Databases Biosciences

Kei Cheung, Ph.D.
Professor
Department of Emergency Medicine
Yale Center for Medical Informatics

The 4th paradigm: data-intensive scientific discovery

- It expands the vision of Jim Gray (Mr. Database)
- His vision of a Personal Memex as well as a World Memex
  - Memex (originally coined by Vannevar Bush in 1945) is a device in which an individual stores all his books, records, and communications

Healthcare and life sciences data sources

Drug Research

Social Media

Patient Records

Gene Sequencing

Test Results

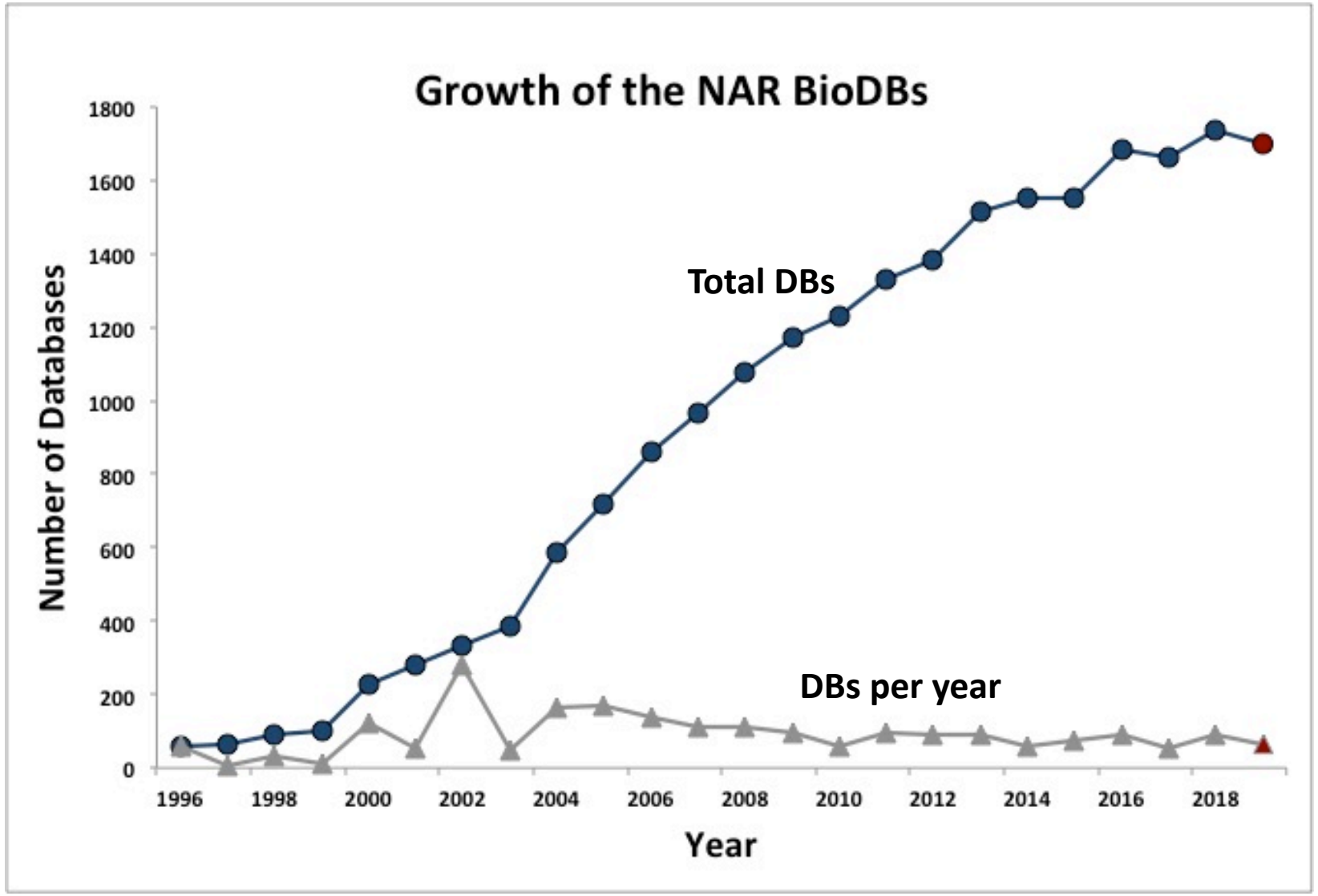
Claims

Home Monitoring

Mobile Apps

4Vs:

- Volume – high-throughput technologies
- Variety – diverse data types, different formats, structured vs. unstructured data
- Velocity – data streaming
- Veracity – trust worthiness of data


Growth of the NAR BioDBs
Number of Databases
1800
1600
1400
1200
1000
800
600
400
200
0
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018
Year
Total DBs
DBs per year

