CSCI 5408 Winter, 2017

Part II:

Business Data Analytics – DM & DW Technologies

(Week 8-13)

Part II: Introduction

- Instructor: Dr. Qigang Gao (q.gao@dal.ca)
  - Class time:
    - Section-1: TR, 2:35-3:55PM, LSC 208
    - Section-2: MW, 6:05-7:25PM, CS 127
  - Office hours: MR, 12:30-2:00PM, CS 219
  - Course management systems:
    - *Brightspace:* submission & learning material, etc.
    - *Bluenose/prof5408/Doc:* DM & DW these examples, etc.
- TA: Abhinav Kalra (abhinav.kalra@dal.ca)
  - **Ass 4-Tutorial**: Mar 1, 1:00-2:30PM, MANAG 1020
  - **Ass 4-Help Hours**: Wed/Fri, 1:00-2:30PM, CS 134/CS 233

Important Dates

- **Final Exam:** Apr 20, 3:30-5:30 PM
- **Deadlines:** Mar 14 (A4), Mar 28 (A5), Apr 11 (A6)
- **Tutorials:** Wed, 1:00-2:30PM, MANAG 1020
  - Ass4-Tut: Mar 01, ETL/DW/OLAP
  - Ass5-Tut: Mar 15, Association DM
  - Ass6-Tut: Mar 29, Classification DM

Learning Materials

- Textbook:
  - "***Data Mining: Concepts and Techniques***" (3rd Edition, 2011). (*The 2nd Edition, Textbooks & References/DM&DW Textbook.pdf)
  - A reference book example: *"Data Mining: The Textbook"*, 2015 (Textbooks & References//DM Reference Book.pdf).
- Lecture slides & assignments: via Brightspace
- Web resource examples:
  - General Data Mining Site: http://www.kdnuggets.com
  - Data Mining Software - Weka: http://www.cs.waikato.ac.nz/ml/weka/
  - Database Repository: http://www.ics.uci.edu/~mlearn/MLRepository.html
  - Wikipedia: http://en.wikipedia.org/wiki/Data_mining

Part II Outline

Overview: (Week 8)

1. Introduction: Overview on DM&DW ***Ass4:*** *ETL/DW/OLAP*

2. Data preprocessing

DW & OLAP: (Week 9)

3. Data warehousing and OLAP

Basic DM Tasks & Algorithms:

4. Association pattern mining (Week 10-11) ***Ass5:*** *Association DM*

5. Classification/prediction (Week 11-12) ***Ass6:*** *Classification DM*

6. Clustering analysis (Week 13)

7. Characterization/Generalization (Week 13)

1. Introduction (Text: Ch1)

- **Challenges to information systems**
- **Data, information and knowledge**
- **Traditional DB Technology & Limitations**
- **Why DM and DW becoming important tools of DSS?**
- **DW/OLAP: store and query on summarized data**
- Typical DM tasks and applications
- Three real world examples
- Machine learning's role in DM
- Relationship of statistics and DM, IR vs. DM
- Challenges of DM: 3Vs (Volume, Variety, Velocity), etc.

Challenges to Conventional Information Systems: DBMS

- Data is growing at a phenomenal rate ("the yotta world")
- Data types are getting more mixed and complicated
- Users expect the availability of more sophisticated information and deep hidden knowledge, and for quick and easy access
- The limitations of traditional DBMS and statistics analysis
- Data rich but information poor!
- Who are the drivers (business) and the enablers (new technologies)
- New strategies for Decision Supporting Systems (DSS) or BI (Business Intelligence) system
- How? DM and DW Technologies

UNCOVER HIDDEN INFORMATION AND PATTERNS

What Types of Information We Want from Data?

- A DBMS is an information system of some organization for storing/managing data and querying information from the data.

- What typical types of information are needed by various users, including regular customers, business management people and decision makers, etc.?

E.g.1, Query a DB: Types of Information?

NO.	RISK	CREDIT HISTORY	DEBT	COLLATERAL	INCOME
1.	high	bad	high	none	$0 to $15k
2.	high	unknown	high	none	$15 to $35k
3.	moderate	unknown	low	none	$15 to $35k
4.	high	unknown	low	none	$0 to $15k
5.	low	unknown	low	none	over $35k
6.	low	unknown	low	adequate	over $35k
7.	high	bad	low	none	$0 to $15k
8.	moderate	bad	low	adequate	over $35k
9.	low	good	low	none	over $35k
10.	low	good	high	adequate	over $35k
11.	high	good	high	none	$0 to $15k
12.	moderate	good	high	none	$15 to $35k
13.	low	good	high	none	over $35k
14.	high	bad	high	none	$15 to $35k

Data from credit history of loan applications

Which of the following queries can (or can't) be answered by SQL?

1. What is the risk level of the application with id NO=1?
2. Does the application NO=1 have adequate property as security against loan?
3. What is the average income for each risk category?
4. Among the high risk class, what is credit history distribution?
5. What are the associations between high risk and different income (or credited history, ...) classes?
6. How to determine whether or not a loan application should be rejected (i.e. with a high risk)?

NO.	RISK	CREDIT HISTORY	DEBT	COLLATERAL	INCOME
1	high	bad	high	none	$0 to $15k
2.	high	unknown	high	none	$15 to $35k
3.	moderate	unknown	low	none	$15 to $35k
4.	high	unknown	low	none	$0 to $15k
5.	low	unknown	low	none	over $35k
6.	low	unknown	low	adequate	over $35k
7.	high	bad	low	none	$0 to $15k
8.	moderate	bad	low	adequate	over $35k
9.	low	good	low	none	over $35k
10.	low	good	high	adequate	over $35k
11.	high	good	high	none	$0 to $15k
12.	moderate	good	high	none	$15 to $35k
13.	low	good	high	none	over $35k
14.	high	bad	high	none	$15 to $35k

Data from credit history of loan applications

Types of Information

- Type 1: Information about individual data objects
  - *what, when, where, etc*
  - Search for existing values of records, no calculation involved
- Type 2: Information about aggregations
  - *what happened to the business*
  - Define groupings and apply simple statistics functions
  - Can be single attribute based, or combinations of attributes
- Type 3: Information about patterns
  - *why it happened , and what to happen next*
  - Patterns are regularities or knowledge of a give data set
  - Discover hidden patterns about a concept (or called target), or relationships, or abstractive representation of categories

E.g.2, A relational DB of electronics retailer business, such as the store *Best Buy*:

customer

cust_ID	name	address	age	income	credit_info	...
C1	Smith, Sandy	5463 E. Hastings, Burnaby, BC, V5A 4S9, Canada	21	$27000	1	...
...	...	...	...	...	...	...

item

item_ID	name	brand	category	type	price	place_made	supplier	cost
I3	hi-res-TV	Toshiba	high resolution	TV	$988.00	Japan	NikoX	$600.00
I8	multidisc-CDplay	Sanyo	multidisc	CD player	$369.00	Japan	MusicFront	$120.00
...	...	...	...	...	...	...	...	...

employee

empl_ID	name	category	group	salary	commission
E55	Jones, Jane	home entertainment	manager	$18,000	2%
...	...	...	...	...	...

branch

branch_ID	name	address
B1	City Square	369 Cambie St., Vancouver, BC V5L 3A2, Canada
...	...	...

purchases

trans_ID	cust_ID	empl_ID	date	time	method_paid	amount
T100	C1	E55	09/21/98	15:45	Visa	$1357.00
...	...	...	...	...	...	...

items_sold

trans_ID	item_ID	qty
T100	I3	1
T100	I8	2
...	...	...

works_at

empl_ID	branch_ID
E55	B1
...	...

