# Database Systems: Achievements and Opportunities

Avi Silberschatz, Michael Stonebraker, Jeff Ullman

October 1991

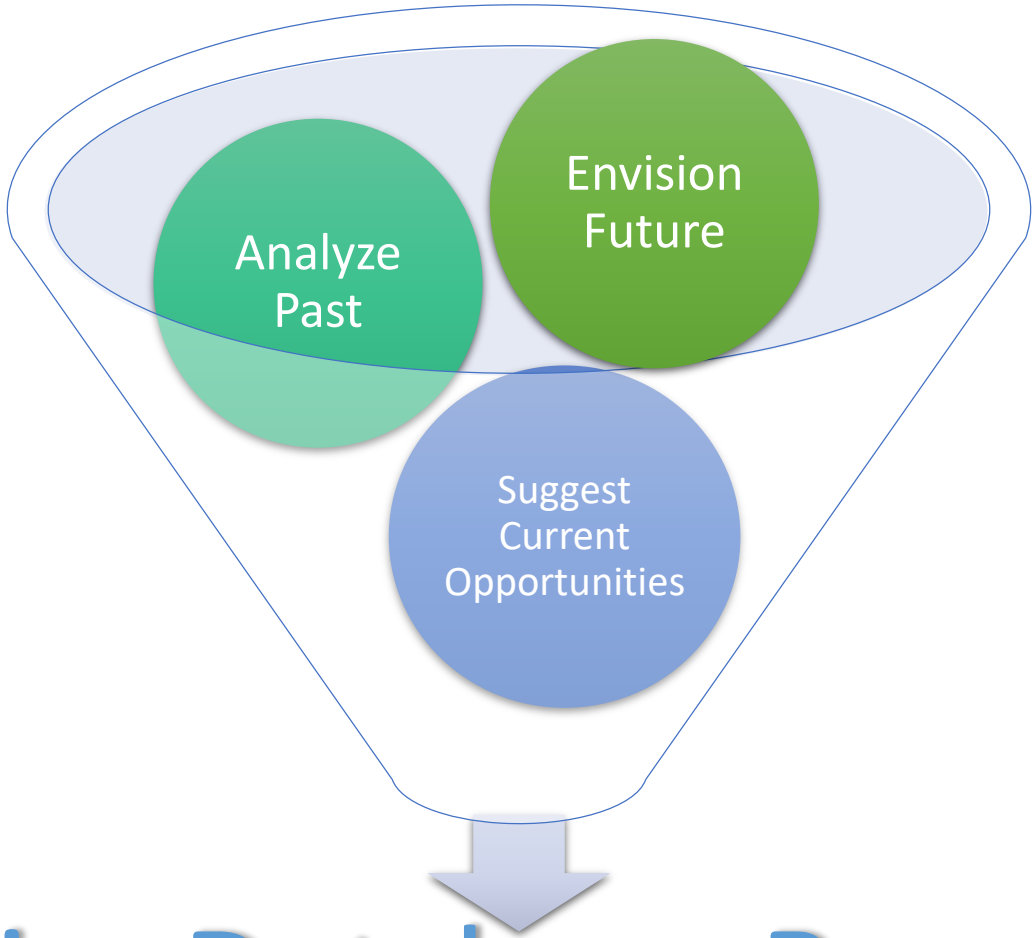
By Uzma Ali

# Agenda

- Motivation
- Database Research Overview
- Research Contributions
- What Does Future Hold?
- Suggested Research Directions
- Conclusion
- Q&A

# Motivation

# Motivation




**Inspire Database Research**

And contradict the general assumption in 1990's that the database technology is mature

# Database Research Overview

# Growing Data

**Library and Archives Canada**

Largest Library in Canada

20 million books and publications
3 million architectural drawings, maps and plans
24 million photographs
350,000 hours of film
More than a billion megabytes of digital content

**AND GROWING…..**

Can Lieutenant Commander Data from Star Trek find relevant information from this library?

His total linear computational *speed* was 60 trillion operations per second. So maybe he could.

# What is a Database?

**Database: The Information you lose, when your memory crashes. Dave Barry**

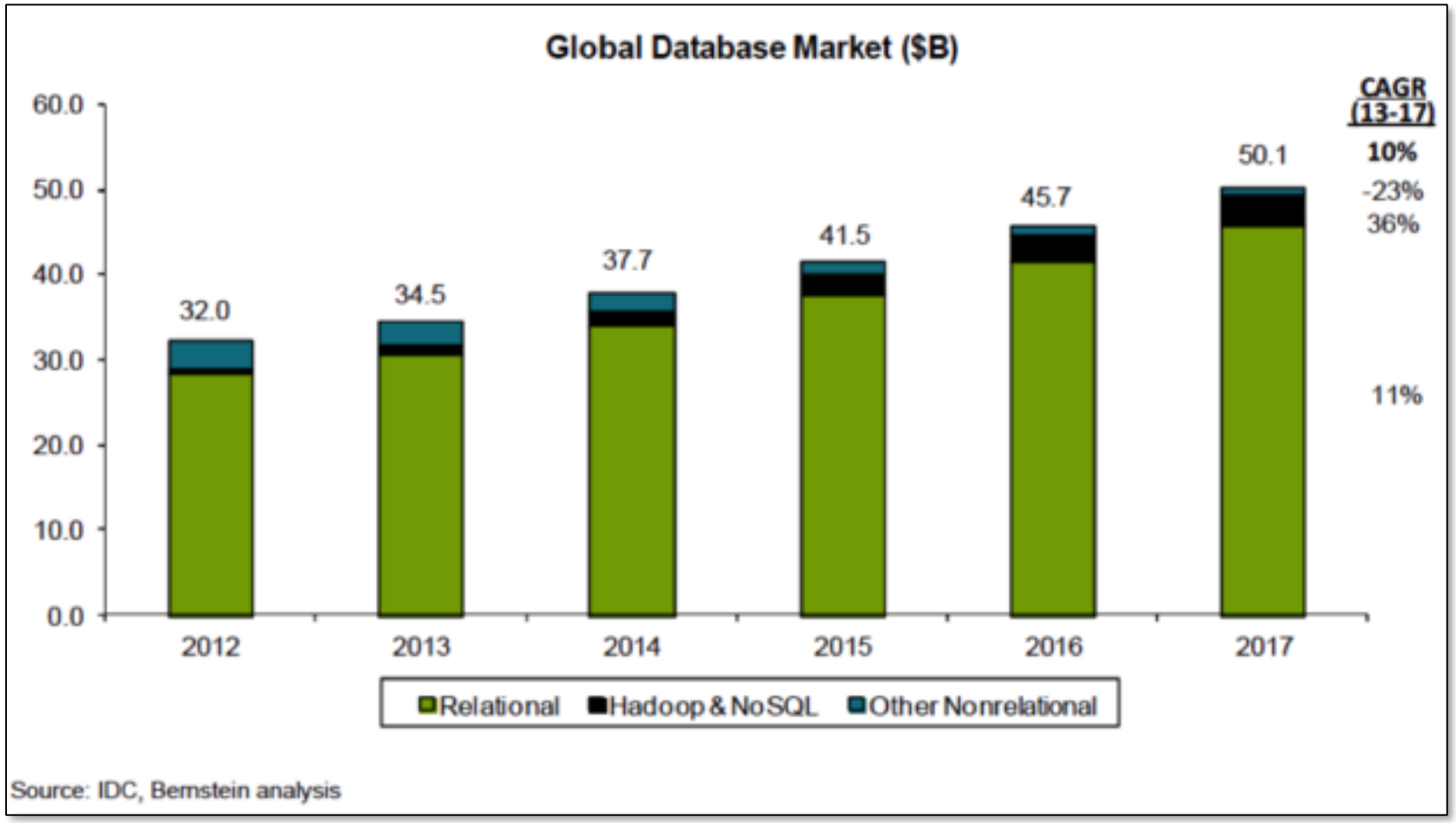
A **database** is an organized collection of data

# Database Management Systems

A **database management system** (**DBMS**) is system software for creating and managing databases.