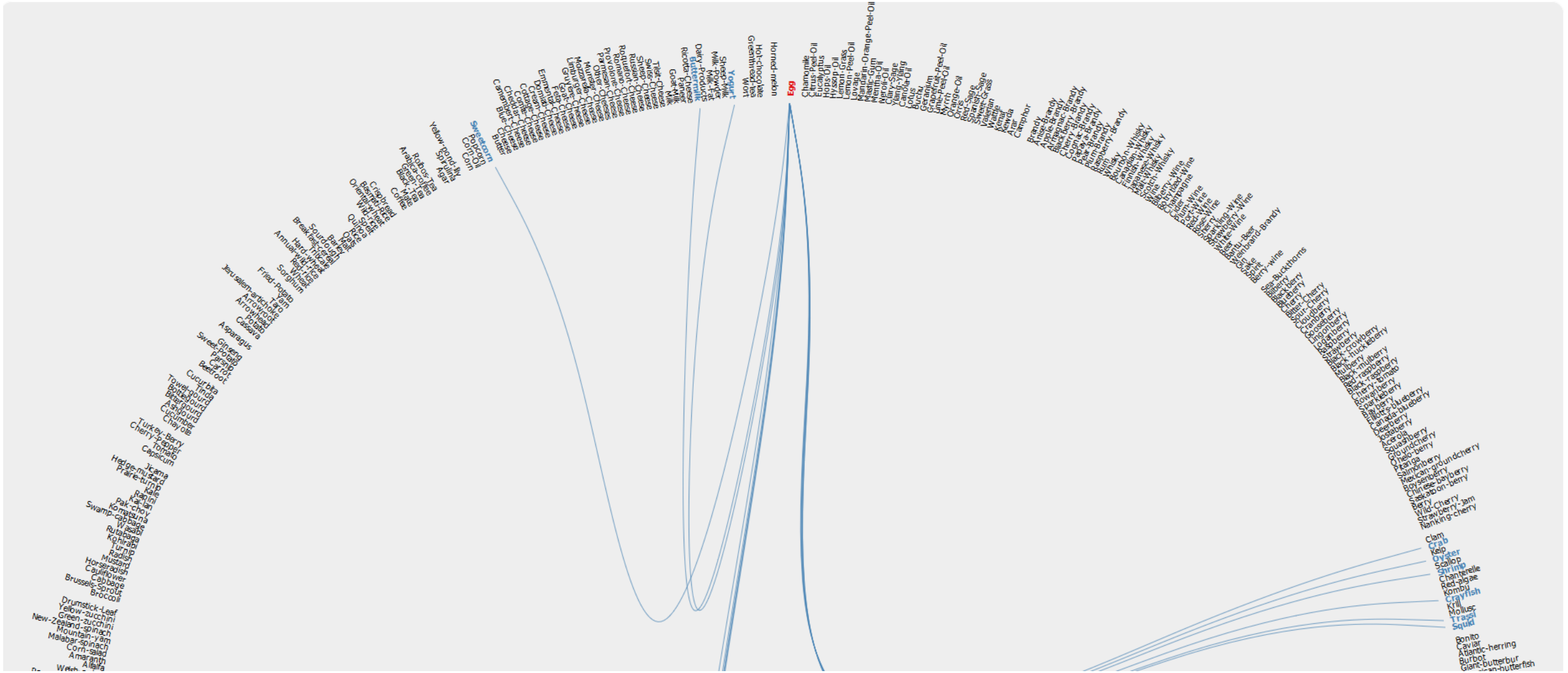
- Compilation Paper
- Category List
- Alphabetical List
- Category/Paper List
- Search Summary Papers

NAR Database Summary Paper

Nucleotide Sequence Databases
RNA sequence databases
Protein sequence databases
Structure Databases
Genomics Databases (non-vertebrate)
Metabolic and Signaling Pathways
ChemProt
FunCoup
Reactome
Enzymes and enzyme nomenclature
Metabolic pathways
BiGG Models
BioCarta Pathways
BioCyc
Bionemo
BioSilico
CeCaFDB
ClusterMine360
ECMDB
HMDB - The Human Metabolome Database
iPAVS
KaPPA-View
LAMP
MEROPS
MetaboLights
Metabolomics Workbench
MetaCrop
MMCD
MMMDB
MNXref/MetaNetX
MODOMICS
MultitaskProtDB-II
Pathguide
Pathway Commons
PMAP
Reactome
Rhea
RNApathwaysDB
SABIO-RK
SMPDB
SYSTOMONAS
UniPathway
WholeCellKB - Model Organism Databases for Comprehensive Whole-Cell Models
WikiPathways
YMDB
Protein-protein interactions
Signalling pathways
Human and other Vertebrate Genomes

The Flavor Network

Click on any Entity/Ingredient to see its flavor sharing



What is (not) a database?

