CSCI6405 Fall 2003
Dta Mining and Data Warehousing

- Instructor: Qigang Gao, Office: CS219, Tel:494-3356, Email: q.gao@dal.ca
- Teaching Assistant: Christopher Jordan, Email: cjordan@cs.dal.ca
- Office Hours: TR, 1:30 - 3:00 PM

Lectures Outline

- **Pat I: Overview on DM and DW**
  1. Introduction (ch1) Ass1: Sep 9 - Sep 23
  2. Data preprocessing (ch3)
- **Part II: DW and OLAP**
  3. Data warehousing and OLAP (Ch2) Ass2: Sep 23 – Oct 7
- **Part III: Data Mining Methods/Algorithms**
  4. Data mining primitives (ch4)
  5. Classification data mining (ch7) Ass3: Oct 7 – Oct 21
  6. Association data mining (ch6) Ass4: Oct 21 – Nov 5
  7. Characterization data mining (ch5)
  8. Clustering data mining (ch8)
- **Part IV: Mining Complex Types of Data**
  9. Mining the Web (Ch9)
  10. Mining spatial data (Ch9)
- **Project Presentations**

Project Due: Dec 8

Analysis Oriented Information Process

Business information process:

- Tracking and analyzing business activities

Business actions: **operations** **+** **decision making**

|	|
Tracking the actions (data collecting, managing, …)	Analyzing the data (get information & knowledge from the data and predicate)

- How to extract information (truthful facts) and knowledge (about why and how) from the databases:

DW: To get fact information by grouping & aggregating data using OLAP.
DM: To get knowledge by discovering patterns/rules.

Business success depends on quick & wise decisions, that in turn needs strong support from the business intelligent information system.

Two Major Types of Information Process

- Operation-oriented information processing:

  - Handling daily based operations
  - On-Line-Transaction Processing (OLTP)

  * DB transaction: the execution of a program that includes database access operations. In OLTP, the users can change the DB.

- Analysis-oriented information processing:
  - Read-only transactions

  - On-Line-Analytical Processing (OLAP) for summarized factual information

  - Data mining for reveal hidden patterns

- Problems:

These differences can make it very difficult to combine operational and decision support processing within a single system, especially with respect to capacity planning, resource management, and system performance tuning.

The difficulty of defining analytical queries from the operational database.

Goal: Make analytical information available easily any where at any time.

Evolution of Information Systems

- **Specific application oriented DB systems (1960s)**

  Application: data file + special designed software

- **Database management systems (DBMS) (1970s-1980s)**

  Relational DB and SQL

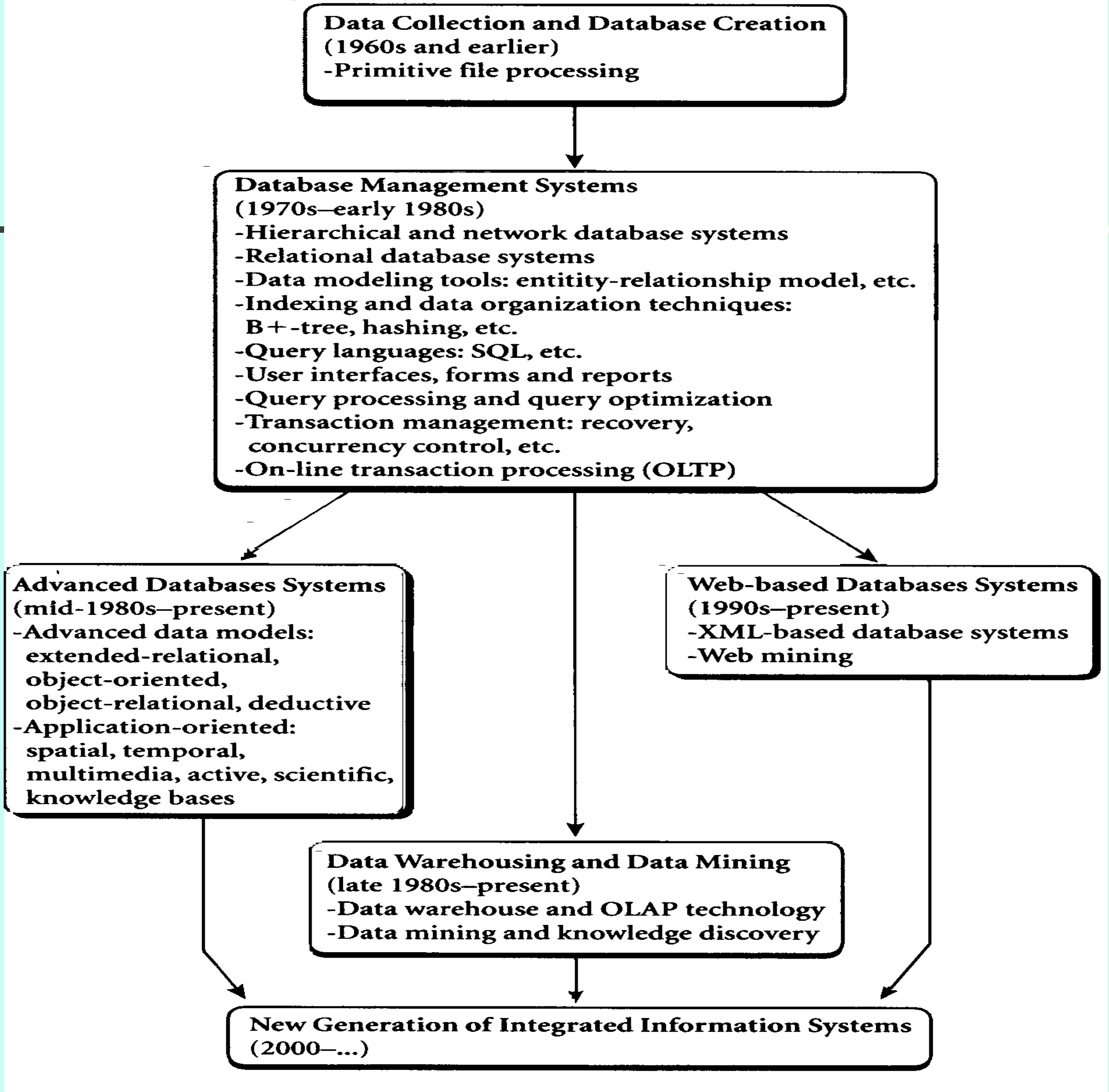
  Application: relational database + DBMS