Classify queries into general types of information retrieval tasks

Q1: List all digital camera products that have a price tag > $300.

Q2: Find who bought 'Samsung 52 inch LED HDTV' on Sep 12 2015.

Q3: Find total sales of LED TV in Dec 2015 in Halifax.

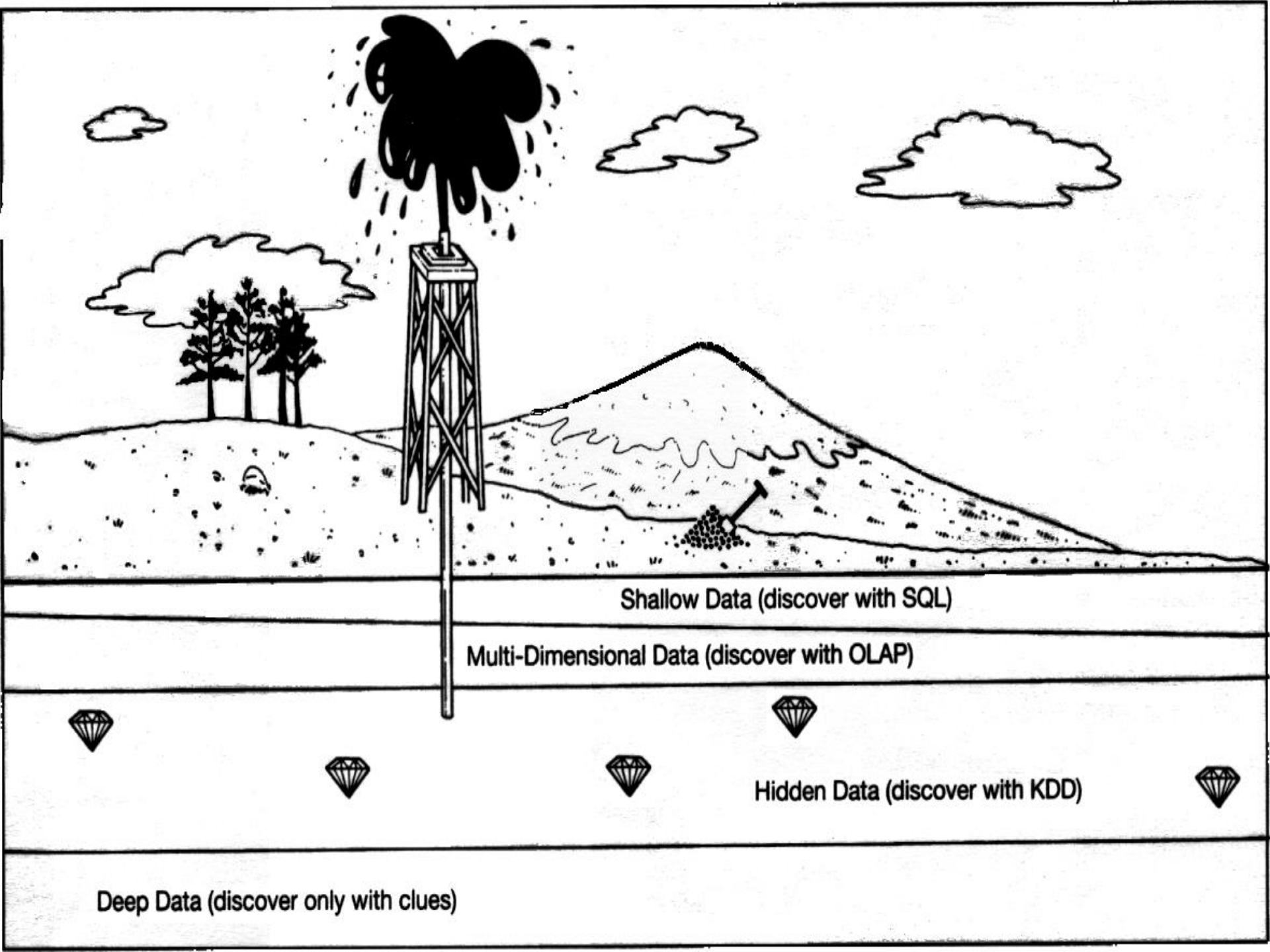
Q4: See each month's sales of different brand LCD and LED TVs from Jan. to Dec in 2014 and 2015 respectively for each city of HRM.

Q5: What are the customers' shopping trends of purchasing digital camera based on the last three years' sales data?

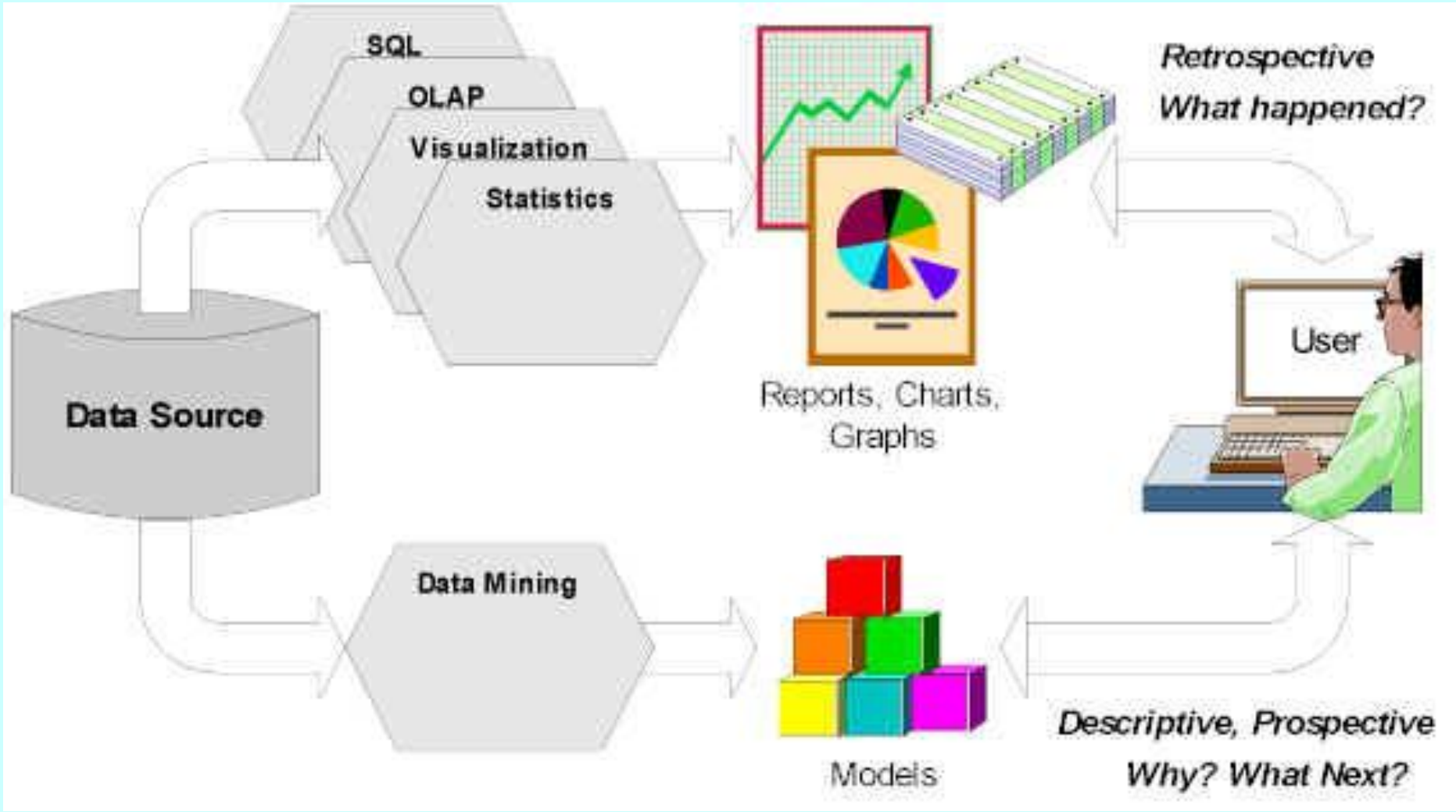
Q6: Find important royalty costumer groups and their profiles.