- It’s not just a file
- It’s not just an Excel spreadsheet
- It’s an organized collection of related information that can easily be accessed, managed, and updated

Difference between Spreadsheet and Database

Spreadsheet	Database
Data analysis	Data management
Mathematical calculation	Structuring data and querying data to create subsets
Typically single user	Database management with multiple users
Formatting and chart display	Reports for data summarization
Limited in scale	Scalable

Worksheet size:	1,048,576 rows by 16,384 columns
Column width:	255 characters
Total no. of characters that a cell can have:	32,767 characters

Some key database concepts

- **Data integrity** is the assurance that data are correct and consistent (data correctly reflects the real world)
- **Data redundancy** occurs if data are duplicated between files
- **Data dependency** defines linkage between data files and their order of entry
- **Data security** refers to data being protected so that only authorized personnel can access them

Relational database (SQL database)

- The relational model was introduced by E.F. Codd in 1970, which is based on the mathematical set theory
- A relational database management system (RDBMS) is a computer application (software) of the relational data model (e.g., MS SQLServer, MySQL, Oracle, …)
- It has become an industry standard with a standard query language (SQL)
- Relational databases have widely been used to manage data in different domains

Components of Relational Database

- A table (relation) represents some class of objects (e.g., patients, doctors, drugs, hospitals)
- Each table consists of columns (attributes) and rows (tuples).
  - Each column represents some attribute of the object represented by the table (e.g., patient id, patient name)
  - Each row corresponds to an instance of the object represented by the table (e.g., each row in the Patient table represents a patient who has a specific patient id and name.)

How to organize data into tables

Keys

- Primary key: Every table should have a primary key comprising a single or multiple columns that contain unique values. A primary key is the unique identifier of a table row (e.g., “sample id” is the primary key for the **Sample** table)
- Foreign key: it is a key taken from a different table. For example, in the **Experiment** table, the “sample id” is the foreign key to the **Sample** table.

Addition, Deletion and Modification Anomalies

Student ID	Name	Address	Subject
401	Adam	Noida	Biology
402	Alex	Panipat	Math
403	Stuart	Jammu	Math
404	Adam	Noida	Physics

Normalization

- Normalization is a *process* in which we systematically organize columns and tables to eliminate anomalies due to data redundancy
- It involves decomposing a (de-normalized) table into less redundant (smaller) tables without losing information
- The objective is to isolate data so that additions, deletions, modifications of data can be made in just one table and then propagated to other tables using foreign keys.
- Normalization is a trade-off between data redundancy and performance.
  - Normalizing a table reduces data redundancy but introduces the need for joins when all of the data is required for a report query.
- **Normal Form**: A set of tables free from a certain set of addition, deletion and modification anomalies.

Different Normal Forms

- **First normal form (1NF)**
- **Second normal form (2NF)**
- **Third normal form (3NF)**
- Boyce-Codd normal form (BCNF)
- Fourth normal form (4NF)
- Fifth normal form (5NF)
- Domain-Key normal form (DK/NF)
- …

First Normal Form

- Each column value must be a single value only.
- All values for a given column must be of the same data type.
- Each column name must be unique.
- The order of columns is insignificant
- The order of the rows is insignificant
- No two rows in a table can be identical.

First Normal Form Example

ID	Student	Age	Subject
401	Adam	15	Biology
404	Adam	15	Physics
402	Alex	14	Math
403	Stuart	17	Math

Second Normal Form

- A table is in second normal form (2NF) if it is in 1NF and if all of its non-key columns are dependent on all of the *key*.
  - A table is in second normal form if it is free from partial-key dependencies
- Tables that have a single column for a key are automatically in 2NF.
  - This is one reason why we often use artificial identifiers (non-composite keys) as keys.
- To achieve second normal form, we may need to split a table into multiple tables and match rows between tables using primary and foreign keys

Second Normal Form Example

Student	Age
Adam	15
Alex	14
Stuart	17

Enroll_id	Student	Subject
1	Adam	Biology
2	Adam	Physics
3	Alex	Math
4	Stuart	Math

Third Normal Form

- Every non-primary key column must be dependent on primary key
- There should not be the case that a non-primary key column is determined by another non-primary key (*transitive dependency)*
  - Student (ID, Name, DOB, City, State, Zip)
- *A table is in 3NF if the following are true:*
  - *it is in 2NF*
  - *All transitive dependencies* are removed

Student (ID, Name, DOB, Zip)

Address (Zip, City, State)

Entity Relationship Diagram (ERD)

What is ERD

- It is a data model associated with a diagrammatic method (P. Chen 1976) used to conduct/view data modeling
- It describes the attributes of and the relationship between entities (data objects)
- DBA uses ERD to perform data modeling and explain the diagram to stakeholders