- **Business analysis oriented information systems (1990s-2000s)**

  DW (OLAP system, SQL server) and DM systems

  Application: DW + OLAP + DM, or DB + DM

Evolution of DB Systems



Data Collection and Database Creation
(1960s and earlier)
-Primitive file processing

Database Management Systems
(1970s—early 1980s)
-Hierarchical and network database systems
-Relational database systems
-Data modeling tools: entitity-relationship model, etc.
-Indexing and data organization techniques:
B+-tree, hashing, etc.
-Query languages: SQL, etc.
-User interfaces, forms and reports
-Query processing and query optimization
-Transaction management: recovery,
concurrency control, etc.
-On-line transaction processing (OLTP)

Advanced Databases Systems
(mid-1980s—present)
-Advanced data models:
extended-relational,
object-oriented,
object-relational, deductive
-Application-oriented:
spatial, temporal,
multimedia, active, scientific,
knowledge bases

Web-based Databases Systems
(1990s—present)
-XML-based database systems
-Web mining

Data Warehousing and Data Mining
(late 1980s—present)
-Data warehouse and OLAP technology
-Data mining and knowledge discovery

New Generation of Integrated Information Systems
(2000—...)

Objectives of DM and DW

- Develop systems and tools which are optimized for handling business analysis oriented information for supporting decision making

- Make it easier to handle very large data repositories in an integrative manner

- Make complex queries to be easily defined and the information to be retrieved fast and accurately

- Make previously unknown and valuable knowledge hidden in data mountains to be quickly available

Data Warehouses and OLAP

Why we need DW and OLAP?

- Different nature of the two processes:

Operational information process, such as OLTP, usually have strict performance requirements, predictable workloads, small units of work, and high utilization.

E.g. Customer account databases of banks.

Decision support information process, such as OLAP, typically have varying performance requirements, unpredictable workloads, large units of work, and erratic utilization.

What is a data warehouse?

- **Data Warehouse:**

Data warehouse is a special kind of database for making analytical information available easily by using OLAP operations on informational data.

