

Agenda

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

Definitions

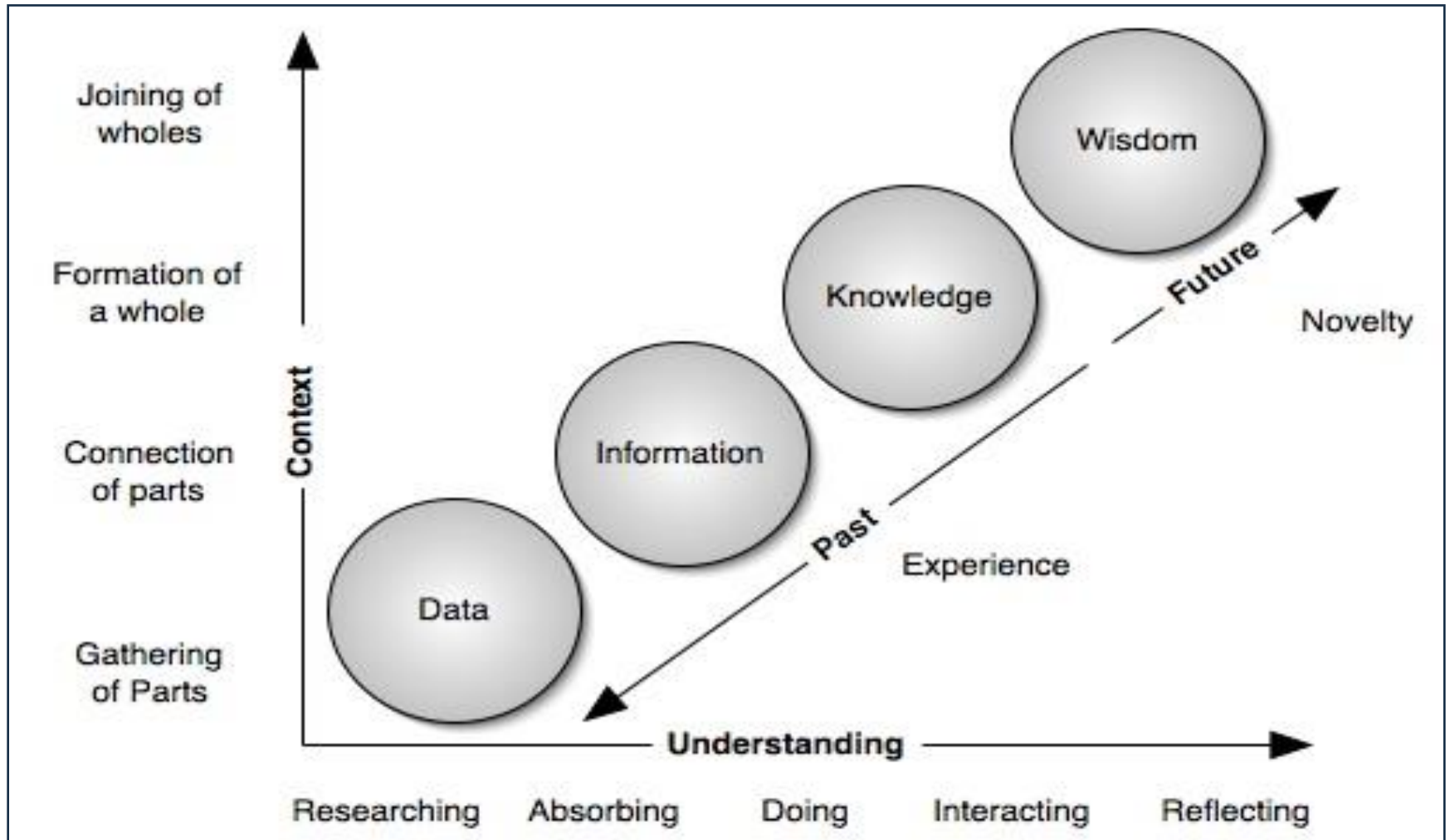
- Data
 - Representations of Facts

- Information
 - Data with “Relevance and Importance”
 - Any datum (and/or data) that changes the probability distribution (chances) of a relevant outcome.