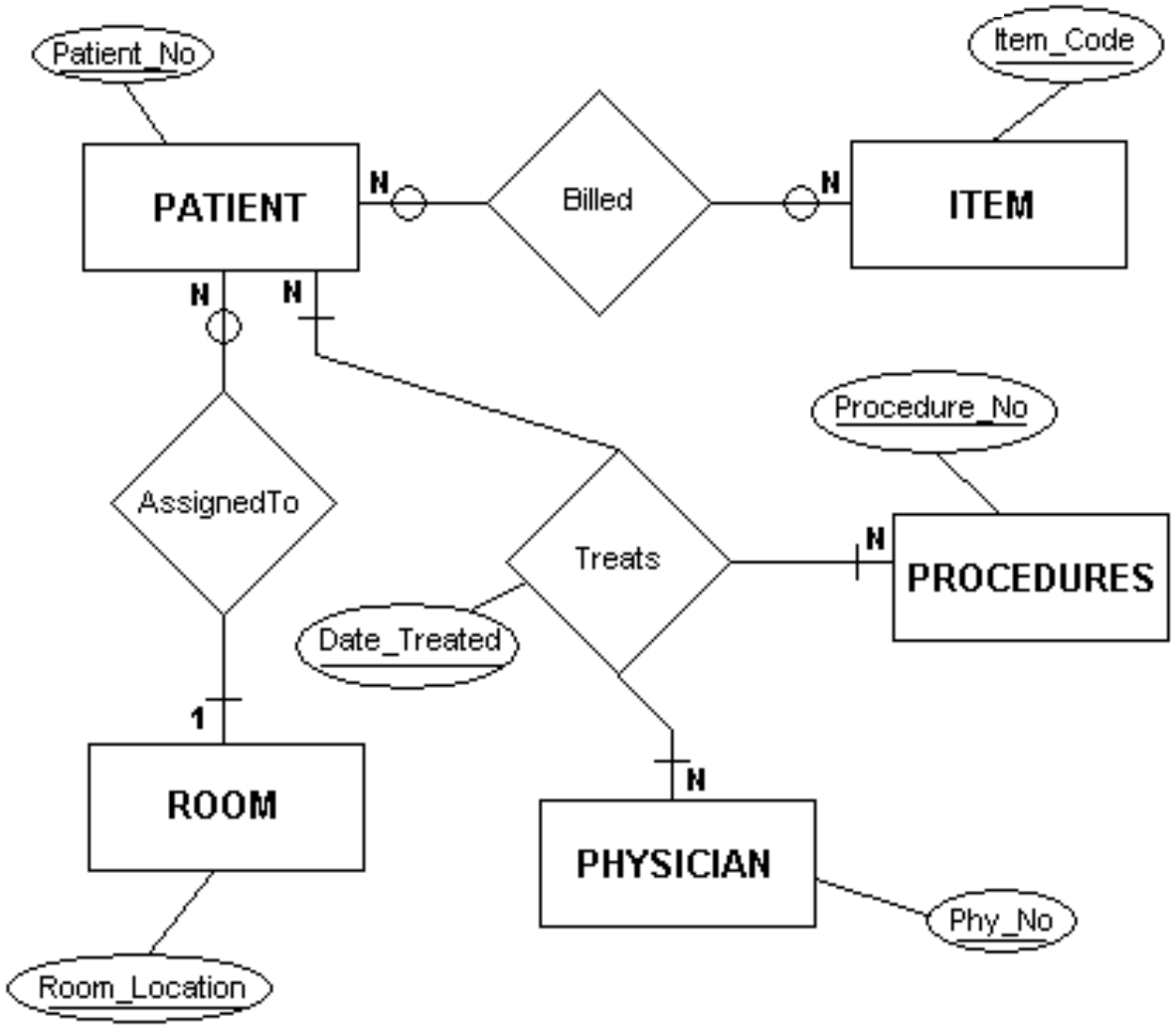
Primary Components of ERD

- **Entity** represents a collection of objects in the real world (e.g., person, place, event)
- **Attribute** is a named property or characteristic of an entity
- **Relationship** is an association between the instances of one or more entities