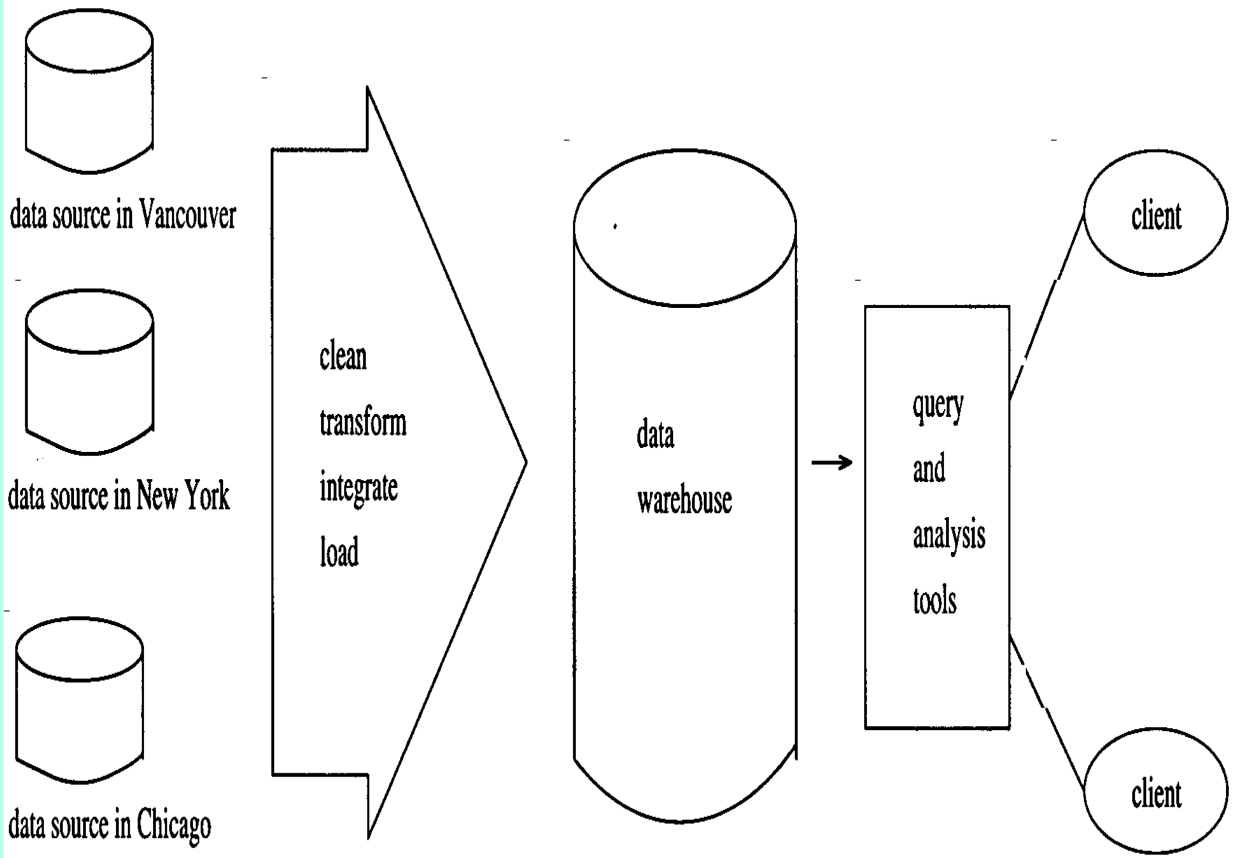
- **key features:**

Subject-oriented, integrated, nonvolatile, time-varied data store in support of management's decisions.

