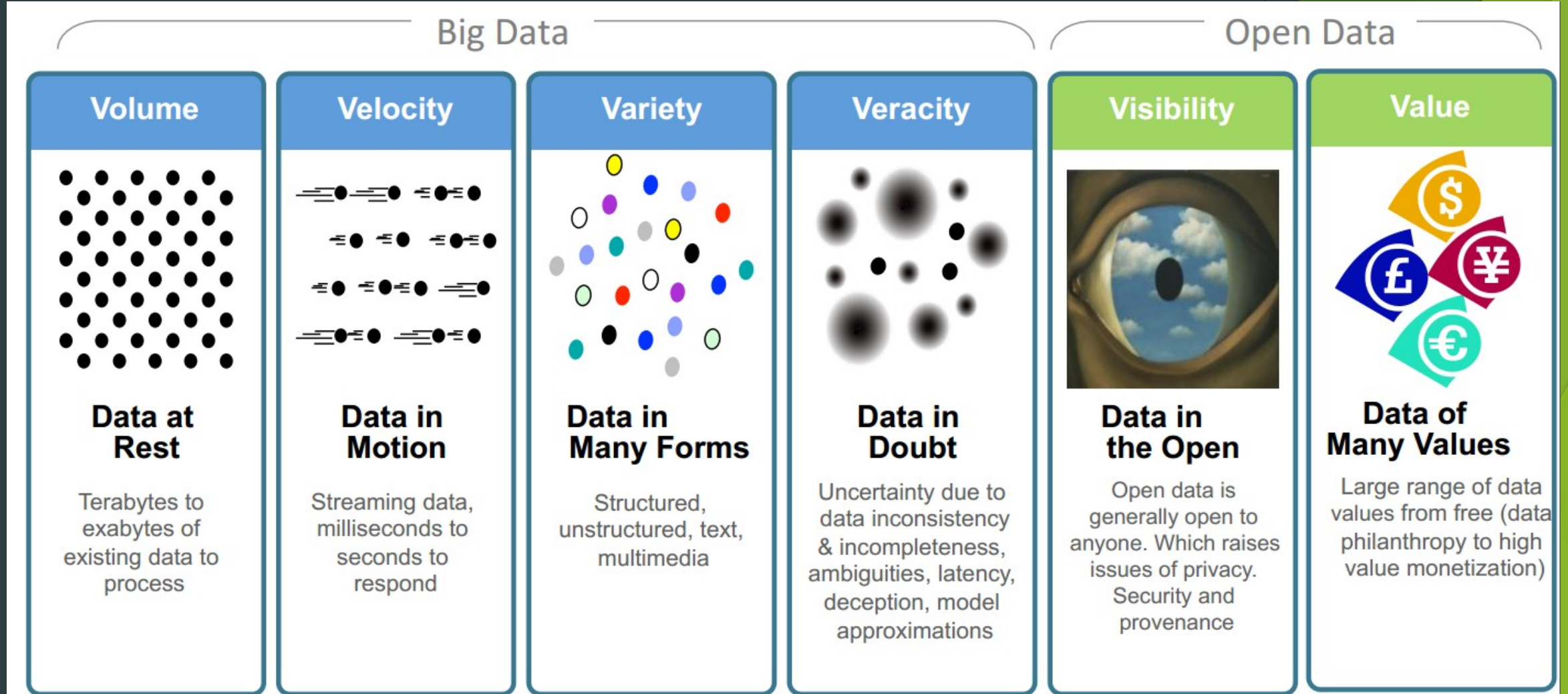
Veracity



Data uncertainty

Managing the reliability and predictability
of inherently imprecise data types

Big Data: 6V



Who is generating Big Data?

