

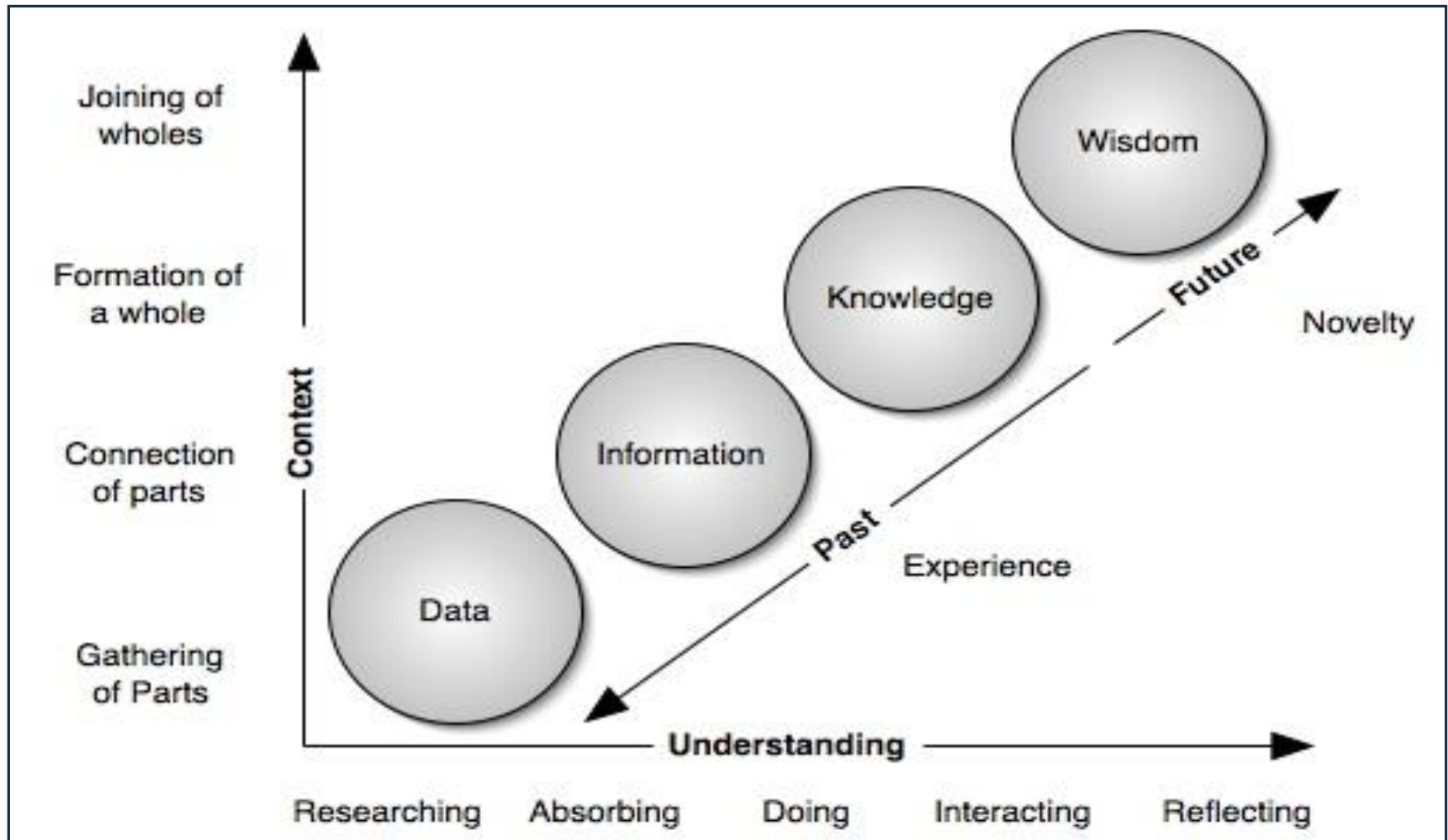
# Agenda

- **Housekeeping**
- **Lecture 1 :**
  - **Intro to data Mining**
  - **Intro to Probability**
- **R down loads**