- **Main advantages:**

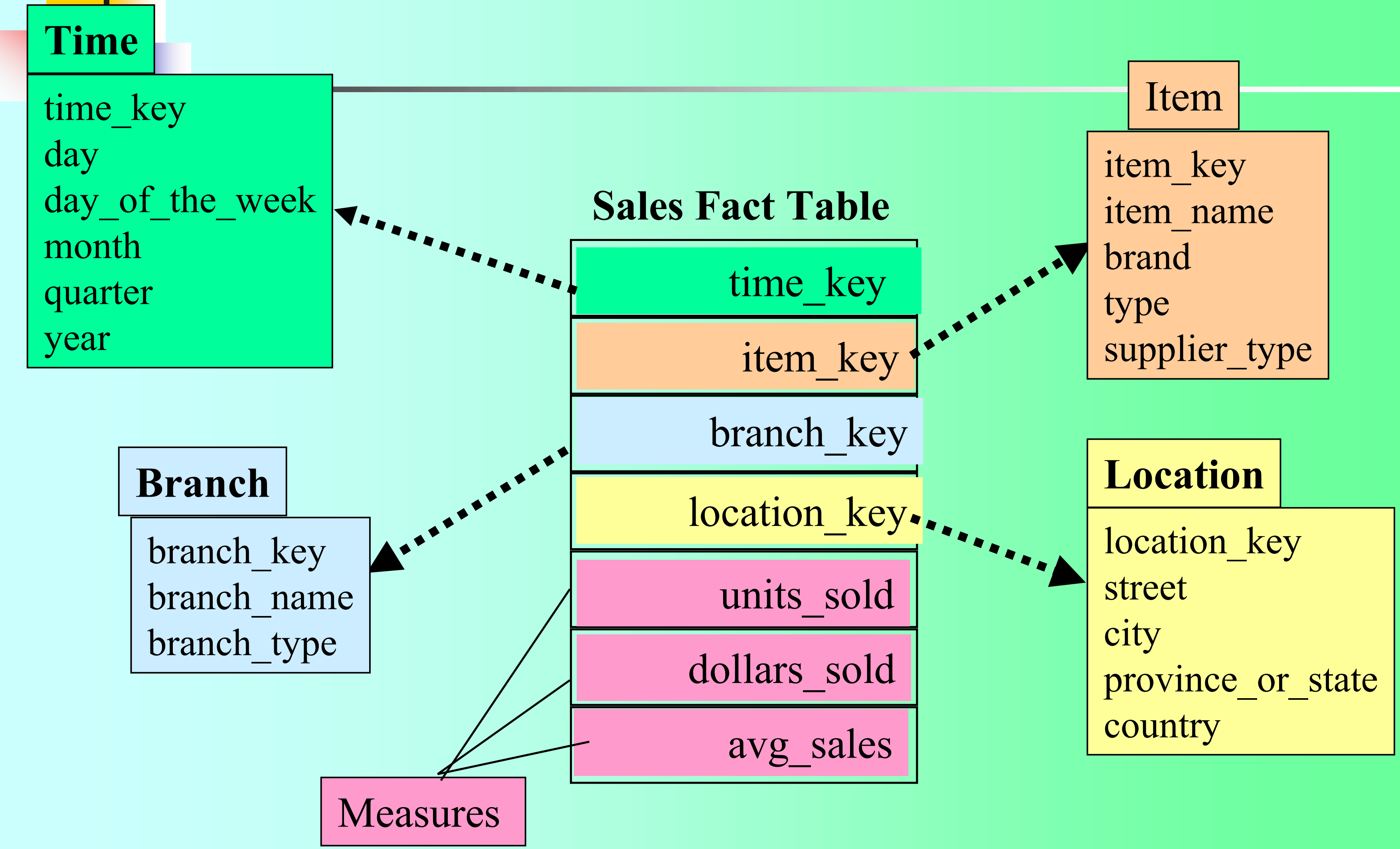
- Integrated data for more complete pictures
- Quick and accurate factual information for ad hoc questions
- Historical data for trend analysis
- Data is more suitable for data mining

A Simplified DW Architecture



Example of DW Star Schema

(The structure supports OLAP operations)



Time

time_key
day
day_of_the_week
month
quarter
year

Sales Fact Table

time_key
item_key
branch_key
location_key
units_sold
dollars_sold
avg_sales

Item

item_key
item_name
brand
type
supplier_type

Branch

branch_key
branch_name
branch_type

Location

location_key
street
city
province_or_state
country

Measures

DW Provides Good Information for DSS

What is good information?

Accurate: truthful and accurate on both raw data and calculations.

Timely: fast on both data access and data calculations.

Understandable: easily reorganized (comprehensively) views and friendly user interfaces.

Complete: integration of databases for a more complete picture of the information needed .

Overview on Data Mining

What's Data Mining?

Data mining is the process of extracting previously unknown, valid, action-able knowledge from large databases.

- Information Retrieval is not DM

- Statistics is part of DM

- Why machine learning plays an important role in DM?

Information Retrieval (IR) Is Not DM

- **IR:**

Finding some desired information in a DB or a store of information.

Selectivity: search and selection process (+ indexing tech).

It more emphases on finding the original stored information, and it is not involved in discoverring process.

E.g., Look for a specific set of records from a DB, or a set of files from a text DB (e.g., search web sites or articles from Internet).

- **DM:**

Discovering the patterns (regularities) hidden in the data. It emphases on new type of information which was not explicitly stored: knowledge.

Discovery: reasoning for new high-level representation (knowledge) from the the originally stored information.

E.g., Find the profile of a particular customer group, or the categories of the files in a text database, or the functions for detecting fraud (transactions).

DM result may help IR process (such as generate categories of text data, etc).

Stats Methods vs. DM

- **Hypothesis**

  Statistics: hypothesis testing.

  DM: a hypothesis-free approach.

  It is relative ease with which new insight can be gained.
- **Assumption on distribution**

  Statistics: need strong assumption.

  DM: fewer assumptions or no assumptions at all.
- **Data input**

  Statistics: mainly constrained to numerical data.

  DM: can be any data types.

- It is fair to say that statistics traditionally has been used for many of the analyses that are now done with data mining, such as building predictive models or discovering associations in databases.

Why Machine Learning (ML)?