The **DBMS** provides users and programmers with a systematic way to create, retrieve, update and manage data..

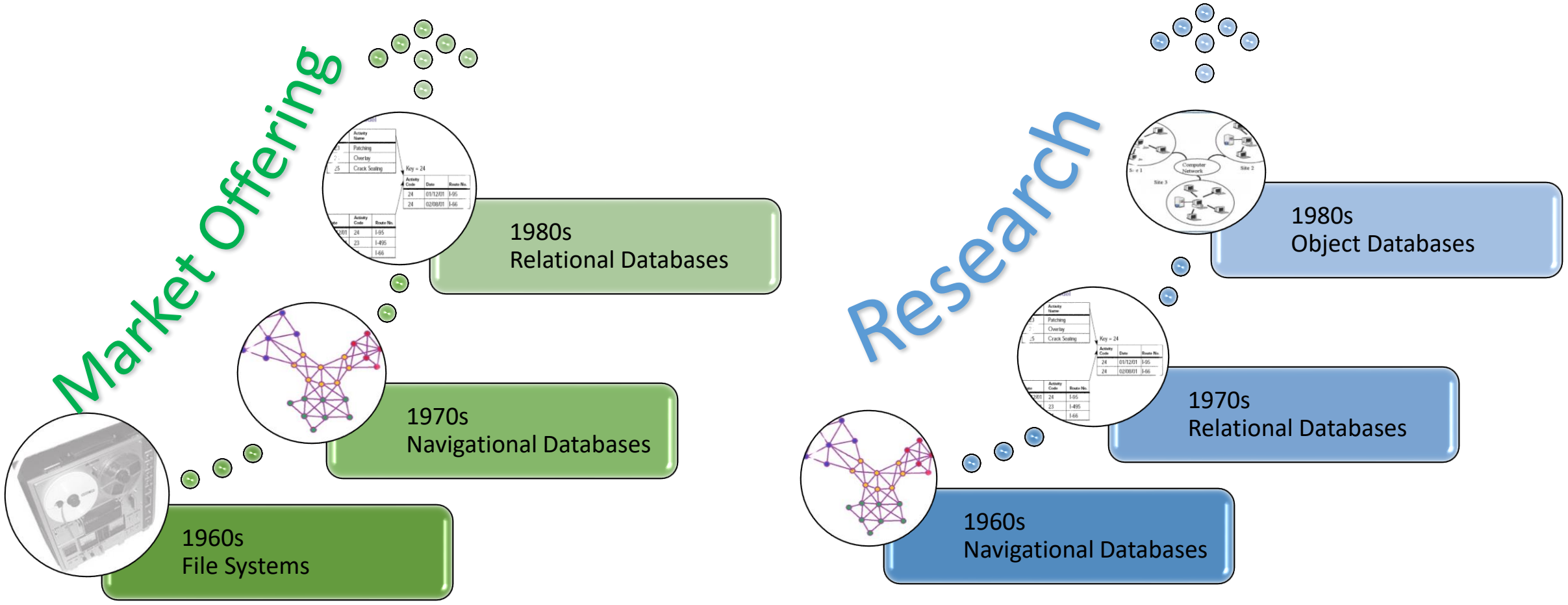
# Database Industry Revenue Trend




**The compound annual growth rate** (**CAGR**) from Year 2013 to Year 2017 is 10%

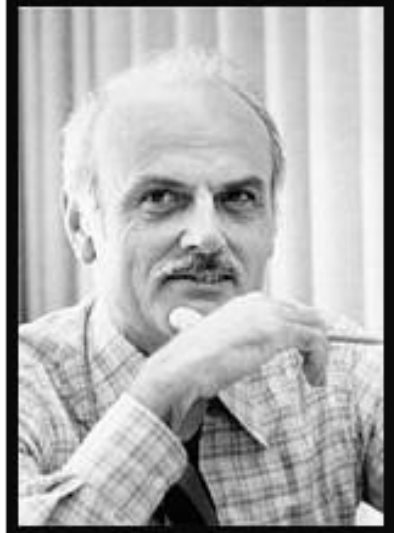
http://www.infoworld.com/article/2916057/open-source-software/open-source-threatens-to-eat-the-database-market.html

# Database Research vs Commercialization




Technological Advancements in the industry are followed by years of research

# Example: Database Research Impact



The most important motivation for the research work that resulted in the relational model was the objective of providing a sharp and clear boundary between the logical and physical aspects of database management.

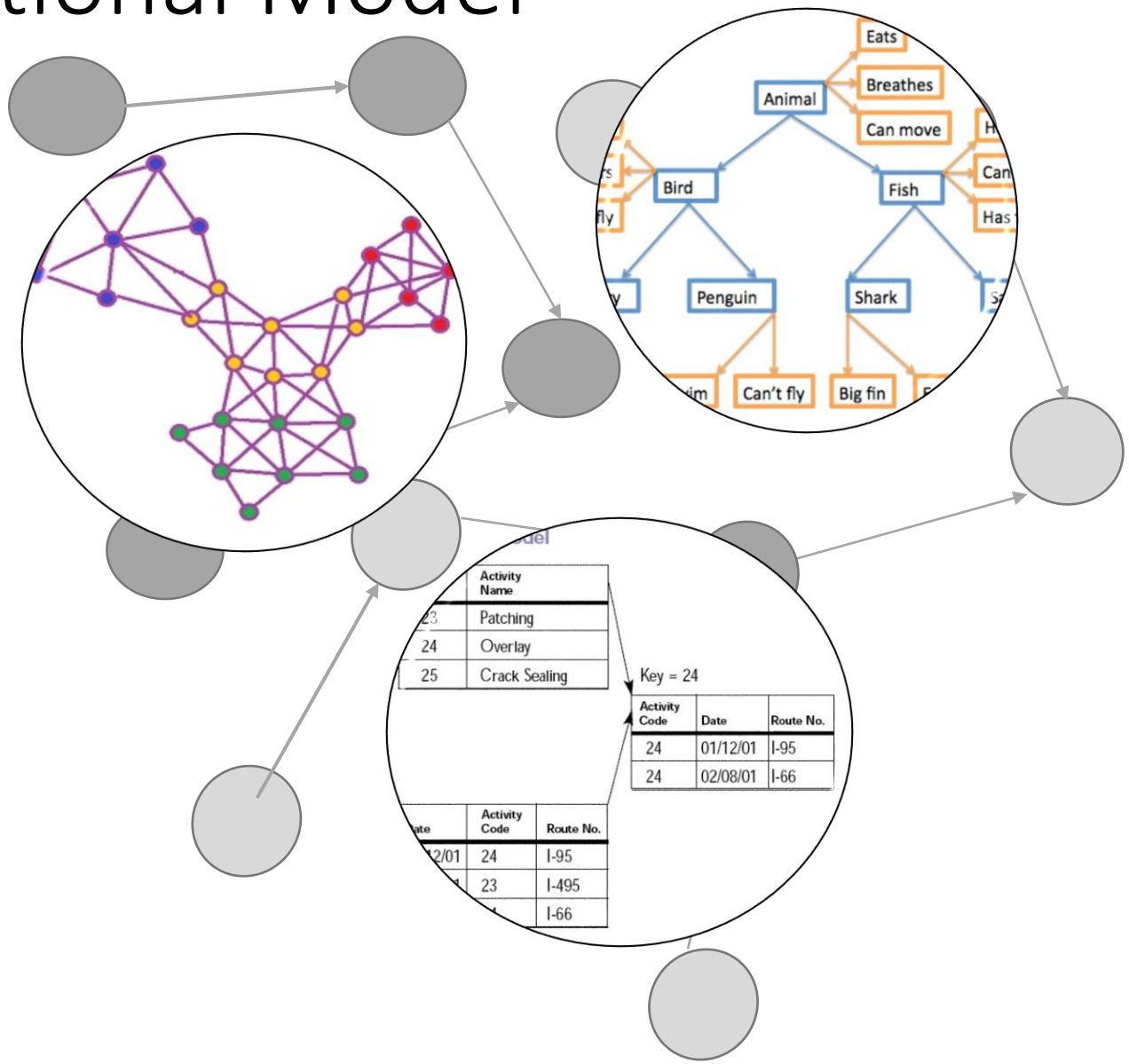
(E. F. Codd)