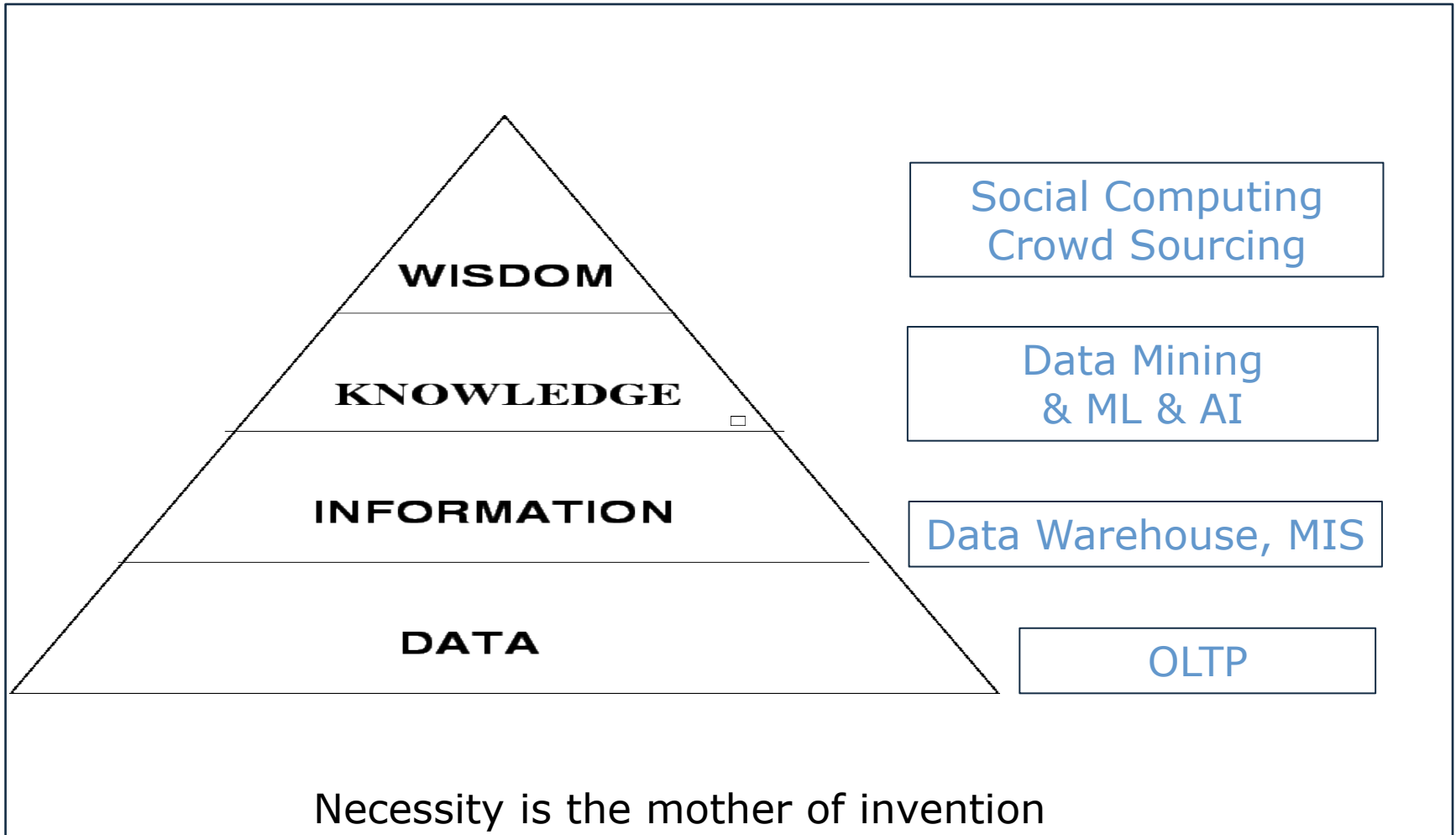
Definitions

- Knowledge
 - ▶ Ability to use information to act (or not), in order to achieve objectives.
 - ▶ The ability to understand and explain, relationship between different phenomena (usually as a rule)
- Wisdom
 - ▶ Ability to synthesize information and knowledge, to create a framework for optimal actions.
- Intelligence
 - ▶ The ability to apply knowledge