Types of Information Process for DSS

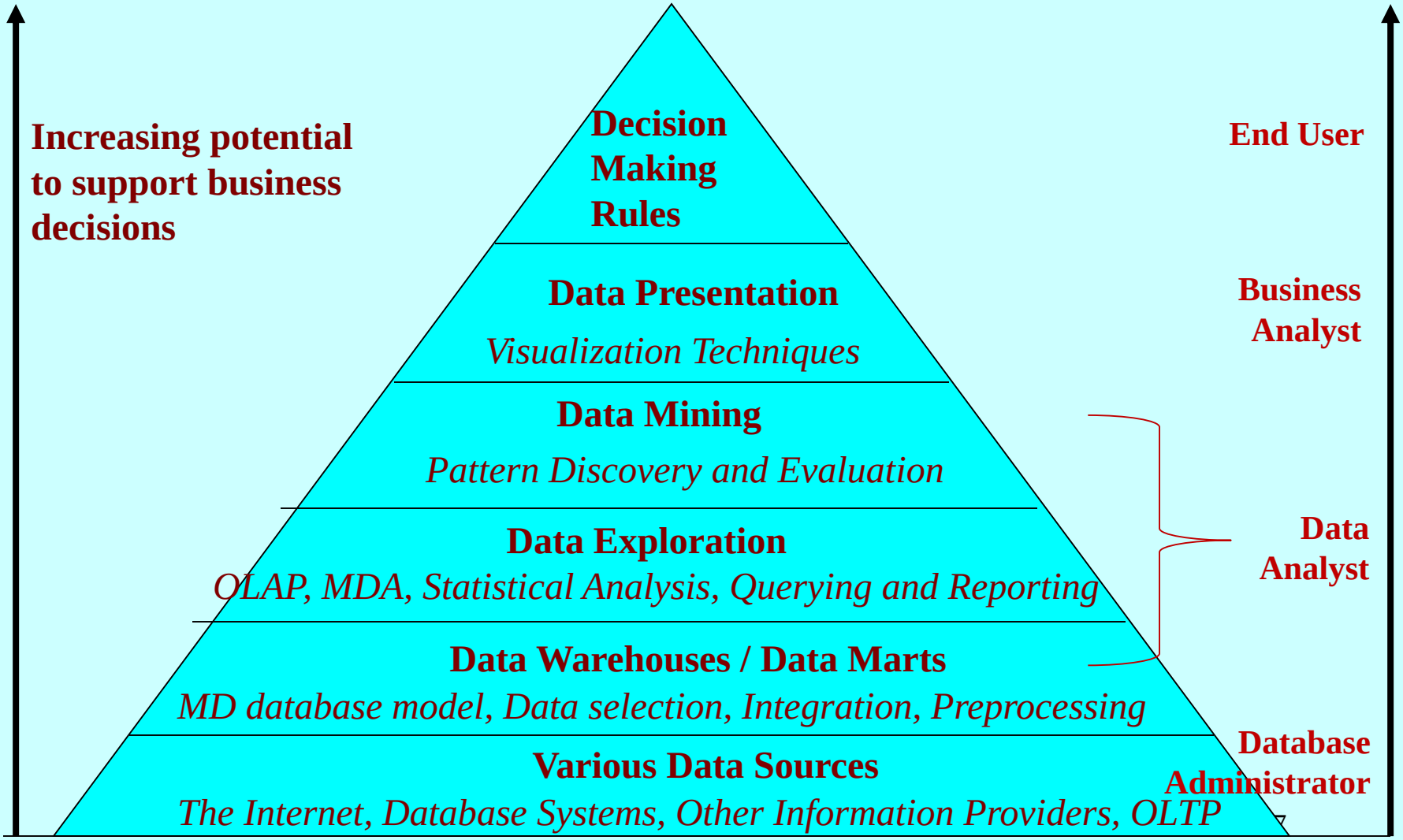
- **On Line Transactional (information) Process (OLTP)**
  - Track/record/retrieve original data records of every day business operations for answering *"what, when, where"* type of questions: Operational databases (Relational DB and SQL)
- **On Line Analytical (information) Process (OLAP)**
  - Store & manipulate summaries of various groupings of original data records for answering *"what happened to the business"* type of questions - Analytical databases: Data warehouses and OLAP
- **Knowledge discovery from data**
  - Discover/analyze hidden patterns of abstractive information (knowledge) for answering *"why and what to happen next"* type of questions: Data Mining (DM)


Shallow Data (discover with SQL)
Multi-Dimensional Data (discover with OLAP)
Hidden Data (discover with KDD)
Deep Data (discover only with clues)

DM/DW: Part of DSS Architecture



A Pyramid View of DSS: *Business Intelligence*



Data, Information, Knowledge

- **Data** is raw measures, or unprocessed facts, that have some relevancy to an individual or organization. Data without structural metadata can not be used to answer questions.
- **Information** is data which has been given some structural metadata, or processed that brings meaning. Information can be used to answer questions.
- **Knowledge** is laws or rules which are generalized higher-level information, corresponding to various regularity patterns hidden in datasets, that can be used to explain why on what happened, and predict what next.

Can you ask any questions and get answers from the following data set & why?

1.	high	bad	high	none	$0 to $15k
2.	high	unknown	high	none	$15 to $35k
3.	moderate	unknown	low	none	$15 to $35k
4.	high	unknown	low	none	$0 to $15k
5.	low	unknown	low	none	over $35k
6.	low	unknown	low	adequate	over $35k
7.	high	bad	low	none	$0 to $15k
8.	moderate	bad	low	adequate	over $35k
9.	low	good	low	none	over $35k
10.	low	good	high	adequate	over $35k
11.	high	good	high	none	$0 to $15k
12.	moderate	good	high	none	$15 to $35k
13.	low	good	high	none	over $35k
14.	high	bad	high	none	$15 to $35k

Can you ask any questions and get answers from the following data set & why?

1.	high	bad	high	none	$0 to $15k
2.	high	unknown	high	none	$15 to $35k
3.	moderate	unknown	low	none	$15 to $35k
4.	high	unknown	low	none	$0 to $15k
5.	low	unknown	low	none	over $35k
6.	low	unknown	low	adequate	over $35k
7.	high	bad	low	none	$0 to $15k
8.	moderate	bad	low	adequate	over $35k
9.	low	good	low	none	over $35k
10.	low	good	high	adequate	over $35k
11.	high	good	high	none	$0 to $15k
12.	moderate	good	high	none	$15 to $35k
13.	low	good	high	none	over $35k
14.	high	bad	high	none	$15 to $35k

They are data, but not information!