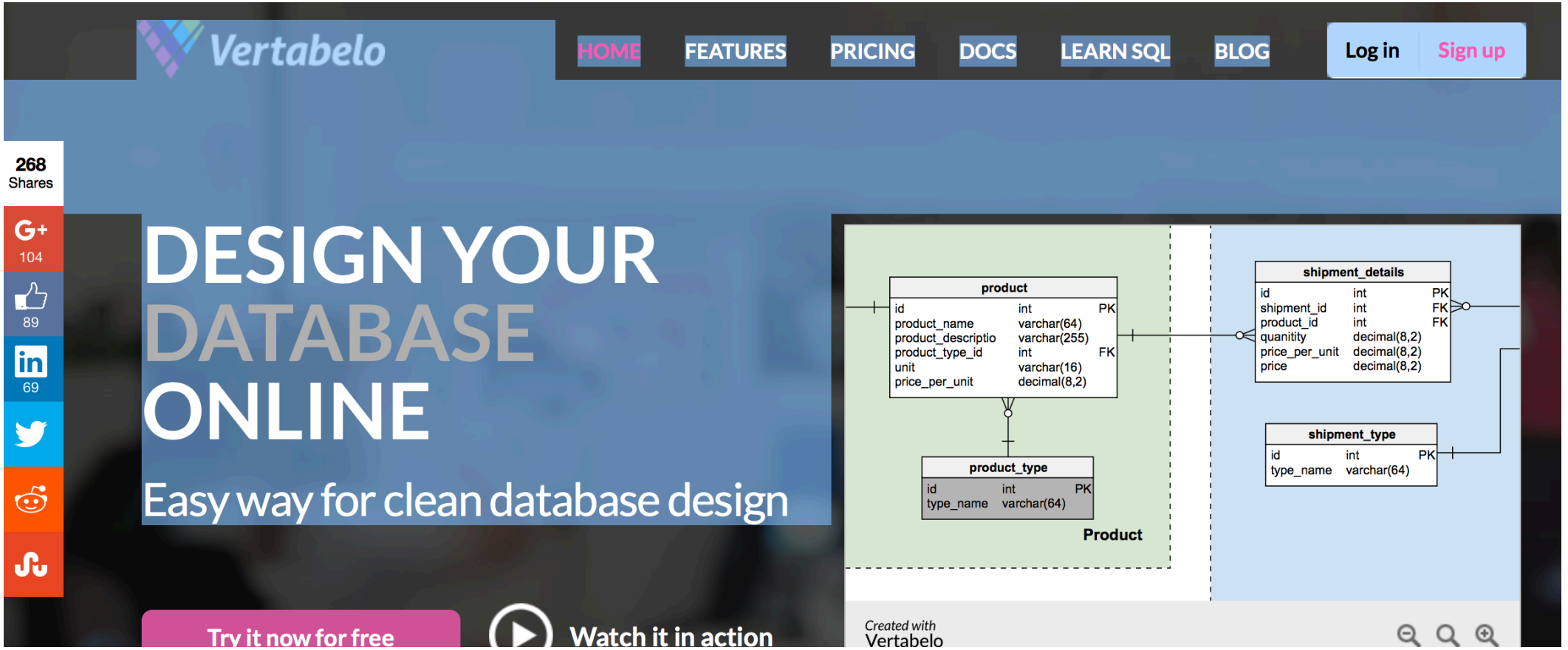
Relationship Cardinality

- It expresses the minimum and maximum number of occurrences of one entity for a single occurrence of the other
  - One-to-One (1:1)
  - One-to-Many (1:N)
  - Many-to-Many (M:N)

Example ERD (Hospital Database)



Vertabelo (https://www.vertabelo.com/)



Data index

- It is a data structure that is added to a file to provide faster access to the data (search)
- It reduces the number of block that the DBMS has to check