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 and ML with a variety of applications**
 - **Web technology and global information systems**

Data Explosion Problem ("Big" Data)

- Snapchat users share 527,760 photos
- More than 120 professionals join LinkedIn
- Users watch 4,146,600 YouTube videos
- 456,000 tweets are sent on Twitter
- Instagram users post 46,740 photos

Data Explosion Problem (“Big” Data)

- **Snapchat users share 527,760 photos**
- **More than 120 professionals join LinkedIn**
- **Users watch 4,146,600 YouTube videos**
- **456,000 tweets are sent on Twitter**
- **Instagram users post 46,740 photos**

Source: <https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#6cc1dc7d60ba>

Data Explosion: Facebook

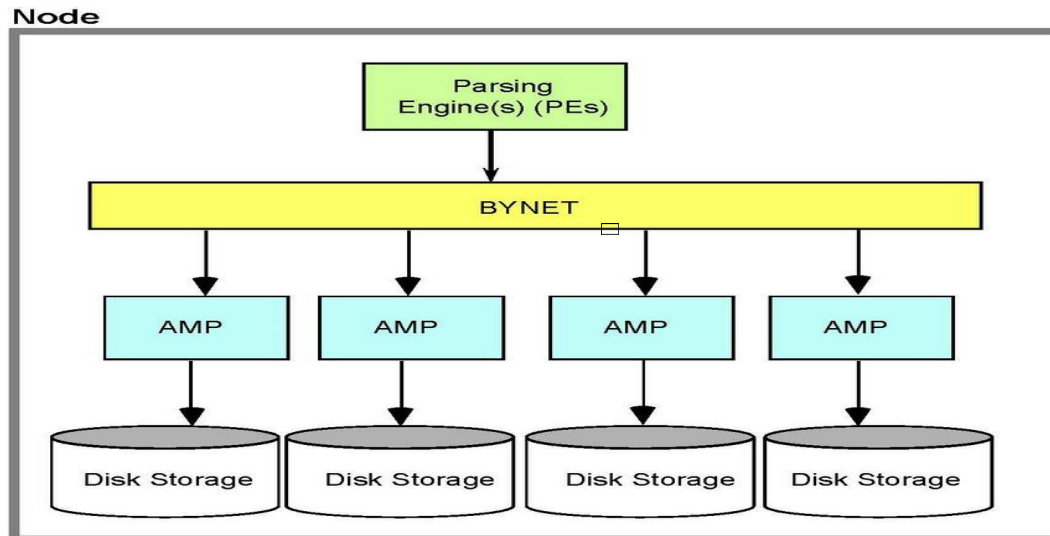
- **1.5 billion people are active on Facebook daily**
- **Europe has more than 307 million people on Facebook**
- **There are five new Facebook profiles created every second!**
- **More than 300 million photos get uploaded per day**
- **Every minute there are 510,000 comments posted and 293,000 statuses updated**

<https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#6cc1dc7d60ba>

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, ML & Knowledge Discovery (DM & KD)

DM vs. ML vs. AI vs. DL

- **Data Mining :**

- ▶ finding patterns in data to explain some phenomenons. (e.g. loan default)

- **Machine Learning:**

- ▶ enable machine to "learn" using statistical methods(e.g. Netflix)

- **Artificial Intelligence:**

- ▶ create ways that machine can mimic human behaviors. (e.g. Deep Blue)

- **Deep Learning:**

- ▶ creating a machine that mimics the working of our brains (image recognition).