Data

- **Data:** i.e. raw data, a string of elementary symbols as digits or letters which has no meaning without specifying its context.
  - Data are particularly as measurements or observations of a set of variables (or called attributes).
  - E.g., 30, $45000, Sally, etc.
  - Data can be a value of an attribute: i.e. a simple measure of something, such as Sally's **age** = 30, Sally's **salary** ($) = 45000, person's **name** = Sally, etc.
  - Data without context can not be used for answering any questions.

Put data in context: **Information**

NO.	RISK	CREDIT HISTORY	DEBT	COLLATERAL	INCOME
1.	high	bad	high	none	$0 to $15k
2.	high	unknown	high	none	$15 to $35k
3.	moderate	unknown	low	none	$15 to $35k
4.	high	unknown	low	none	$0 to $15k
5.	low	unknown	low	none	over $35k
6.	low	unknown	low	adequate	over $35k
7.	high	bad	low	none	$0 to $15k
8.	moderate	bad	low	adequate	over $35k
9.	low	good	low	none	over $35k
10.	low	good	high	adequate	over $35k
11.	high	good	high	none	$0 to $15k
12.	moderate	good	high	none	$15 to $35k
13.	low	good	high	none	over $35k
14.	high	bad	high	none	$15 to $35k

Data from credit history of loan applications

Attributes & Metadata

Information: data in context and convey meaning to people

- In DB concept, information corresponds to
  - **Rows** (records, or tuples) in a table. A record represents a fact about an object, and a table is the collection of same type of objects.
  - **Columns** (attributes and value measures) in a table.
  - Information: a single cell value or a value derived from a group of cells

NO.	RISK	CREDIT HISTORY	DEBT	COLLATERAL	INCOME
1.	high	bad	high	none	$0 to $15k
2.	high	unknown	high	none	$15 to $35k
3.	moderate	unknown	low	none	$15 to $35k
4.	high	unknown	low	none	$0 to $15k
5.	low	unknown	low	none	over $35k
6.	low	unknown	low	adequate	over $35k
7.	high	bad	low	none	$0 to $15k
8.	moderate	bad	low	adequate	over $35k
9.	low	good	low	none	over $35k
10.	low	good	high	adequate	over $35k
11.	high	good	high	none	$0 to $15k
12.	moderate	good	high	none	$15 to $35k
13.	low	good	high	none	over $35k
14.	high	bad	high	none	$15 to $35k

Data from credit history of loan applications

Context: schema/metadata

Information rule: "All information in the database should be represented in one and only one way – as values in a table."
- Edgar Frank Codd., 1969
(http://www.15seconds.com/issue/020522.htm)

Examples of information

- **Information: data in context, has meaning**
  - Each value in the loan application table represents a piece of information about a specific characteristic of a particular record (i.e. object).
- **You can ask questions about the loan records for answers**
  - What is the debt and the risk of the client with NO=1?
  - Who are the clients with high risk?
  - What are the income groups for the high risk clients?
- **DB information retrieval:** (e.g. the loan information system)
  - For answering various ad hoc queries from the DB system.
  - Basic requirement: the loan table + query langue/operators + interface.

NO.	RISK	CREDIT HISTORY	DEBT	COLLATERAL	INCOME
1.	high	bad	high	none	$0 to $15k
2.	high	unknown	high	none	$15 to $35k
3.	moderate	unknown	low	none	$15 to $35k
4.	high	unknown	low	none	$0 to $15k
5.	low	unknown	low	none	over $35k
6.	low	unknown	low	adequate	over $35k
7.	high	bad	low	none	$0 to $15k
8.	moderate	bad	low	adequate	over $35k
9.	low	good	low	none	over $35k
10.	low	good	high	adequate	over $35k
11.	high	good	high	none	$0 to $15k
12.	moderate	good	high	none	$15 to $35k
13.	low	good	high	none	over $35k
14.	high	bad	high	none	$15 to $35k

Data from credit history of loan applications

Conventional Information System: DBMS

- Information system as DBMS
  - A database management system (DBMS) is also called information system since it is mainly for storing the information about a business, and answering specific questions by retrieving the stored information, i.e. table cell values, or simple derived values.
- The information to be retrieved can be a piece of fact, or an aggregation result of a group of facts.
  - E.g., From the Statistics Canada DB, query questions:
    - How many people in NS are 80, or older in 2016?
    - What is the population of baby bummers (borned from 1945 to 1965) in Nova Scotia?
    - What is the average starting salary of CS Bachelor graduates in Halifax in 2016?
  - Each aggregation result represents a piece of analytic information.
- When many aggregations are needed, from multiples sources: OLAP tool is needed.
  - E.g. The analysis of the Wealth and Health of 200 Countries, over 200 Years (http://www.youtube.com/watch?v=jbkSRLYSojo)

A historical global economic trends analysis:

Hans Rosling's 200 Countries, 200 Years, 4 Minutes - The Joy of Stats - BBC Four

(An animation with 120000 aggregated values from different data sources)

https://www.youtube.com/watch?v=jbkSRLYSojo

Knowledge

- **Knowledge:** general laws, rules, or patterns generalized from large group of factual information.
  - Physics laws of mechanics, thermodynamics, electricity and magnetism, etc. Such as Newton's Law of Gravitation can be applied to any objects on earth.
  - If temp > 30º C, then it is hot (dress in short).
  - If *education = CS Bachelor*, then *starting salary ≥ $45,000* (P=.78)
  - Knowledge is used to help making new **predications**, and for better **explanations**.
- **Knowledge acquisition**: machine learning/data mining
- **Knowledge based systems**: automatic problem solving by applying the built-in knowledge to the sensed environment, i.e. input data
  - E.g. Robot navigation, engine diagnose system, self-driving vehicles, bank load approval system, AlphaGo (using ML to master the ancient game of Go), etc.

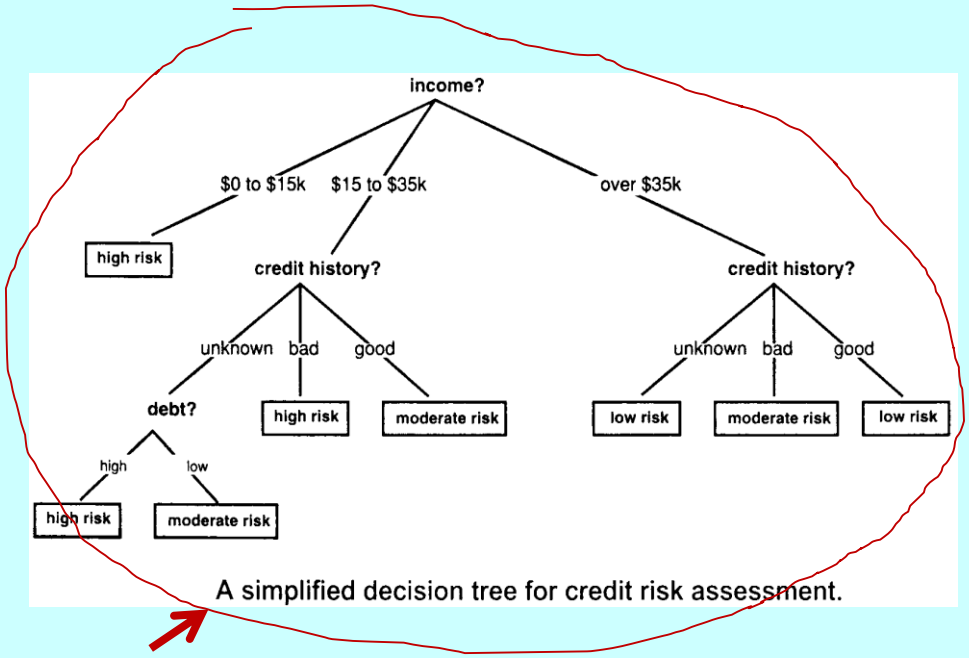
E.g., Finding the knowledge for risk predication.