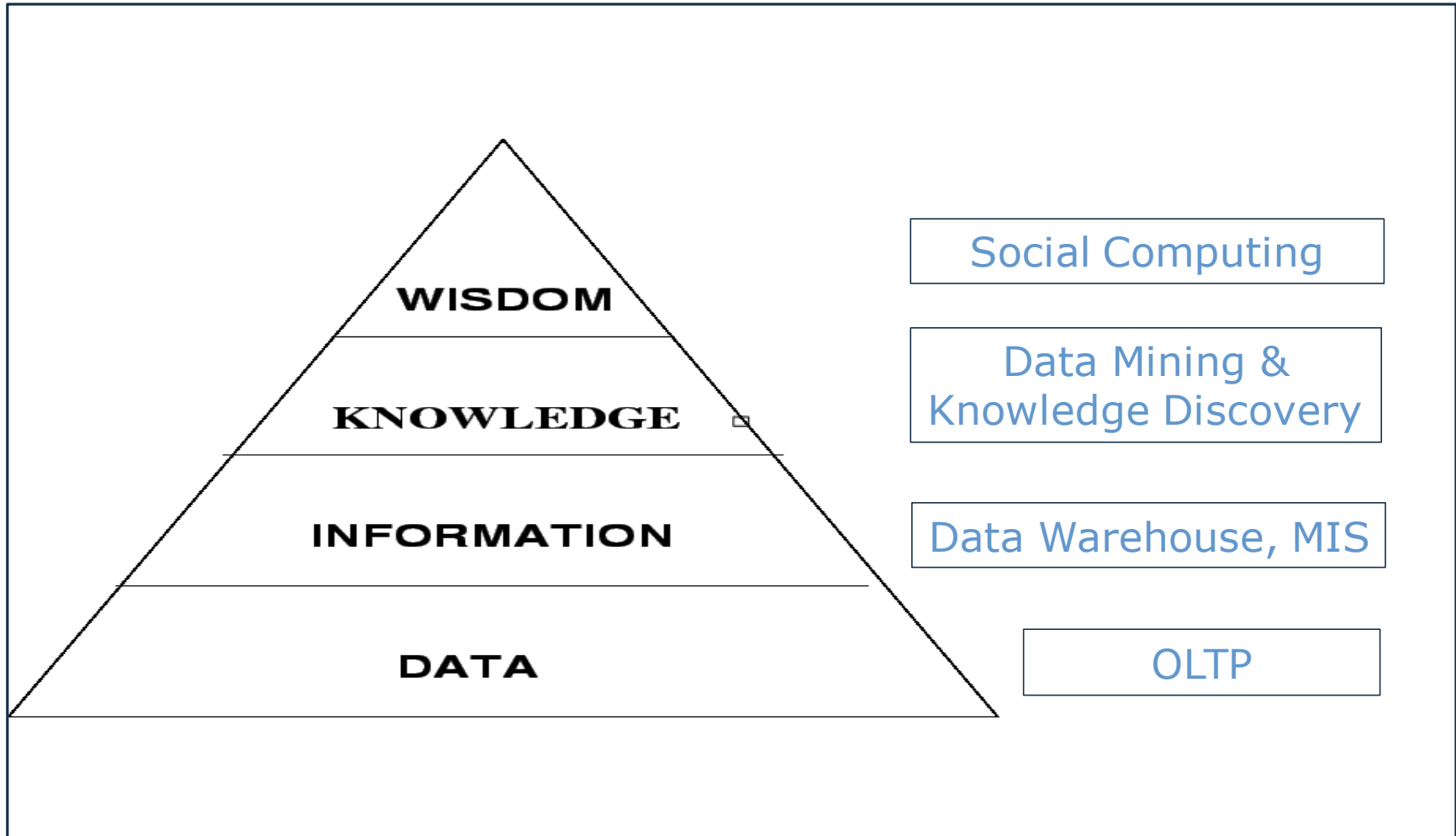
# Definitions

- Data
  - Representations of Facts
- Information
  - Data with “Relevance and Importance”
  - Data that changes the probability of a relevant outcome.
- Knowledge
  - Ability to use information to act (or not), in order to achieve objectives.
- Wisdom
  - Ability to synthesize information and knowledge, to create a framework for optimal actions.