- Source: Linked in

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)*

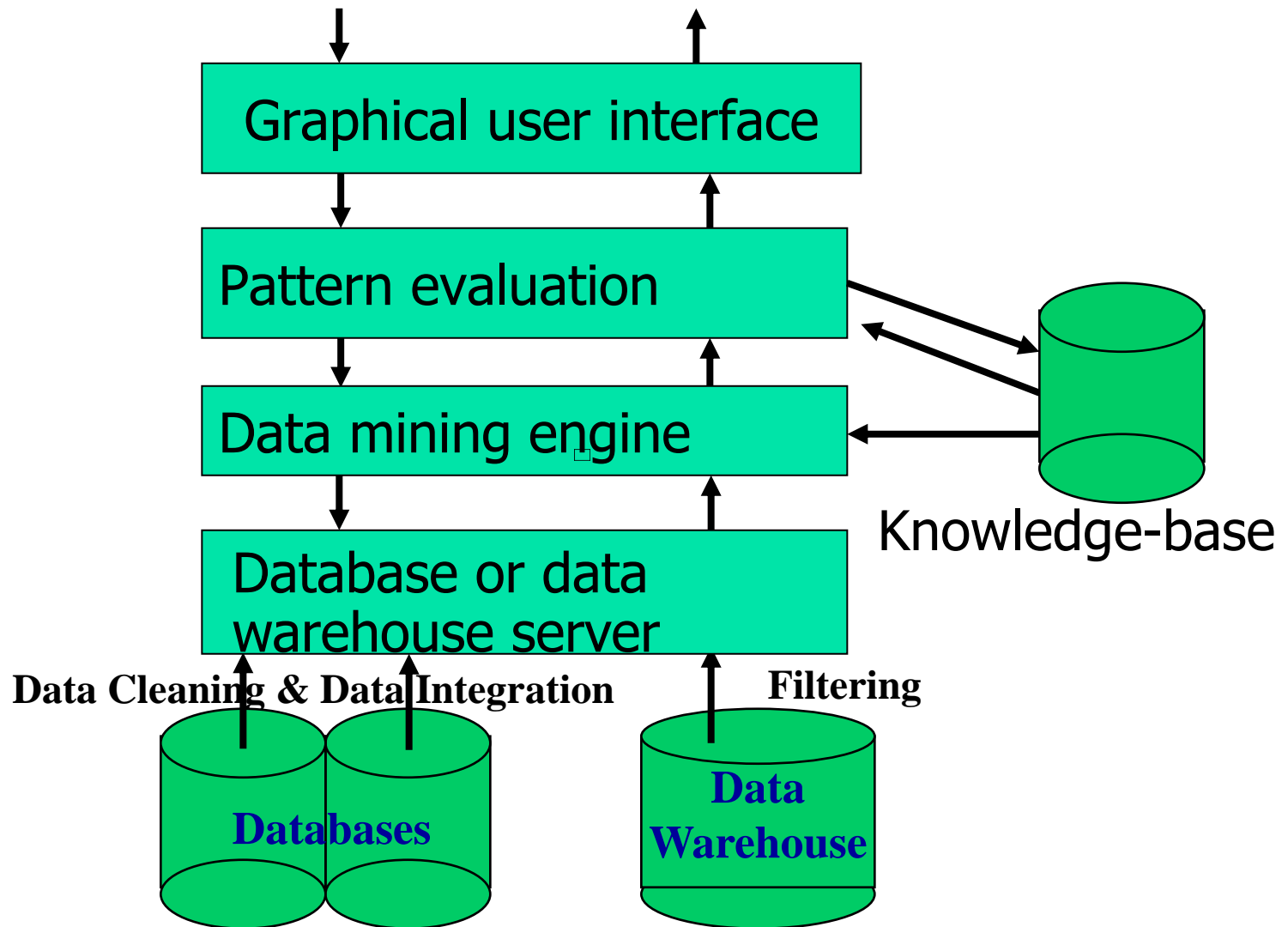
- ***process***: knowledge discovery is iterative, as you uncover “nuggets” in the data, you learn to ask better questions
- ***valid***: generalize to the future
- ***novel***: not something we already know
- ***useful***: actionable, can be used for a task
- ***understandable***: process leads to human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

Intelligence Chiefs Testify At Senate Hearing

- <https://www.youtube.com/watch?v=70VVbrTP18g> **40 minute**