Input: <CREDIT HIS=good, DEBT=low,
COLLATERAL=unknown, INCOME=$30k>.
Output: RISK=?

How to acquire the knowledge: Data mining.

NO.	RISK	CREDIT HISTORY	DEBT	COLLATERAL	INCOME
1.	high	bad	high	none	$0 to $15k
2.	high	unknown	high	none	$15 to $35k
3.	moderate	unknown	low	none	$15 to $35k
4.	high	unknown	low	none	$0 to $15k
5.	low	unknown	low	none	over $35k
6.	low	unknown	low	adequate	over $35k
7.	high	bad	low	none	$0 to $15k
8.	moderate	bad	low	adequate	over $35k
9.	low	good	low	none	over $35k
10.	low	good	high	adequate	over $35k
11.	high	good	high	none	$0 to $15k
12.	moderate	good	high	none	$15 to $35k
13.	low	good	high	none	over $35k
14.	high	bad	high	none	$15 to $35k

Data from credit history of loan applications

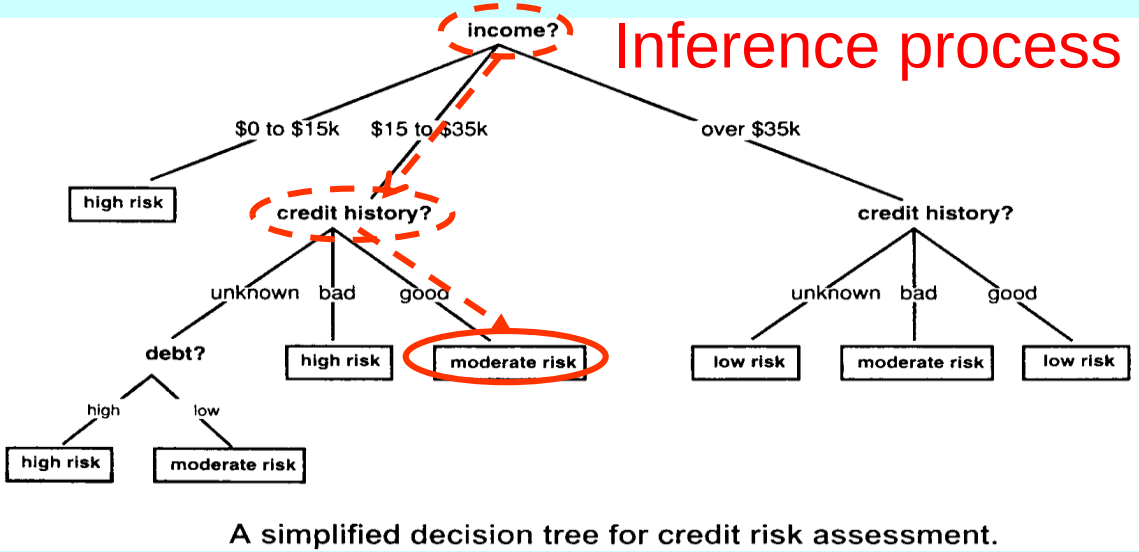


A simplified decision tree for credit risk assessment.

Knowledge representation: decision tree.

Business intelligent system: uses build-in knowledge to perform problem solving

- E.g. A system for automatic risk classification:
  - apply the build-in risk classification knowledge to make predication on risk for new applications.
  - Prediction is the inference process



A simplified decision tree for credit risk assessment.

Classification for new application:

Input: <CREDIT HIS=good, DEBT=low, COLLATERAL=unknown, INCOME=$30k>

Output: RISK=moderate risk

Inference process: if INCOME=$30k and CREDIT HIS=good, then RISK=moderate risk

Conventional Information System: DBMS

- **Using relational DB model**

  Structure: tables & links.
  Operations: (basic relational algebra operators)

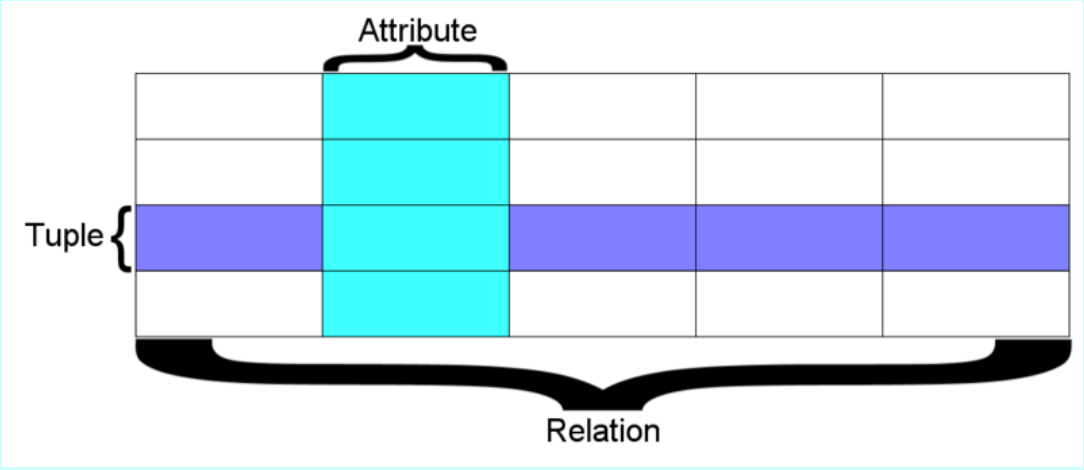
  - Selection, Projection (for single table)
  - Union, Intersection, Difference (for union compatible tables)
  - Cross Product, **Join** (for none union compatible tables)

- **Information is queried in SQL (Structured Query Language)**

  Each query is described by a SQL statement, which specifies sequence of the relational algebra operations for retrieving a defined information in terms of what has been stored.

Two Basic Rules of Relational DB Model

- **Information Representation Rule:**
  All information in the database should be represented in one and only one way -- as values in a table.
- **Information Access Rule:**
  Each and every datum (atomic value) is guaranteed to be logically accessible by resorting to a combination of table name, primary key value, and column name.



Basic SQL Clause Structure

- SQL is based on set and relational operations with certain modifications and enhancements
- A typical SQL query has the cause form:
  SELECT $A_1, A_2, ..., A_n$
  FROM $r_1, r_2, ..., r_m$
  WHERE P
  - A_is represent attributes
  - r_is represent relations
  - P is a predicate.
- This query is equivalent to the relational algebra expression:

$$\pi_{A1, A2, ..., An}(\sigma_P(r_1 \times r_2 \times ... \times r_m))$$

- The result of an SQL query is a new relation, i.e. a result table

DBMS Is for Operational DBs

- Operational DBs are transaction-based and often designed using relational DB model
- A such DB will contain several normalized tables. The objectives are to reduce redundancy and promote quick access to individual records
- It is not efficient for dynamically changed grouping operations from large data sets, in particular if the queried information is stored in different tables, event in different DBs of the same organization
- It is difficult to use SQL define complex queries

  Analyzing data and exploring relationship are not part of the SQL vocabulary.