Social



User Tracking & Engagement



Homeland Security



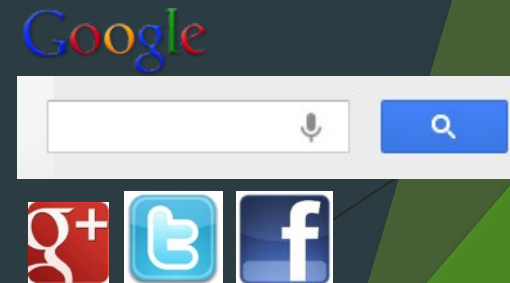
eCommerce



Financial Services



Real Time Search



Data Mining

- **Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns**
- Definition (Fayyad et. al): The non-trivial discovery of *novel, valid, comprehensible* and *potentially useful* patterns from data.
- What is a pattern? A relationship in the data. E.g.,
 - On Thursday nights people who buy diapers also tend to buy beer
 - People with good credit ratings are less likely to have accidents
 - Male consumers, 37+, income bracket 50K-75K □ spend between \$25-\$50 per catalog order

Web Mining

- Discovering interesting and useful information from Web *content* and *usage*
- *Three types: Web usage, Web structure, Web content*

Application Examples:

- ▶ Web search, e.g. Google, Yahoo, MSN, Ask, ...
- ▶ eCommerce :
 - ▶ Recommendations: e.g. Netflix, Amazon
- ▶ Advertising, e.g. Google AdSense
- ▶ Improving Web site design and performance

How does it differ from “classical” Data Mining?

- ▶ The web is not a relation
 - ▶ Textual information and linkage structure
- ▶ Usage data is huge and growing rapidly
 - ▶ Google’s usage logs are bigger than their web crawl
 - ▶ Data generated per day is comparable to largest conventional data warehouses
- ▶ The content is diversified
- ▶ Ability to react in real-time to usage patterns
 - ▶ No human in the loop

Opinion mining

- ▶ A field of study with many names
 - ▶ Sentiment analysis
 - ▶ Sentiment mining
 - ▶ Subjectivity detection
 - ▶ ...
- ▶ Often used synonymously
- ▶ Some shadings in meaning
- ▶ “**sentiment analysis**” describes the current mainstream task best

Cont....

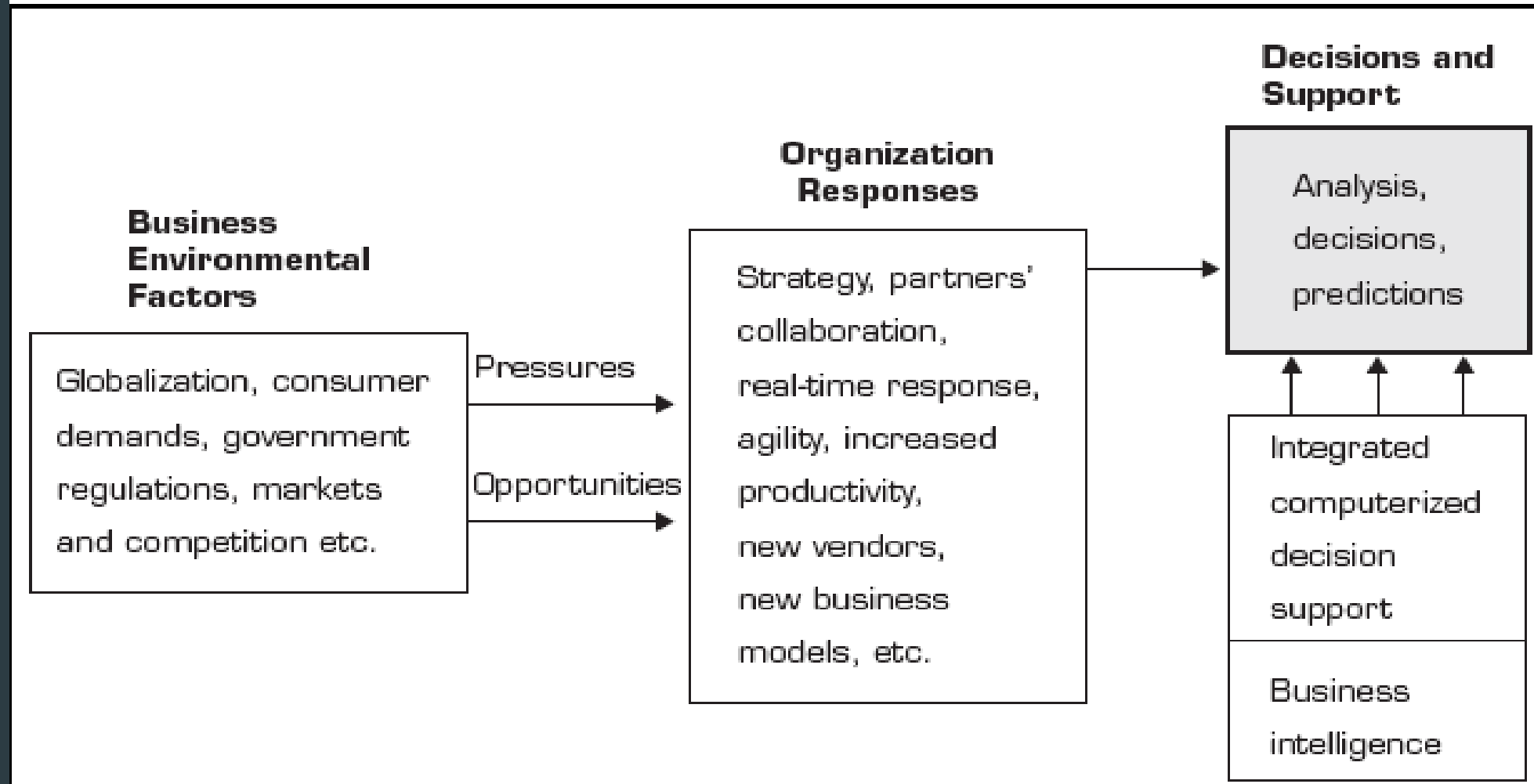
- ▶ **Opinion mining** (also known as **sentiment analysis**) refers to the use of **natural language processing, text analysis, computational linguistics, and biometrics** to systematically identify, extract, quantify, and study affective states and subjective information.
- ▶ Sentiment analysis is widely applied to **voice of the customer** materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine
- ▶ Generally speaking, sentiment analysis aims to determine the attitude of a speaker, writer, or other subject with respect to some topic or the overall contextual polarity or emotional reaction to a document, interaction, or event.