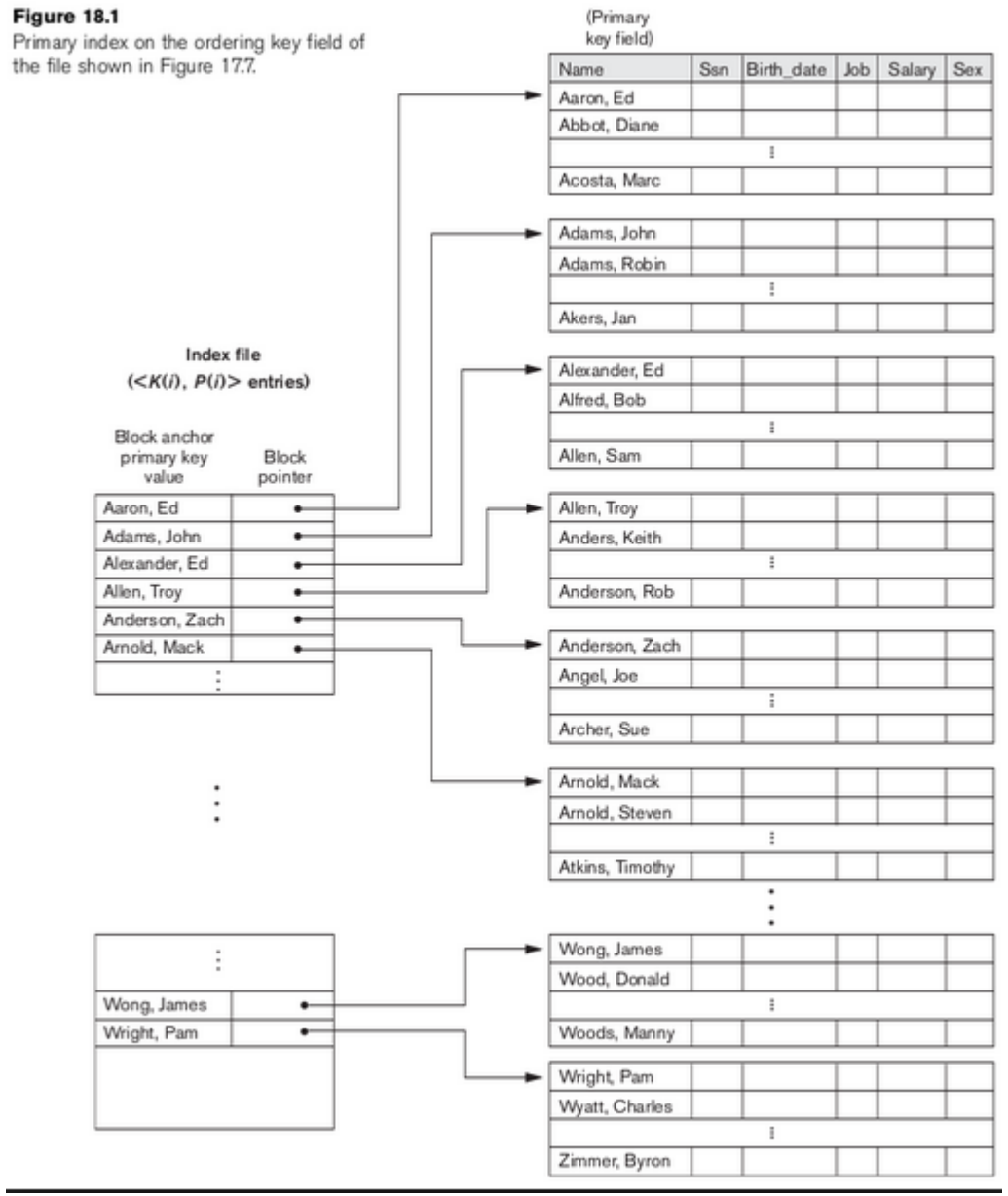
Properties of data index

- It contains a search key and a pointer
- Search key – an attribute or set of attributes that is used to look up the records in a file
- Pointer – contains the address of where the data is stored in memory.
- It can be compared to the card catalog system used in the libraries