- It is constrained to retrieve information from single database

How to harvest useful information from accumulated big data?

1. 1.8 trillion searches, 146 languages. What did the world search for, trends in 2016? (Google searches (84.5%), the Year in Review)
   - https://www.google.com/trends/yis/2016/GLOBAL

   Top Canadian Google searches
   - https://www.google.ca/trends/

How to derive useful info from data?

2. **Why most polls predicated wrongly about the outcome of the 2016 presidential election? How many eligible voters didn't vote in the election?**

   The Elections Project notes that there were 251,107,404 people who classify as members of the voting-age population, therefore 115,449,897 of the voting-age population (or 46.3 percent) did not vote. (The U.S. Census bureau estimates that there are over 320 million people living in the U.S.)

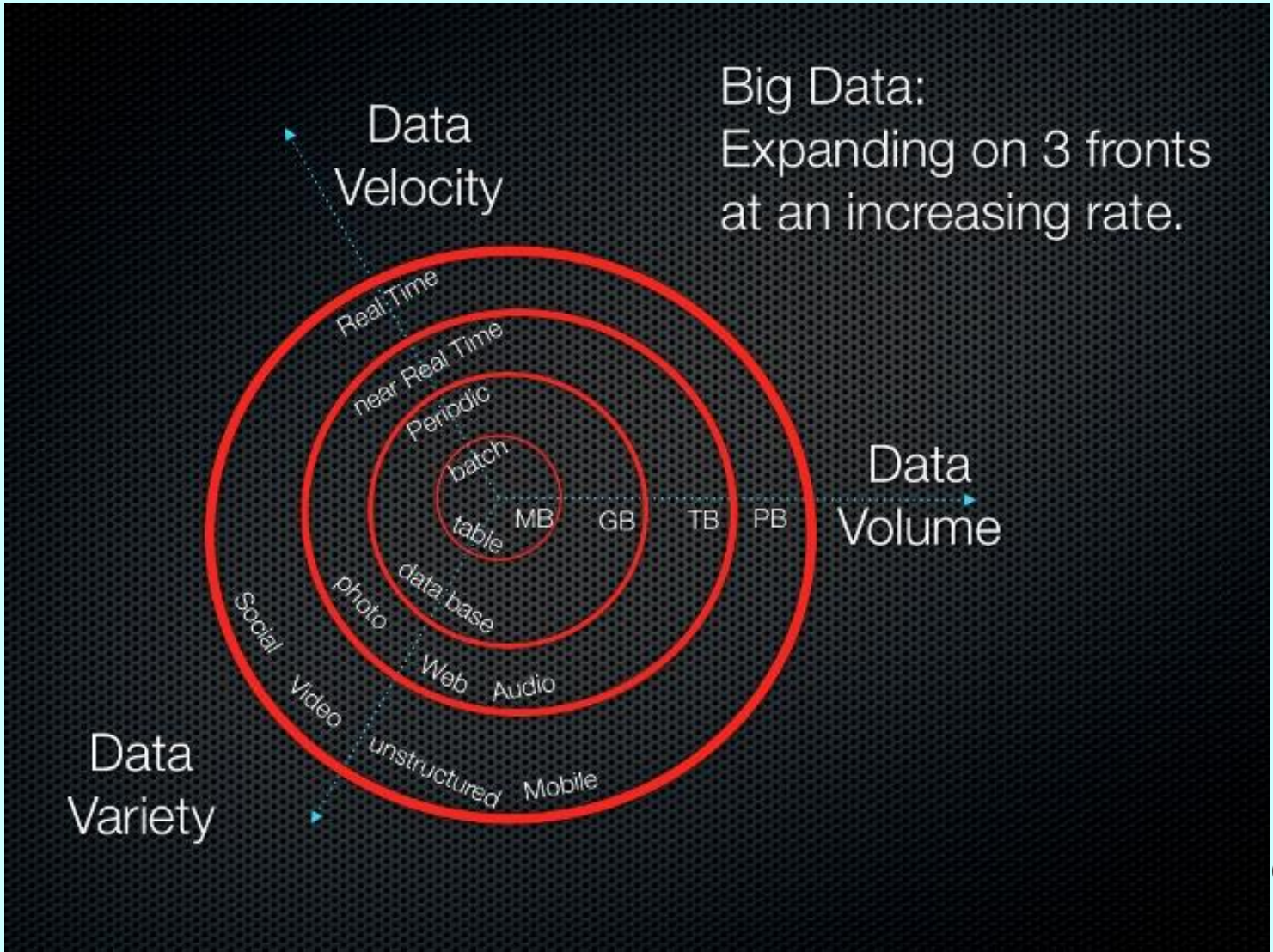
3. **350 million photos and videos sent by snapchatters every day (2013), what trends and interests shown in the data?**

   "Snapchat is a photo messaging application ("app") developed by Evan Spiegel and Robert Murphy, then Stanford University students".

4. "**The Wealth and Health of 200 Countries, over 200 Years...**" http://www.youtube.com/watch?v=jbkSRLYSojo

   A great demonstration on trend analysis using aggregated data with graphical animation too.

Challenges: 3Vs that define Big Data



Evolving DB/Analysis Technologies

- Database models and management technologies:
  - Relational database
  - Data warehouse models
  - Internet DB technology (Web 2.0)
  - NoSQL (Data modeled in other than the tabular relations)
  - MapReduce/Hadoop, Spark, Cloud computing, etc.
- Data analysis & information retrieval technologies:
  - SQL, Elasticsearch (a distributed, full-text search engine)
  - Data indexing & Pattern matching
  - Statistics
  - Data warehouse/OLAP
  - Machine learning & Data mining