izquotes.com

# Research Contributions

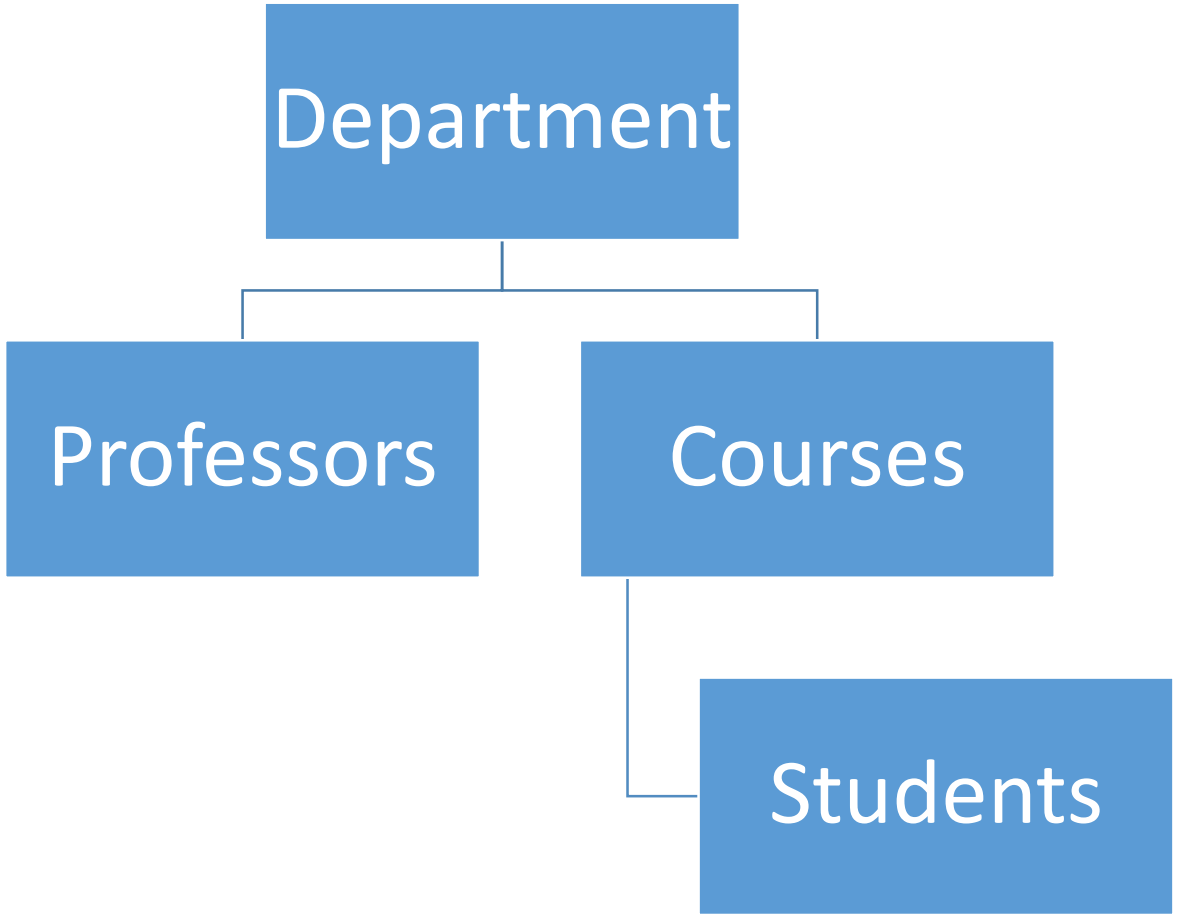
# Evolution of the Relational Model

- Hierarchical Model
- Network Model
- Relational Model



# Hierarchical Model

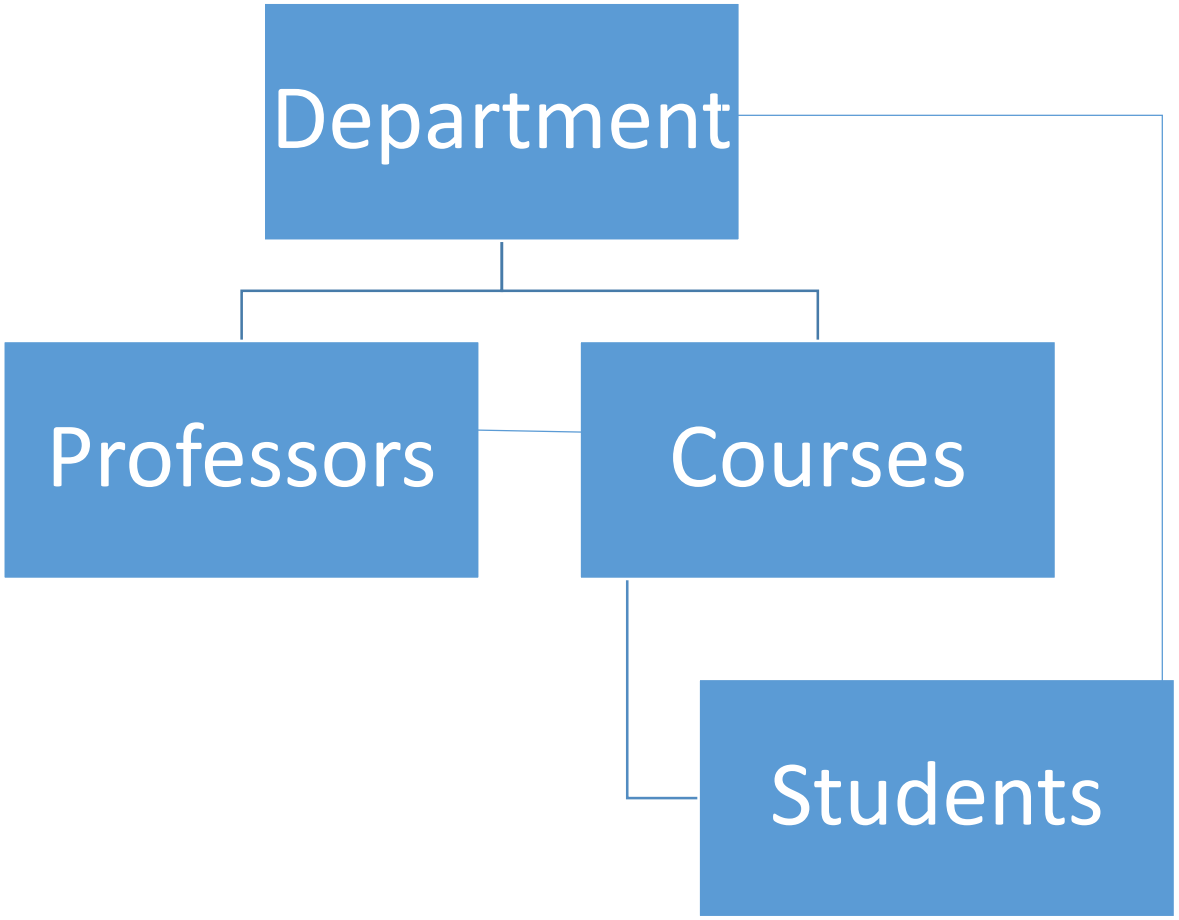
IBM Information Management System (IMS) – still in use




- All data records are assembled into a collection of trees
- Programmer defined physical storage format
- Complex application programs required to navigate the database for simple queries