Motivation (Business intelligence/, Data science and Big data, web mining, data mining, opinion mining)

Drivers

Motivation for BI- Changing Business Environments and Computerized Decision Support

FIGURE 1.1 The Business Pressures–Responses–Support Model



Cont...

- ▶ **The Origins and Drivers of Business Intelligence is thus;**
 - ▶ Organizations are being compelled to **capture, understand, and harness their data** to support decision making in order to improve business operations
 - ▶ Managers need the *right information* at the *right time* and in the *right place*

Why BI/Data Analytics in general?

- ▶ Lots of data is being collected and warehoused
 - ▶ Web data, e-commerce
 - ▶ purchases at department/grocery stores
 - ▶ Bank/Credit Card transactions
- ▶ Computers have become cheaper and more powerful
- ▶ **Competitive Pressure** is Strong
 - ▶ Provide better, customized services for an *edge* (e.g. in Customer Relationship Management)

Some difference- Statistics Vs DM/DA

Statistics	Data Analytics/Mining
Confirmative	Explorative
Small data sets/ File-based	Large data sets/ Databases
Small number of variables	Large number of variables
Deductive	Inductive
Numeric data	Numeric and non-numeric (including text, networks)
Clean data	Data cleaning

Database Processing vs. Data Mining as a data analytics Process

	Database	Data analytics	Comments
Query	Well defined Structured Query Language	• Poorly defined • No precise query language	The data miner might not know what he exactly wants to see
Data	Operational data	Not Operational data	The data have been cleansed and modified to better support the mining process
Output	Precise and Subset of database	Not a subset of database	The output is some hidden useful patterns & knowledge in the database

Review questions

- ▶ Define Business intelligence, data science, data mining, Big data, opinion mining and web mining
- ▶ Explain some basic motivations for data analytics in general
- ▶ Develop a justification for big data within an organization
- ▶ Identify three “big data” sources, either within or external to your organization that would be relevant to your business.
- ▶ Explain the difference b/n DM and Database

Summary

- ▶ Conceptualizations of key concepts
 - ▶ BI, data science , data mining, web mining, opinion mining
- ▶ Motivation for BI/data analytics
- ▶ Difference between DA/DM and statistics, data base systems

Thank you