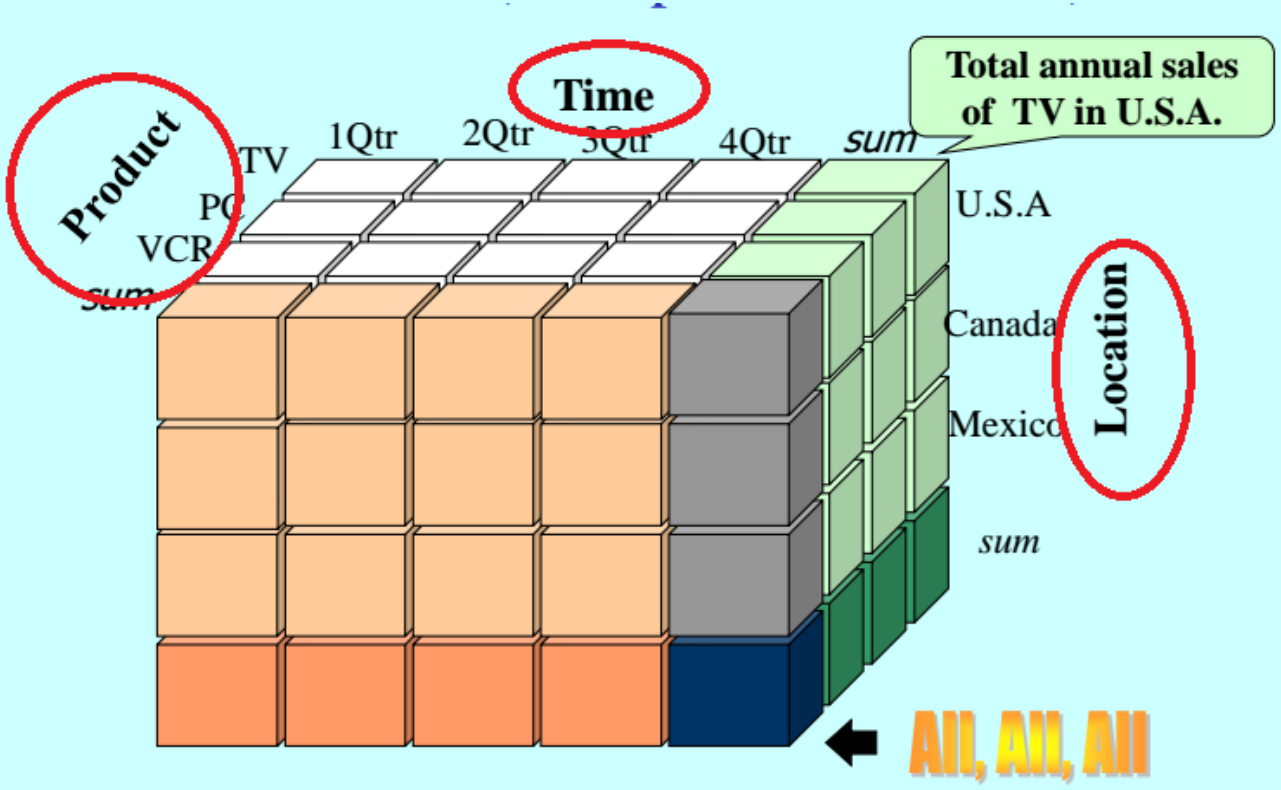
Data Warehouse (DW) Concept

DWs are central repositories of ***integrated*** data from one or more ***disparate sources***. They store ***current and historical summarized*** data and are used for creating analytical reports for business information workers throughout the enterprise. Examples of reports could range from annual and quarterly *comparisons* and trends to *detailed daily sales analyses*.

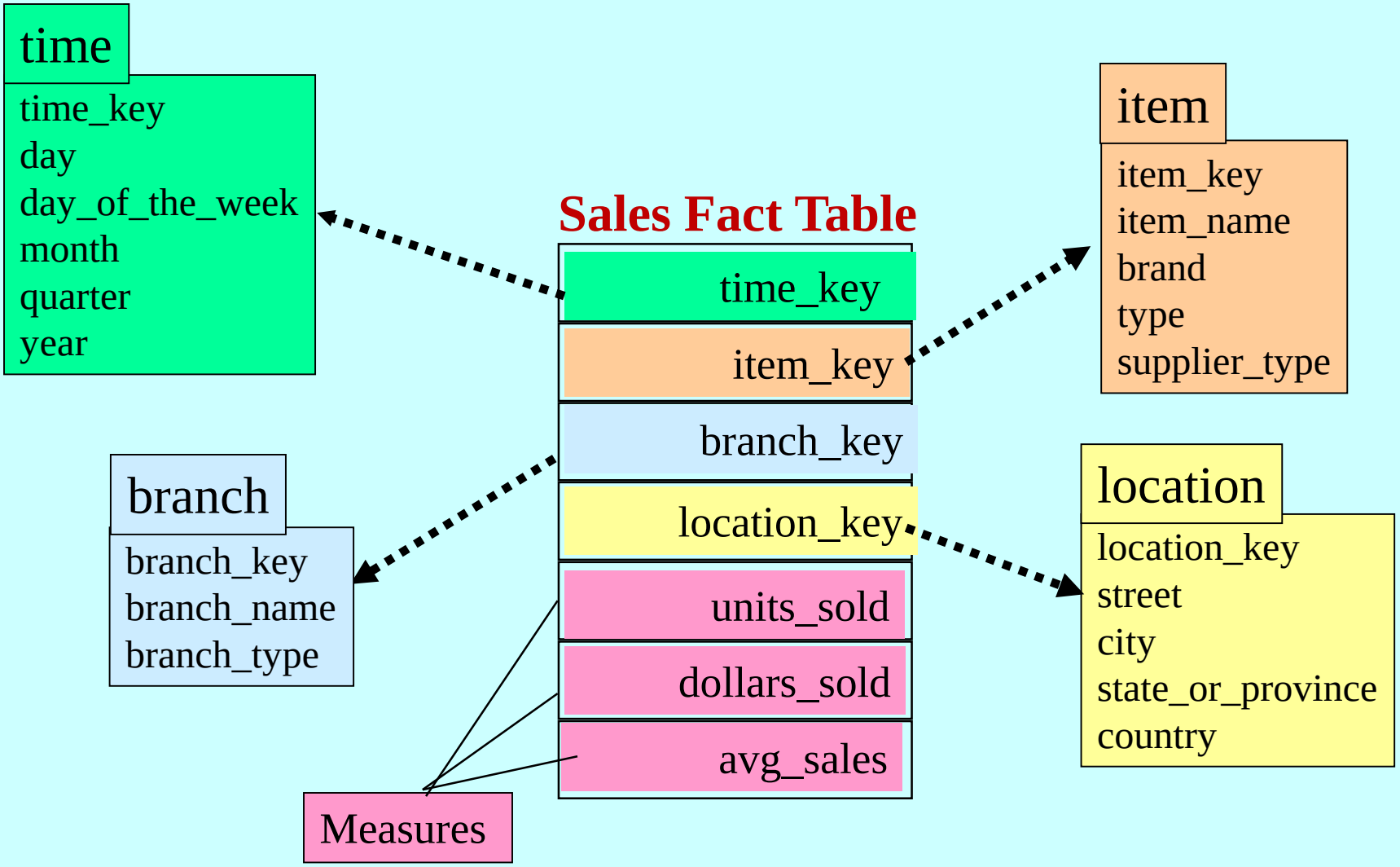
- DW is subject oriented (data are organized around business focus)
- The data stored in DW are aggregated information data (*i.e. various summarized data*)
- OLAP tools are used to quickly generate report for answering business ad hoc queries, etc