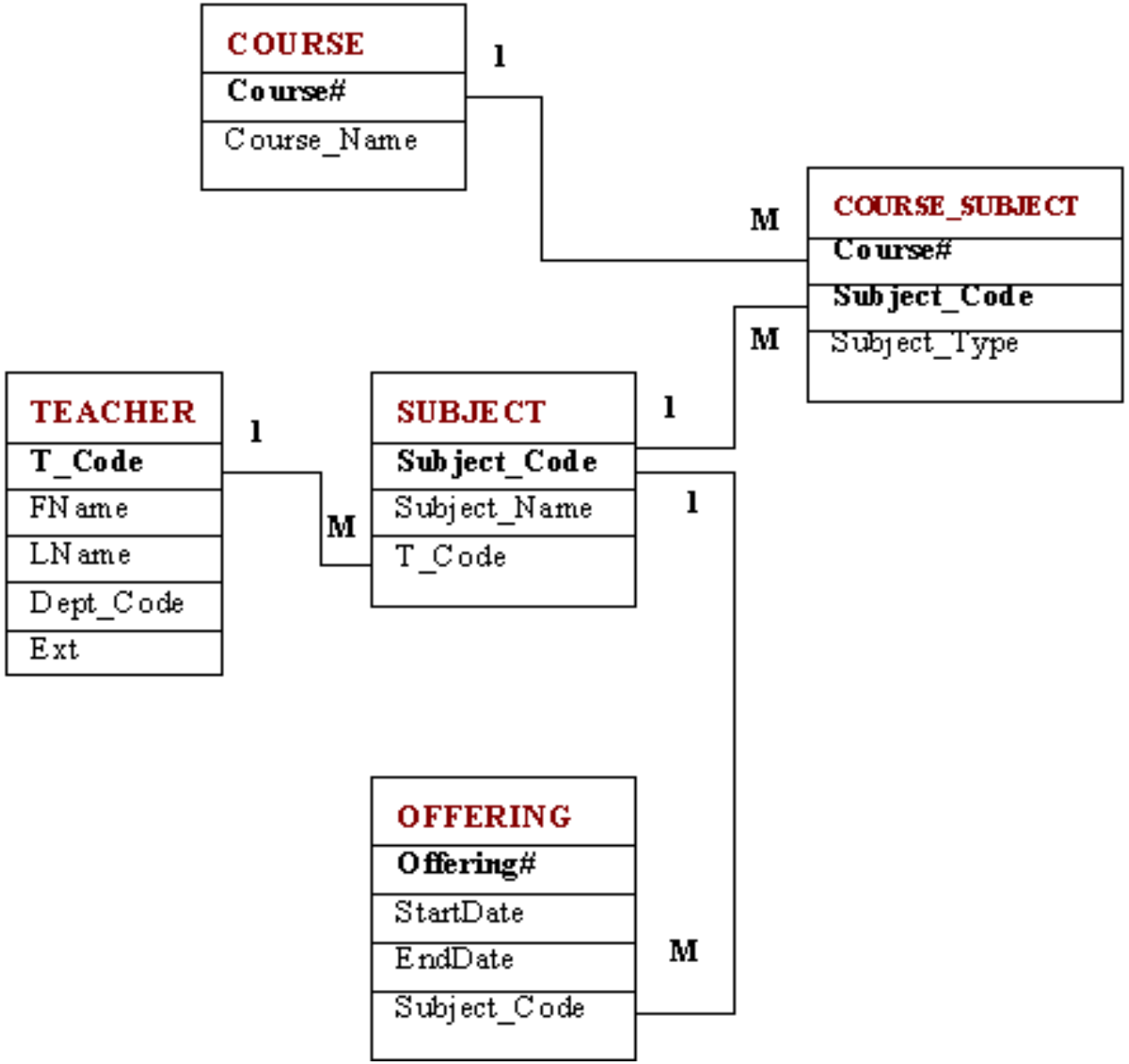
# Network Based Model

Integrated Data Store (IDS) was the basis of CODASYL standards




- Allows generalized graph structure to represent objects and relationships
- It was not easy to use or implement applications with IDS, because it was designed to maximize performance using the hardware available at that time

# Relational Model



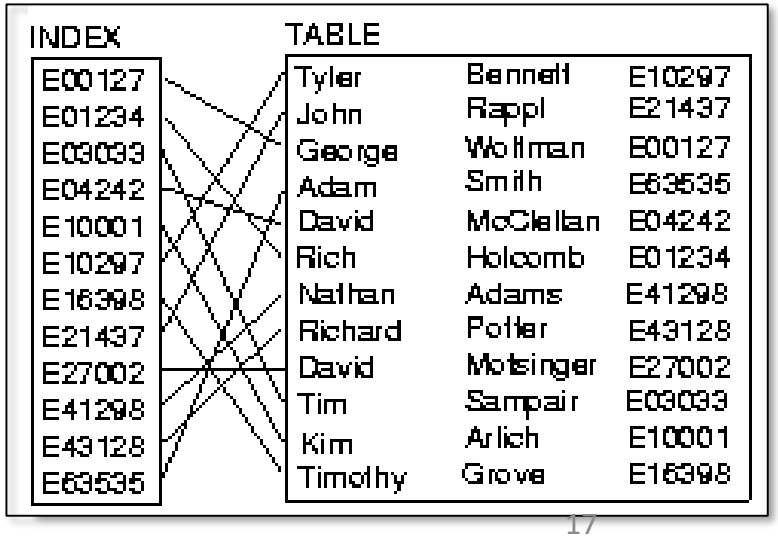

- All Data is stored into tuples and then grouped into tables
- The tables are stored independently
- Users would use a high level, non-procedural, declarative language to provide the required information
- The DBMS would take care of storing the data and retrieval procedures for answering queries by using query optimizer

# Research on Relational Model

- Invent of high level relational query language eg SQL.
- Development of algorithms to optimize queries. That is translate queries into plans that are as efficient as programs written for Navigational DBMS.
- Formulated theory of normalization to eliminate redundancy.
- Created indexing techniques for fast access to data records
- Developed algorithms to move data between disk and main memory pool understanding the access patterns