# What are Data, Information, Knowledge, & Wisdom?



# Support Systems In a Typical Organization



# Evolution of Technology

- **1960s**
  - **Data collection, database creation, IMS and network DBMS**
- **1970s:**
  - **Relational data model, relational DBMS implementation**
- **1980s:**
  - **RDBMS, advanced data models (extended-relational, OO, deductive, etc.)**
  - **Application-oriented DBMS (spatial, scientific, engineering, etc.)**
- **1990s:**
  - **Data mining, data warehousing, multimedia databases, and Web databases**
- **2000s**
  - **Stream data management and mining**
  - **Data mining with a variety of applications**
  - **Web technology and global information systems**

# Data Explosion Problem (“Big” Data)

- Yahoo generates mountains of data daily.
- NASA Earth Observing System (EOS) is projected to generate 50GB of image data hourly.
- Wal-Mart built an 11-Tbyte database of customer transactions (on 700M unique store/item product combinations, 2,900 stores, 52 weeks/year).
- 3.6 Zettabytes ( $\text{ZB} = 10^{21}$  bytes) of data were consumed by individuals in the U.S. in 2008.

***How much is 3.6 zettabytes? If we printed 3.6 zettabytes of text in books, and stacked them as tightly as possible across the United States including Alaska, the pile would be 7 feet high***

# Data Explosion Problem (“Big” Data)

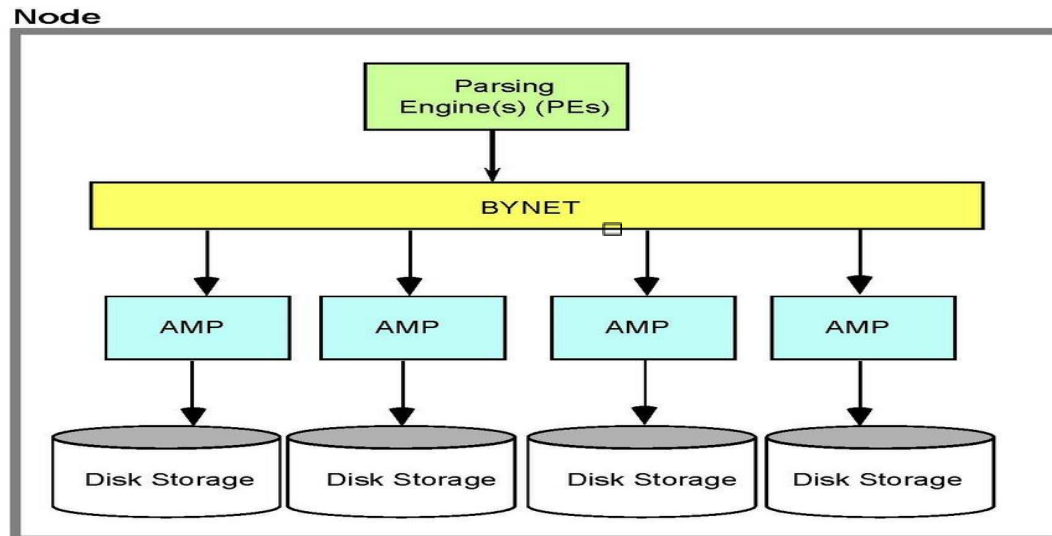
## The AT&T Network Handles:

- ~ 18.7 Petabytes of traffic every business day (One Petabyte is the equivalent of the 100 complete printed collections of the U.S. Library of Congress)
- ~ 2 Billion voice calls per day
- ~ 32 billion text messages traveled over the AT&T’s wireless during the 4th quarter of 2007

# How to Get Information Out of "Big" Data

## New Data Warehouse Architectures

### Major Components of a Teradata System





# How to Get Knowledge Out of “Big” Data

There is a need for a new generation of techniques with the ability to *intelligently and automatically* assist humans in analyzing ‘mountains’ of data for nuggets of useful knowledge (and not just information).