DW: Multi-dimensional data model

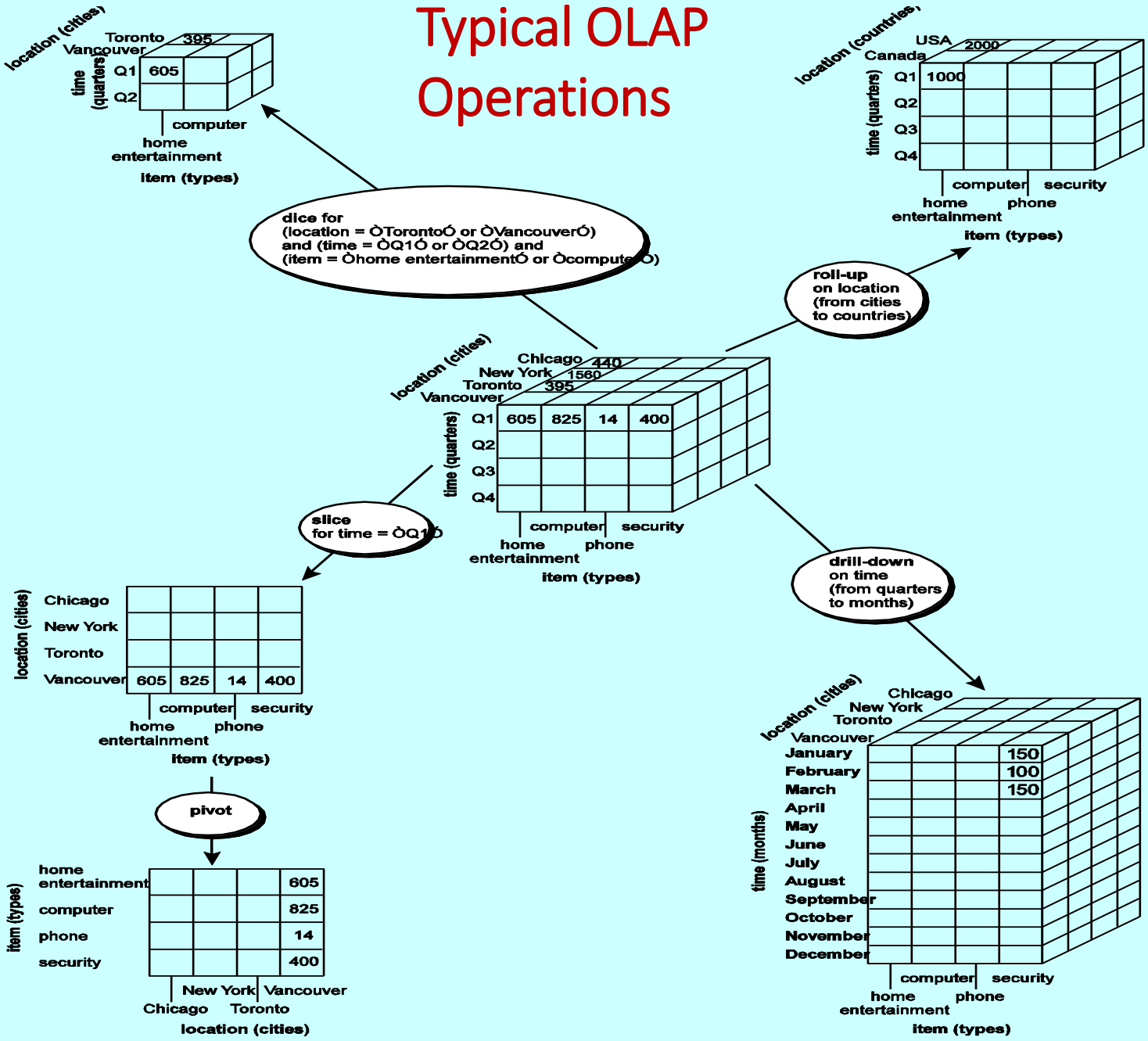
- E.g., Data cube with the subject "**Sales**"



Example of DW Star-Schema



Typical OLAP Operations



Example 1: War-Mart's DW and DSS

(http://derbaum.com/tu/WalMarts%20DWH.pdf,
Vienna University of Technology, 2006)

- Wal-Mart is an exceptional company on information innovation, an information system integrator.
- Fundamentally when you look at the value added by Wal-Mart, it is knowledge assets and how they are able to establish a *global OLAP information network.*
- Wal-Mart's data warehouse, the biggest in the world, enabled it to become a very successful company.

War-Mart's DW and DSS

https://www.healthcatalyst.com/wal-mart-birth-of-data-warehouse/

- Data mountains:

  Wal-Mart DW, from 1990: 300 Gigbyetes to 2001: 104 Terabytes (*1 TB = 10^3 GB = 10^6 MB = 250 million pages of text).

Wal-Mart began to achieve wide acclaim for its mastery of supply chain management. Behind the mastery of their supply chain was *Wal-Mart's data warehouse*. The world's largest retailer leveraged transaction data collected by its point-of-sales systems to achieve unprecedented insight into the **purchasing habits** of its 100 million **customers** and the **logistics guiding** its 25,000 **suppliers**.

"…a single repository for a completely integrated, 360-degree view of your business -one version of the truth."

- The system collects and analyzes item information from more than 4,000 stores to *track buying trends* department by department, shelf by shelf and item by item.

- It handles more than 30 applications and more than 50,000 business ad hoc queries per week. It's the largest commercial computer and data warehouse system in the world.

File Edit View Favorites Tools Help

Back Search Favorites History

Address http://www.ncr.com/products/software/teradata.htm Go

NCR Transforming Transactions into Relationships

Search: Go

Home | About NCR | Solutions | Industries | Products & Services | Contact Us | Buy NCR | Login

Resources

- Support
- Partners
- News and Events
- Library

Software Finder

- Teradata Warehouse
- Store Automation Software
- Self-Service Software
- Web Kiosk Software

Visit today

Teradata Warehouse Software

Teradata Warehouse Software

Successful business relationships today - whether with customers, partners, or suppliers - rely on integrating all the information an organization has gathered, and analyzing it to gain valuable insight, then putting that information into the hands of the people who make the day-to-day decisions that most affect customers. The advent of new technologies such as the Internet, call centers, and information kiosks provides information every minute, from many sources. Teradata, with its high-performance database, supporting tools and utilities and robust data mining capabilities give companies a competitive advantage by harnessing both their operational and historical detailed data in a centralized data warehouse that drives the business.

Visit Teradata.com for all the latest Teradata information, including:

- Teradata Database Software
- Worldmark midrange and enterprise servers
- Tools and Utilities
- Data Mining
- Storage and backup

At Teradata.com you will find the latest news, industry events, analyst reports, whitepapers, and much much more.

Software Products

- Teradata Database
- Teradata Warehouse Tools & Utilities
- Teradata Warehouse Mining
- Teradata CRM

Wal-Mart's DW for Business Management Queries

E.g. Let's look at some sales questions Wal-Mart's data warehouse has to answer:

- How much orange juice did we sell last year, last month, last week in store X?
- Comparing sales data of orange juice in various stores?
- What internal factors (position in store, advertising campaigns...) influence orange juice sales?
- What external factors (weather...) influence orange juice sales?
- Who bought orange juice last year, last month, last week?
- And most important: How much orange juice are we going to sell next week, next month, next year?

Other business questions may include:

- What is the suppliers price of orange juice last year, this year, next year?
- How can we help suppliers to reduce their cost?
- What are the shipping/stocking costs of orange juice to/in store X?
- How can suppliers help us reduce those costs?

(*http://derbaum.com/tu/WalMarts%20DWH.pdf, 2006)

The Merit of DW

A single repository for completely integrated, 360-degree view of your business -one version of the truth.

Review Questions

1. Use examples to explain the differences between terms Data, Information and Knowledge. How does each term/concept link to business information queries according to three types of information processes?
2. Why IT industry needs to develop DM and DW considering that DBMS/SQL are already available for storing and querying information?
3. Why and in which way DW model is more advanced than RDB model in supporting business management queries?
4. What are the major challenges to large corporations in terms of information management for supporting decision making?

Review Questions (cont)

5. What are the main limitations of conventional information systems, i.e. the DBMS technology, in terms of information queries?
6. Name three types of information process which are needed by large corporations, give a brief explanation for each.
7. As a common user of a DB system, such as Dal online or RBC online banking, which type of information process do you deal with & why?
8. For the president of Dalhousie Univ. or the dean of FCS, what type of information they are interested in getting?
9. As a store manager of War-Mart or Superstore, what type of information you need to know all the time?