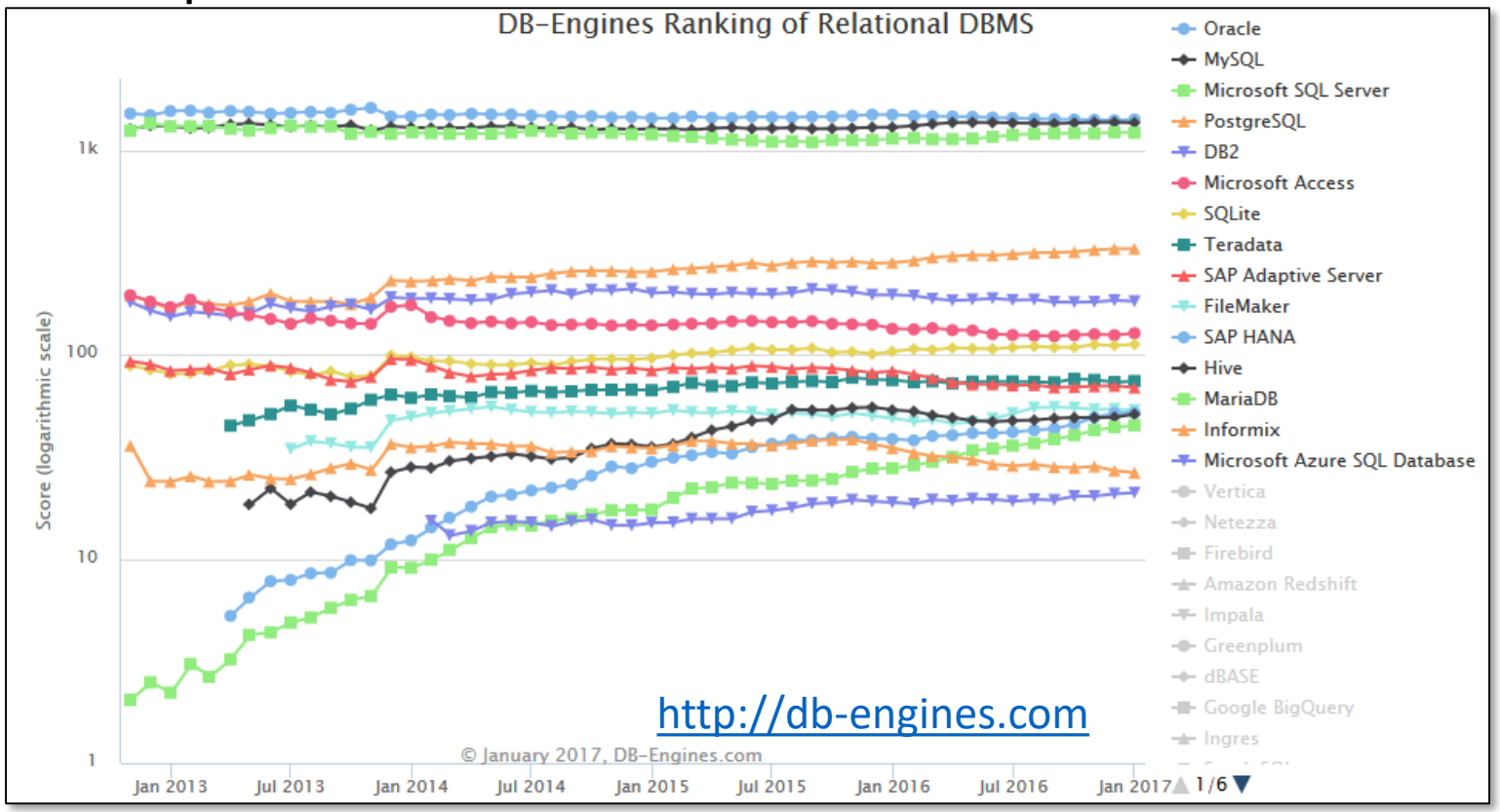
```
SELECT *
FROM Courses c
INNER JOIN Departments d
On c.deptid=d.deptid
```

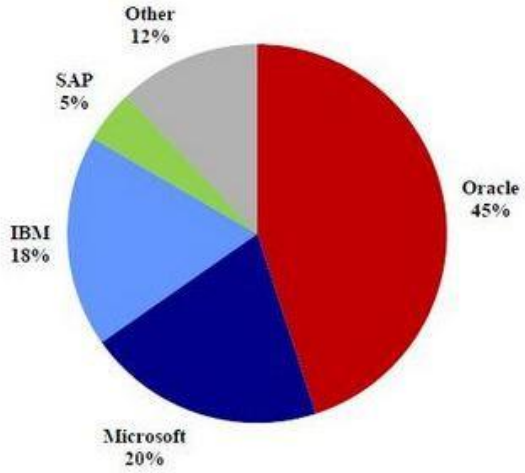
| SSN | Name | Town | Zip |
|---|---|---|---|
| 1234 | Joe | Stony Brook | 11790 |
| 4321 | Mary | Stony Brook | 11790 |
| 5454 | Tom | Stony Brook | 11790 |
| | ..... | | |



# Results of Research on Relational Model

- Numerous commercial relational products came into market in 1980s.
- 1990 onwards, commercial relational databases are available for all hardware platforms ranging from personal computers to mainframes.




*Source: IDC (#241292, May 2013)*

# Transaction Management

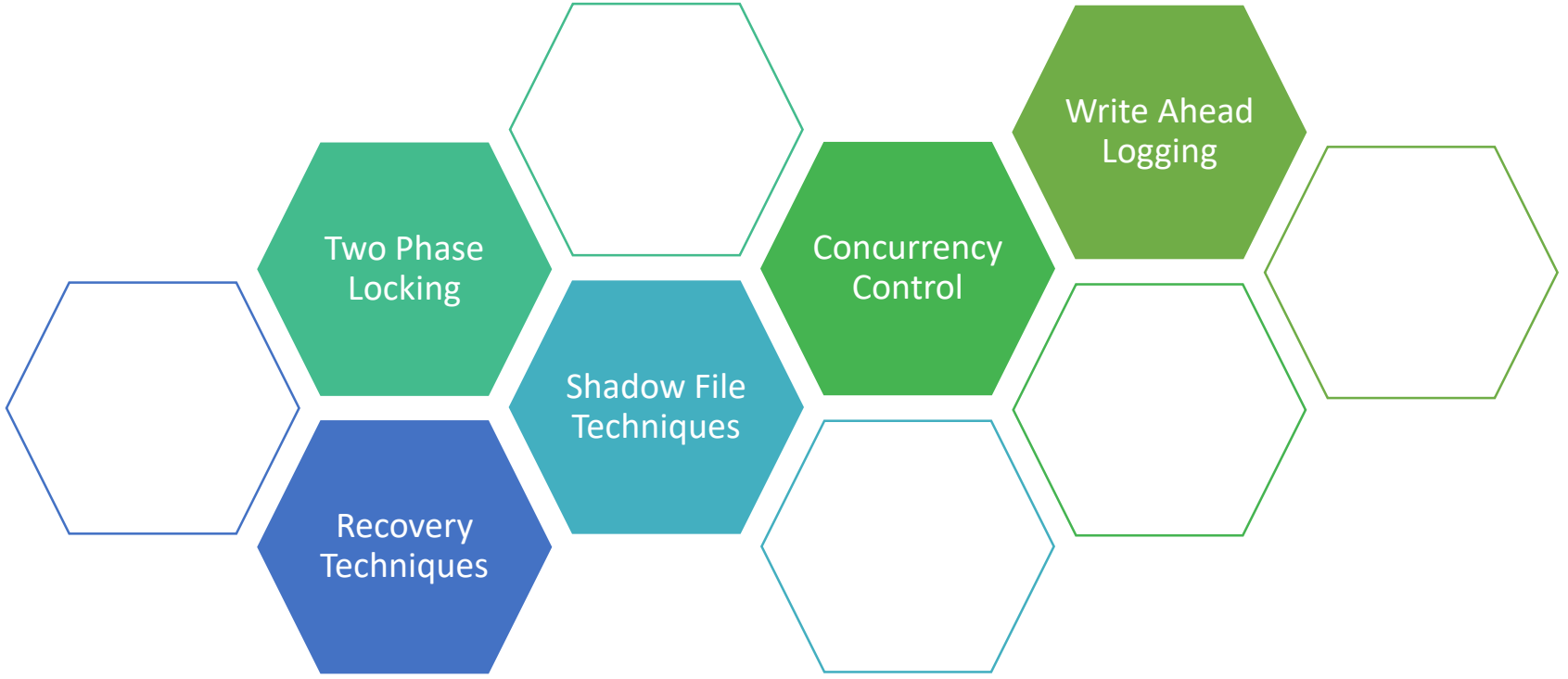
A transaction is a series of operations on a DBMS that follow ACID properties.

**A**tomicity

**C**onsistency

**I**solation

**D**urability



# Transaction Management

**A**tomicity

**C**onsistency

**I**solation

**D**urability

All or nothing

This property states that the transaction is executed as a single unit of work. That is, either all actions within the transaction are carried out or none are.

# Transaction Management

**A**tomicity

**C**onsistency

**I**solation

**D**urability

No constraints are violated

This property states that each transaction run by itself must preserve the consistency of the database.

# Transaction Management

**A**tomicity

**C**onsistency

**I**solation

**D**urability

Users (sessions) don’t affect each other

This property states that due to performance reasons, multiple transactions are interleaved.

# Transaction Management

**A**tomicity

**C**onsistency

**I**solation

**D**urability

Once data is committed, it is permanent

This property states that transactions survive system crashes and failures.