This has led to an emerging field:

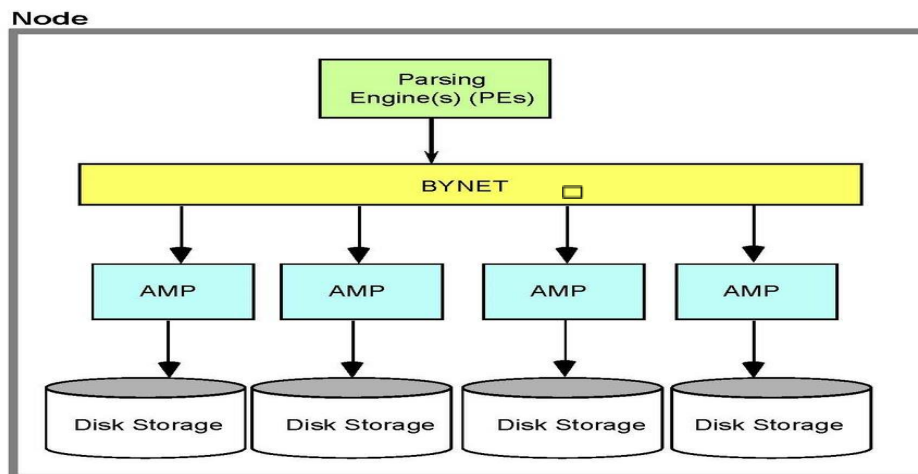


Data Mining & Knowledge Discovery (DM & KD)

# How to Get Knowledge Out of “Big” Data

## Equip the Data Warehouse with Intelligent Algorithms

### Major Components of a Teradata System



# What is Data Mining & Knowledge Discovery ?

## DM & KD Mean Different Things to Different Professionals

- Management: Potentially money making tools
- Computer Scientists: A new Knowledge Discovery breakthrough - NOT STATISTICS
- Statisticians: Not statistically, significantly, new - A computerized statistician
- Electrical Engineers: Another application of Information Theory and Entropy
- Neuroscientists: Neurocomputer - a computer model of the human brain
- Mathematicians: Some weighted average of a bunch of numbers

# Data Mining & Knowledge Discovery

- Underlying Disciplines  
Biology, Neurology, Psychology, Statistics, Computer Science, Engineering
- Artificial Intelligence (AI)  
Integrates the “Underlying Disciplines” for solving various types of problems
- Techniques
  - Symbolic: *Rules Based Systems (RBS)*, *Case-Based Reasoning (CBR)*, *Fuzzy Logic (FL)*
  - Connectionist: *Artificial Neural Networks (ANN)*
  - Inductive (ML): *C4.5*, *CART*
  - Evolutionary: *Genetic Algorithms (GA)*

# What is Data Mining & Knowledge Discovery?

The non-trivial *process* of identifying *valid*, *novel*, potentially *useful*, and ultimately *understandable* patterns in data.