- Leaning is one of basic necessities of life: any living creatures needs to have the ability to adapt themselves to their environment

  E.g., Learn problem solving ability from observations and experiences.

- A part of general methodology of science

  * Self learning is a process of discovery.

- To gain an understanding the methodological issues for adopting ML algorithms

What is learning?

An operational definition:

A certain 'task' to be carried out either well or badly, and a 'subject' that is to carry out the task; how to determine when someone has learned something.

An individual learns how to carry out a certain task by making a transition from a situation in which the task cannot be carried out to a situation in which the same task can be carried out under the same circumstances.

The process of this transition is called Learning, or Training.

Self-learning computer system

A self-learning computer can generate programs itself, enabling it to carry out new tasks.

Computer: speed and accuracy, but lack of creativity/flexibility.
Human : creativity/flexibility.

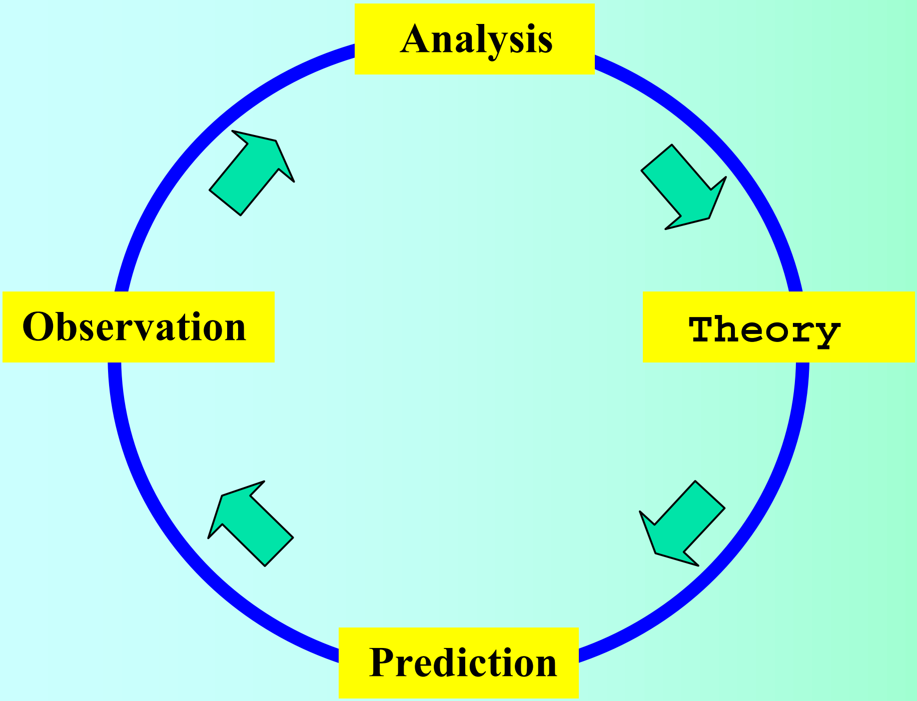
ML methods are developed to obtain knowledge automatically for carrying on new unknown tasks.

General Methodology of Science

- The fundamental task of the modern scientist is to **explain** and to **predict** based on the existing knowledge, and to discovery new knowledge
- A general methodology of science:

  The process of scientific research ideally takes the form so called **empirical cycle.**

Empirical Cycle



1. **Observation:** we start with a number of observations.

2. **Analysis:** we try to find patterns in these observations.

3. **Theory:** if we have found some regularities, we formulate a theory (hypothesis) explaining the data.

4. **Prediction:** our theory will predict new phenomena that can be verified by new observations.

Empirical cycle: An on-going process

In stage 4 of the cycle there are two possibilities:

a) our predictions are correct, in which case our theory is corrected, or
b) the predictions are wrong.

If b), we have to analyze the new observations and try to come up with a new theory. So the whole process starts again.

- This why we speak of an empirical cycle:

The process goes on and on for ever, and we can refine our theories indefinitely.

The same holds, apart from changes of detail, for a manager who tries to analyze a market to develop new products or optimize production.

- We can formulate hypotheses to explain empirical observations but that we can never prove that they are true.

Eg., The evolution of laws of gravity

- The gravity theory: Isaac Newton (1642-1727)

  *Newton's Law of gravitation is very accurate only when gravity is weak – and must be replaced by Einstein's general relativity in strong gravitational field.

- The general theory of relativity: Albert Einstein (1879-1955)

  *Similarly, relativity must be replaced by quantum mechanics when examining interactions on microscope scale, such as the big bang singularity, or at the edge and center of a back hole.

- Quantum gravity theory: Stephen Hawking (1941-), etc

  *The discovery is continuing …