# ACID Techniques

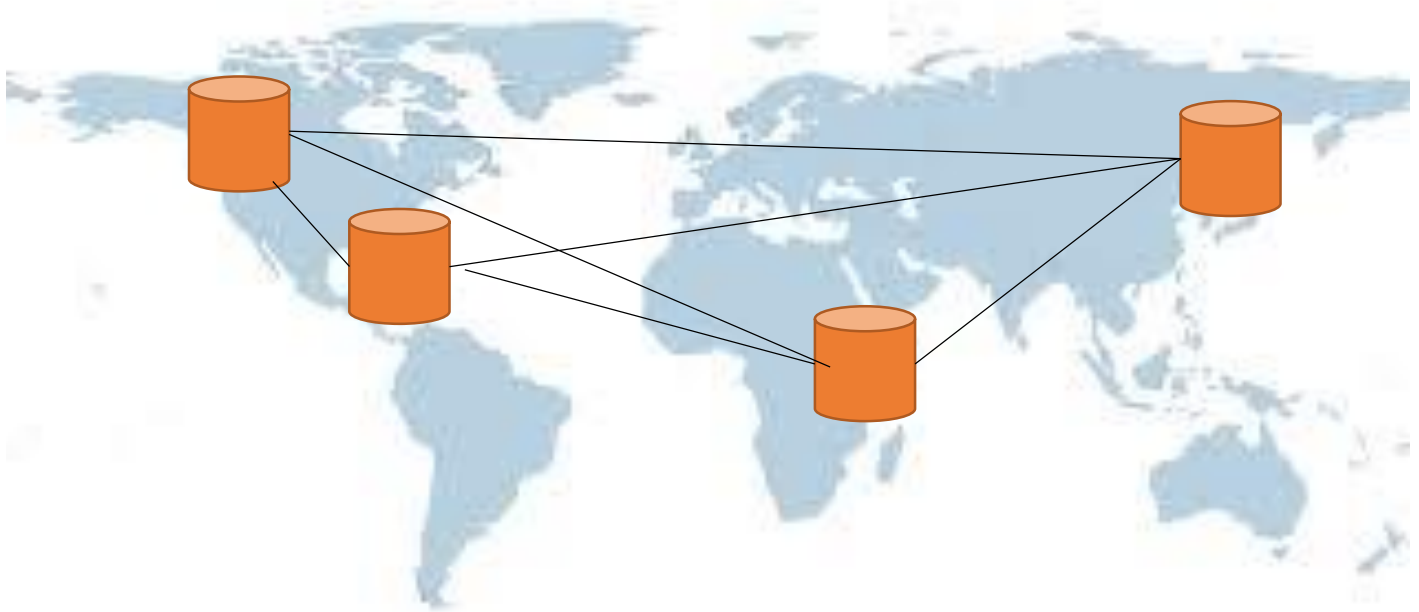
- Concurrency Control – providing Isolation & Consistency assurance
  - Two phase locking

  Requires transactions to obtain all locks before releasing all.
  - Prevent serializability violations by using access timestamps
- Recovery – providing Durability
  - Write-ahead Logging

  Effects of transaction are logged in a sequential file that enables database restoration
  - Shadow file techniques

  Keeping additional copy of the database file(s).

# Distributed Databases



- The data can come closer to the users responsible for maintaining it
- The decentralized system is less likely to crash with all the system wide outage
- Less likely to lose replicated data if one site goes down

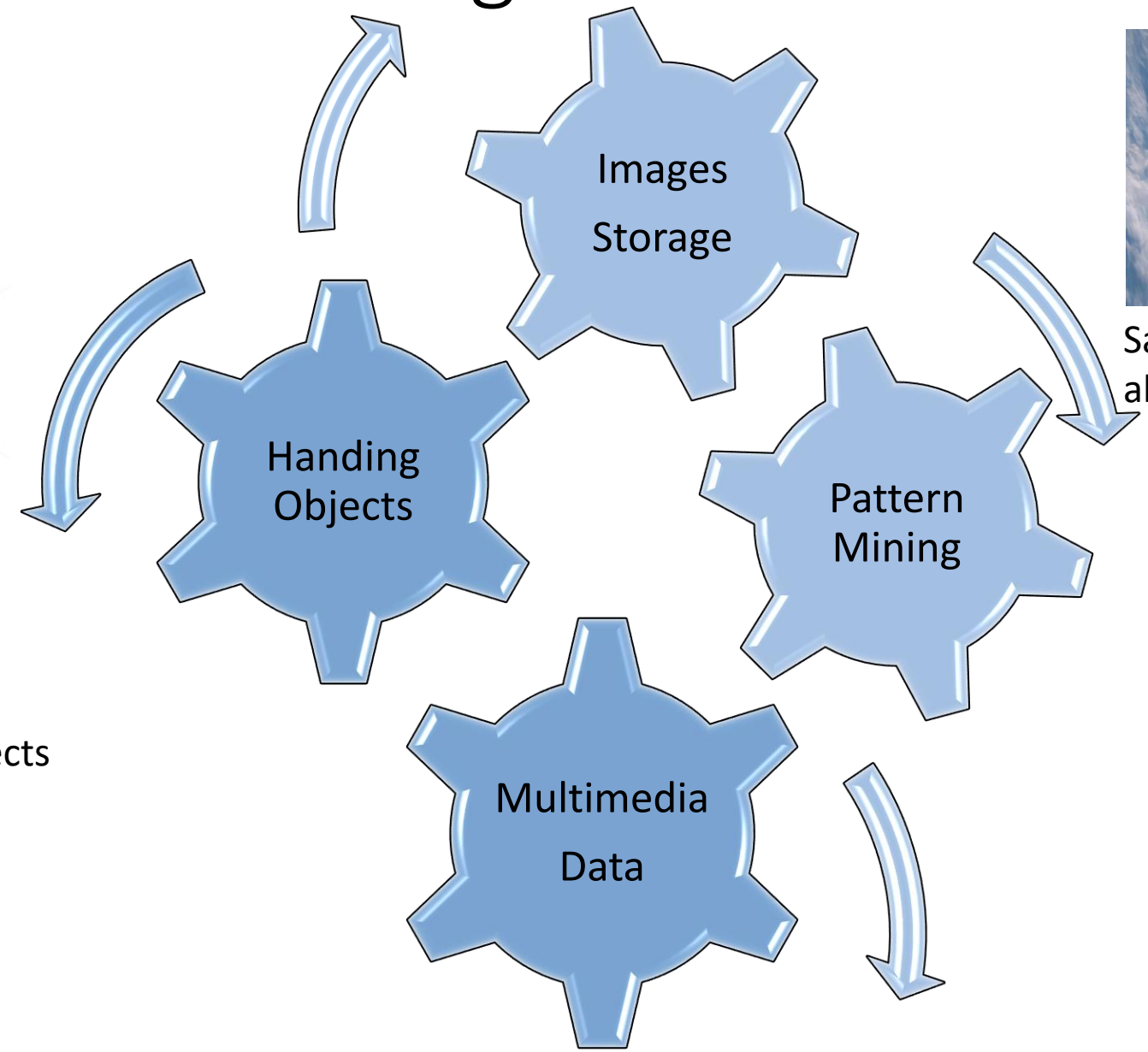
# What Does Future Hold? (late 1990s onwards)

# Future Looks Bright

- As 1990s unfold, computing and hardware will become cheaper
- As a result, relational databases will become cheaper too
- The databases management systems will be much more commonly used in various industries.
- But as computing resources become cheaper, the industry application problems will also expand at the same rate.