-- *Fayad, Shapiro, Smyth (1996)*

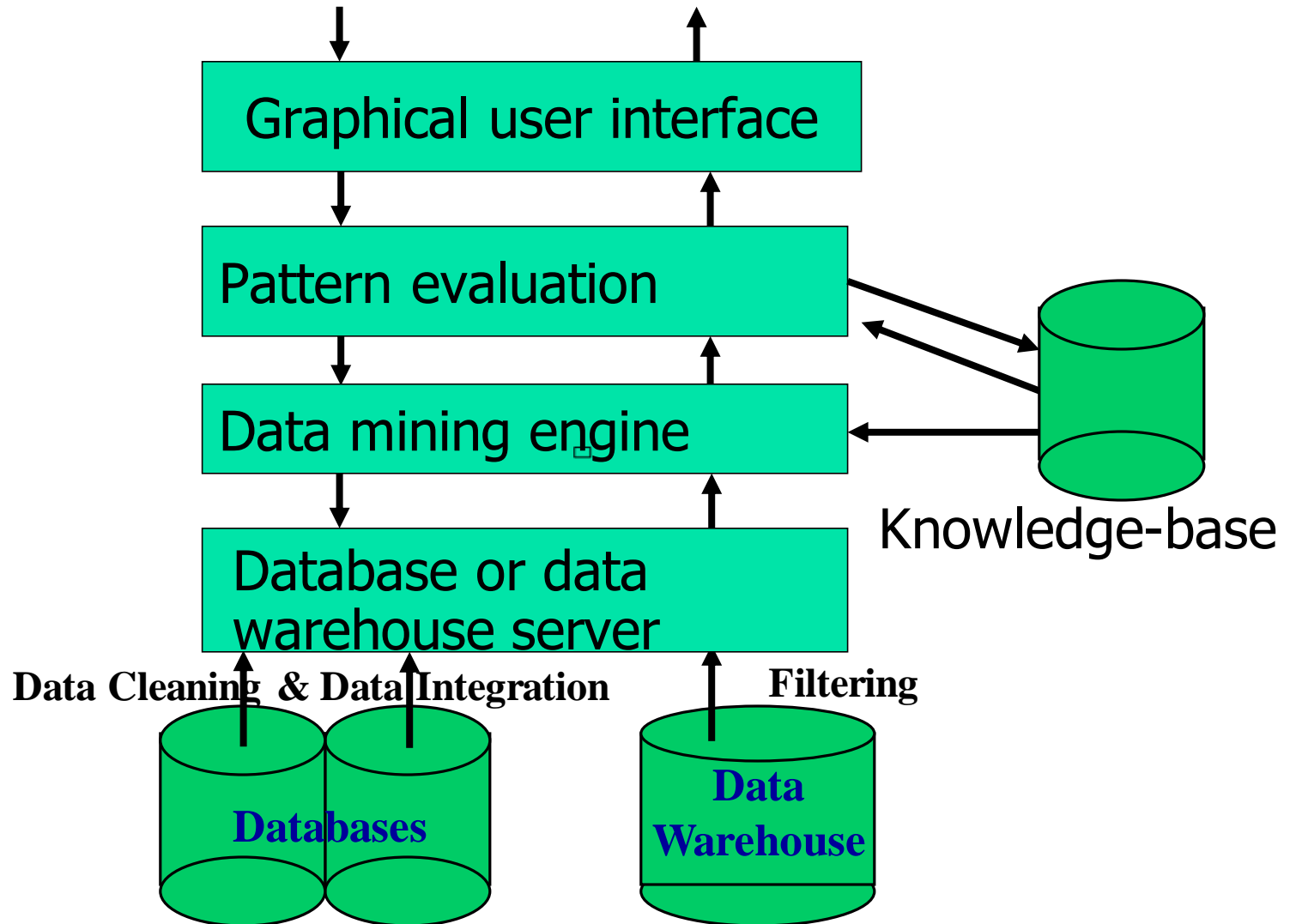
- ***process***: knowledge discovery is iterative, as you uncover “nuggets” in the data, you learn to ask better questions
- ***valid***: generalize to the future
- ***novel***: not something we already know
- ***useful***: actionable, can be used for a task
- ***understandable***: process leads to human insight

# What is Data Mining & Knowledge Discovery ?

The New York Times:

Data mining has entered a golden age, whether being used to set ad prices, find new drugs more quickly or fine-tune financial models. Companies as diverse as Google, Pfizer, Merck, Bank of America, the InterContinental Hotels Group and Shell use it.

# Architecture: Typical Data Mining System



# DM & KD Process: End-to-End Solution

- Pose a Profound Question
- Identify Relevant Data
- Access the Data
- Clean the Data
- Transform & Integrate the Data
- Mine/Discover Knowledge
- Make Intelligent Decisions