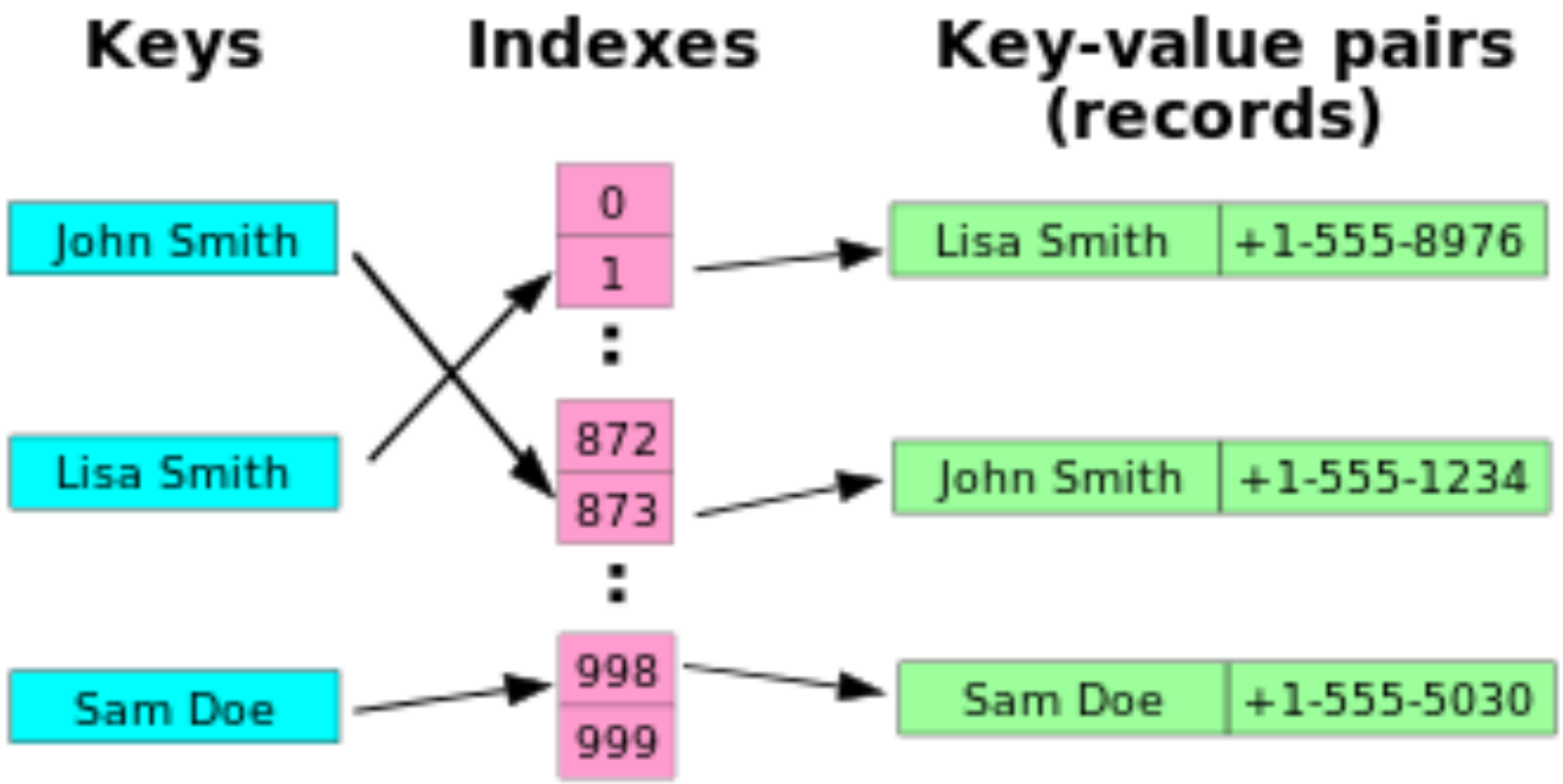
Two types of indices

- Ordered index (primary index or clustering index) – which is used to access data sorted by order of values
- Hash index (secondary index or non-clustering index) – used to access data that is distributed uniformly across a range of buckets

Ordered index



Hash index

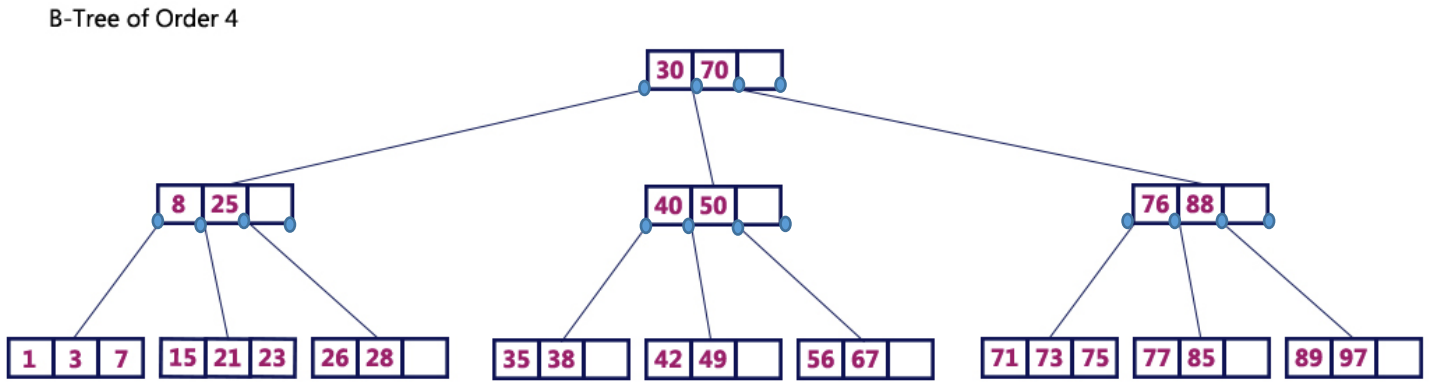

Keys
Indexes
Key-value pairs (records)
John Smith
Lisa Smith
Sam Doe
0
1
872
873
998
999
Lisa Smith
+1-555-8976
John Smith
+1-555-1234
Sam Doe
+1-555-5030

How to choose indexing technique

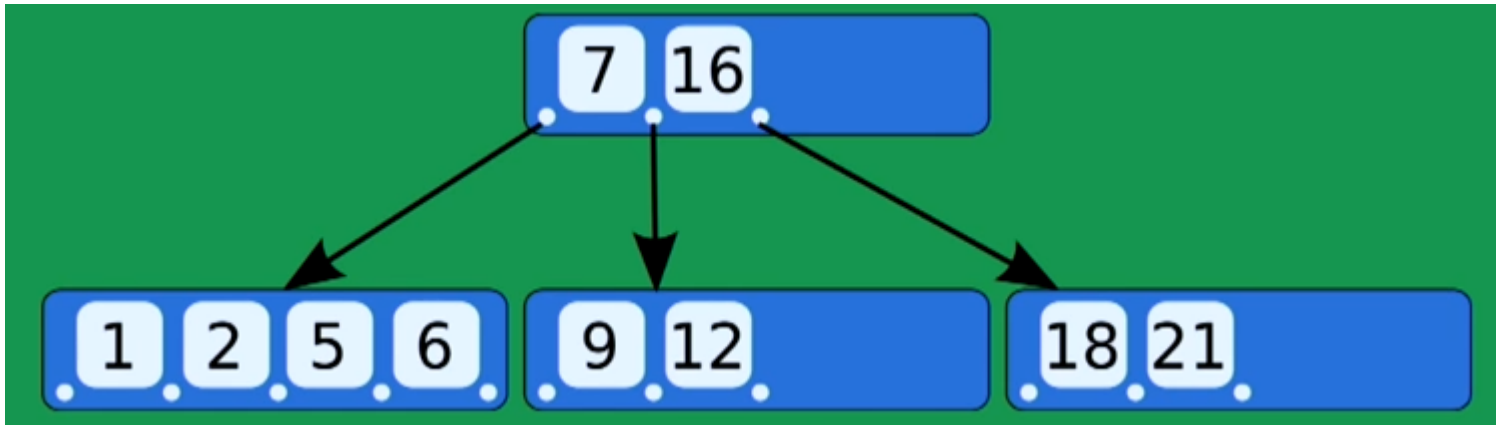
- Five Factors involved when choosing the indexing technique:
  1. access type
  2. access time
  3. insertion time
  4. deletion time
  5. space overhead