# Future Challenges in the Industry

Working with Complex DNA Sequence objects

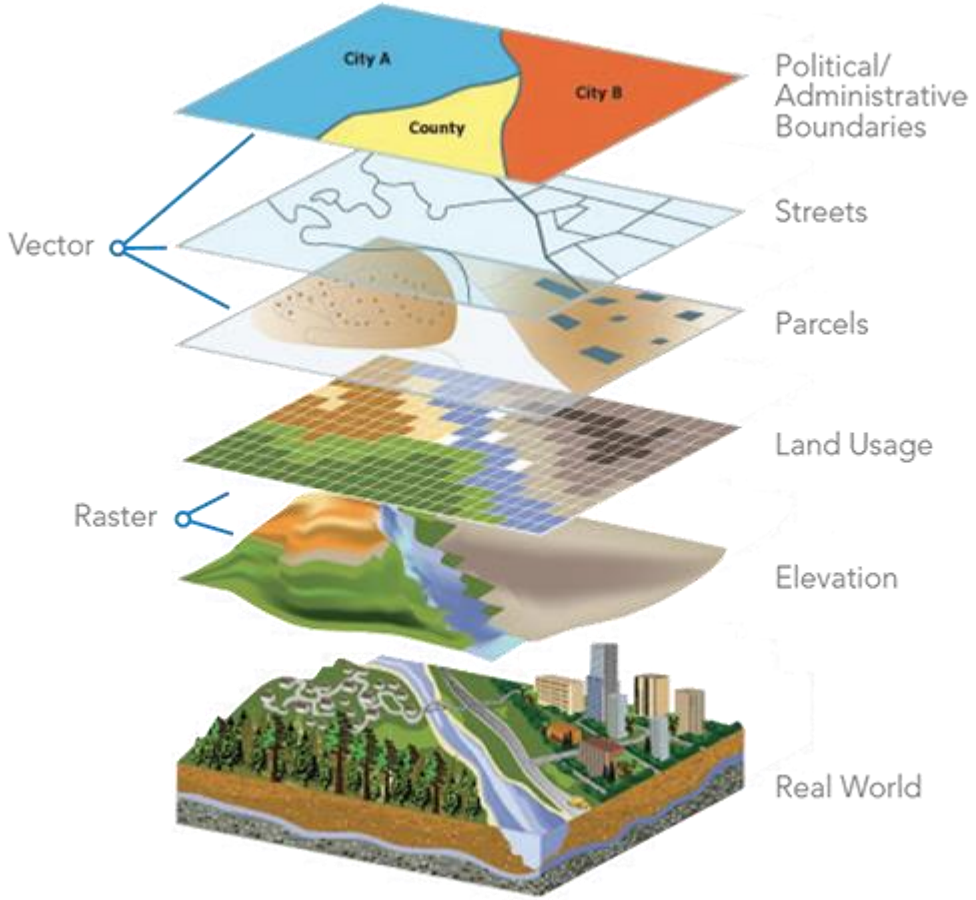




Satellite Images storage and allowing relevant search

Market Basket Analysis

# New Data Modeling Concepts – Spatial Data




learn.arcgis.com

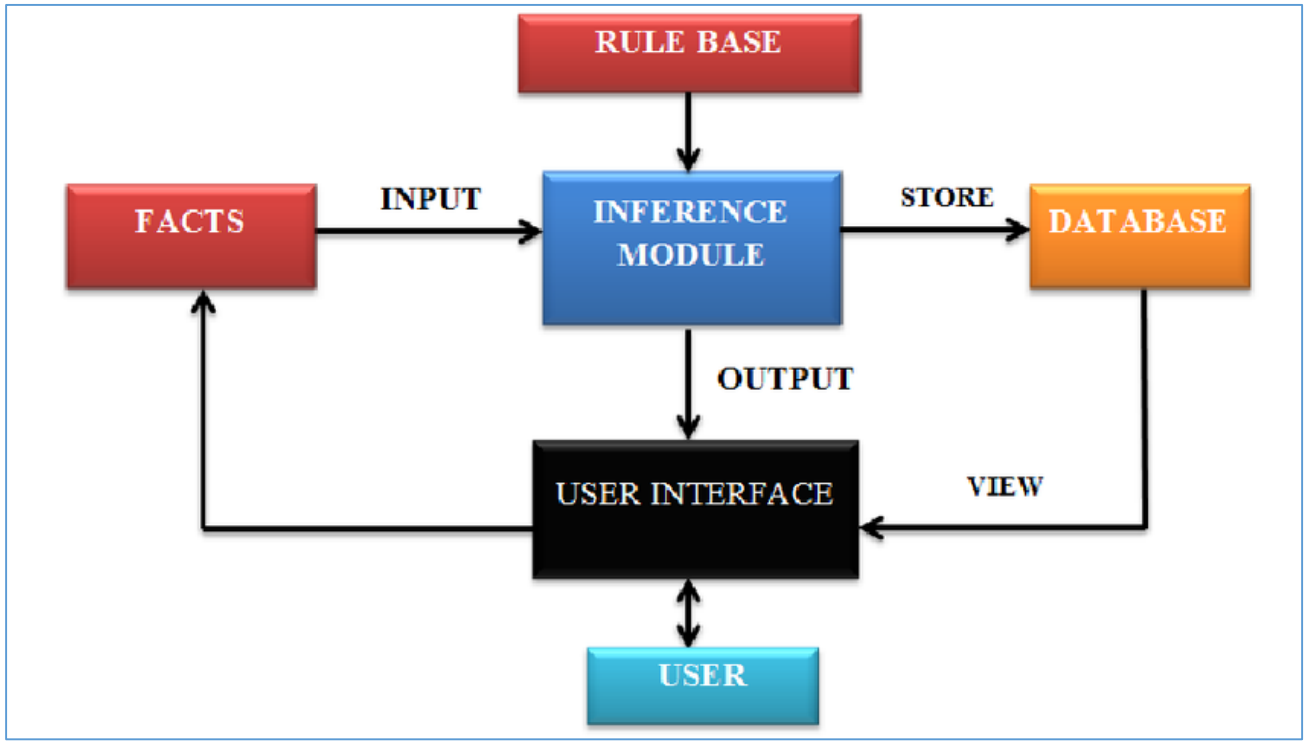
Possible Multi-Dimensional Complex Queries

Can we find 10 closest neighbours to a particular location in the city?

# Suggested Research Directions

# Rule Based System – Artificial Intelligence

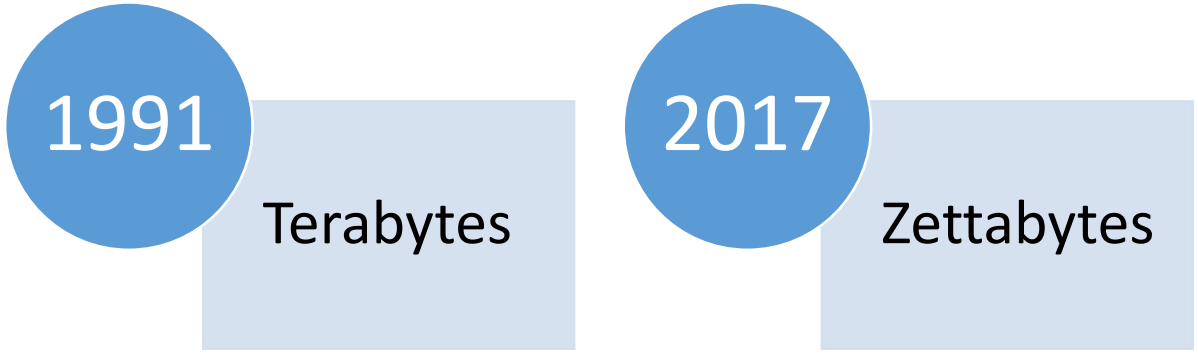
- Classifying a disease based on patients symptoms and lab results
- Provided facts are checked with the rule base by an inference module to come up with conclusions.
- In the future, the whole rule based system could be extensions of a DBMS.
- As the data structures increase over time, the traditional AI tools that assume all relevant data is in main memory wont work




**Reference:** A rule-based expert system for automated ECG diagnosis, Sept 2013

# Scaling Up

- All the DBMS Algorithms (concurrency control, transaction management) also need to scale up as required by the industry in the future




***"There were 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days."***

Eric Schmidt, of Google, said in 2010

# Tertiary Storage

- Ultra large databases will require archive storage
- Since higher latency would work for such data, tertiary storage would be the go-to options
- More research needs to continue as the expansion of the world data is continuing.

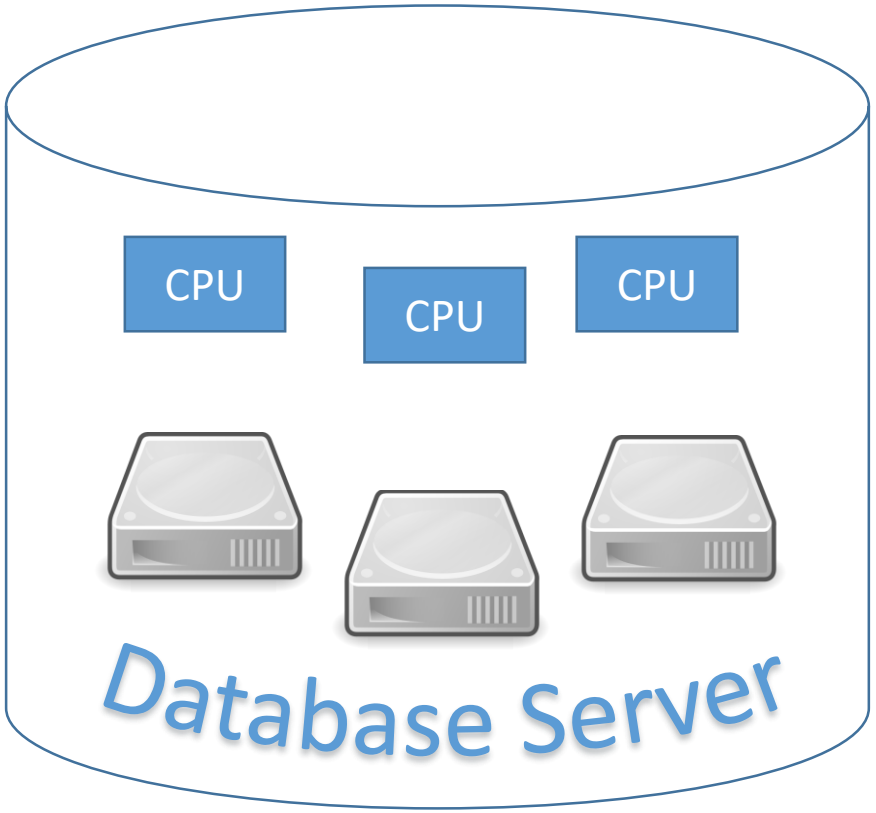
| Storage Device | Cost/MB | Latency Data Rate |
|---|---|---|
| Magnetic Disks | 20c | 25 ms 5-8 MBps |
| Tapes (low end) | 0.5c | 3 rain 1.5 MBps |
| Tapes (high end) | 0.7c | 3 rain 10 MBps |
| Optical Disks | 10c | 1.25 MBps |

Reference

Tertiary Storage Systems: Solving Multimedia DBMS Storage Problems
Greg Magsamen – September 27, 2007

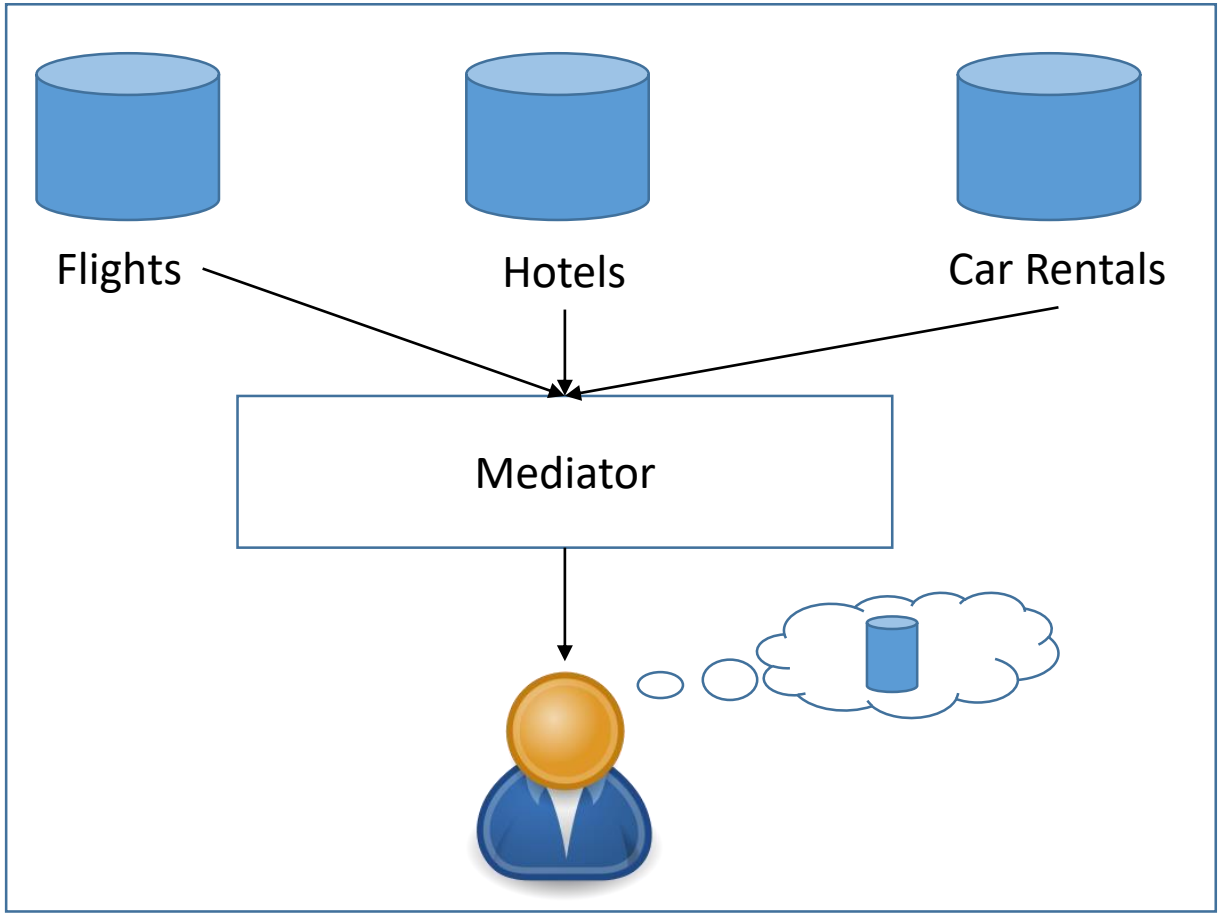
# Parallelism

- Can adding more CPU processors and adding more disks  to a database server infrastructure linearly improve the database query speed?
- To obtain linear speedup, research must be done to optimize the architecture to reduce overheads.
- This will be an evolving area considering optimizations and the reducing costs of hardware



# Heterogeneous, Distributed Databases

- The mediator fuses semantically inconsistent data
- User gets the feeling that interaction is only with a single database with a global schema




Source:
http://people.scs.carleton.ca/~bertossi/talks/dublin08.pdf

# Heterogeneous, Distributed Databases

- Interconnection of Information will be the key for collaborative applications Eg. The Human Genome Project, wikipedia
- Intercompany applications will also pose interoperability issues
- Security requirements for heterogeneous distributed databases could also be complex

# Conclusion

# Conclusion

- The inspirations of the database research suggested by this paper are evident such as the advent of the Big Data technologies.

- The database research still needs to continue even today for the advancement of various industries and evolving new technologies.

Reference
Andy Pavlo, History of Databases, CMU Database Group. Retrieved From https://www.youtube.com/watch?v=MyQzjba1beA

# Q&A