B tree

- A B tree of order m:
  - Every node has at most m children
  - A non-leaf node with k children contains k-1 keys
  - The root has at least two children if it is not a leaf node
  - Every non-leaf node (except root) has at least m/2 children
  - All leaves appear in the same level



B tree: what is the order of the B tree below



B tree visualization: https://www.cs.usfca.edu/~galles/visualization/BTree.html

On-Line Transaction Processing (OLTP)

What is OLTP?

- It is a class of information systems (e.g., databases) that facilitate and manage transaction-oriented applications, typically for data entry and retrieval transactions
- A database that is based on a normalized relational model is considered an OLTP application. It supports the following transactions:
  - Insert new rows
  - Update existing rows
  - Delete rows
  - Select rows
- A database transaction must be atomic, consistent, isolated and durable (ACID)

Structured Query Language (SQL)

- It is a standard programming language for creating (CREATE) relational databases and tables as well as retrieving (SELECT), adding (INSERT), deleting (DELETE) and updating (UPDATE) data in a relational database
- It is compliant with ANSI and ISO standards

SQL Statement (CREATE DATABASE/TABLE)

```
CREATE DATABASE Patient_DB;

CREATE TABLE Patient_DB.Patient
(
        ID int,
        Name varchar (50),
        Address varchar (250),
        Age smallint
        Sex varchar (2)
);
```

INSERT Statement

INSERT INTO Patient_DB.Patient

(ID, Name, Address, Age, Sex)

VALUES (1, ‘John Doe’, ‘XYZ’, 40, ‘M’)

…

ID	Name	Address	Age	Sex
1	John Doe	XYZ	40	M
2	Jane Smith	ABC	34	F
3	Mary Queen	PQSRT	46	F
4	Mike Lee	DWQER	60	M

UPDATE Statement

```
UPDATE Patient_DB.Patient
SET AGE=41
WHERE ID=1
```

ID	Name	Address	Age	Sex
1	John Doe	XYZ	41	M
2	Jane Smith	ABC	34	F
3	Mary Queen	PQSRT	46	F
4	Mike Lee	DWQER	60	M

DELETE Statement

```
DELETE Patient_DB.Patient
WHERE Name='Mike Lee'
```

ID	Name	Address	Age	Sex
1	John Doe	XYZ	41	M
2	Jane Smith	ABC	34	F
3	Mary Queen	PQSRT	46	F

SELECT Statement

```
SELECT ID, Name, Age, Sex
FROM Patient_DB.Patient
WHERE Age>=40
ORDER BY Age
```

ID	Name	Address	Age	Sex
1	John Doe	XYZ	40	M
2	Jane Smith	ABC	34	F
3	Mary Queen	PQSRT	46	F
4	Mike Lee	DWQER	60	M

SELECT Statement (Aggregation)

```
SELECT Sex, avg(Age)
FROM Patient_DB.Patient
GROUP BY SEX
```

Results: M 50
F 40

ID	Name	Address	Age	Sex
1	John Doe	XYZ	40	M
2	Jane Smith	ABC	34	F
3	Mary Queen	PQSRT	46	F
4	Mike Lee	DWQER	60	M

SELECT Statement (JOIN)

```
SELECT A.*, B.Report_Text
FROM Patient_DB.Patient AS A
INNER JOIN Patient_DB.LabTest. AS B
ON A.ID = B.Patient_ID
```

ID	Name	Address	Age	Sex
1	John Doe	XYZ	40	M
2	Jane Smith	ABC	34	F
3	Mary Queen	PQSRT	46	F
4	Mike Lee	DWQER	60	M

Patient_ID	ID	Report_Text
1	1	……
1	2	…….

Other Types of SQL Statements

- TRUNCATE TABLE
- DROP TABLE
- CREATE VIEW
- CREATE INDEX (boost query performace)
  - Full-Text index (e.g., part of MS SQLServer)

From OLTP to OLAP (On-Line Analytical Processing)

OLAP Overview

- OLTP databases are tuned to small/medium size of data with relatively simple queries
- Some applications use fewer but more time-consuming analytic queries
- New architectures (data warehouses) have been developed to handle such analytic queries efficiently (De-normalization)

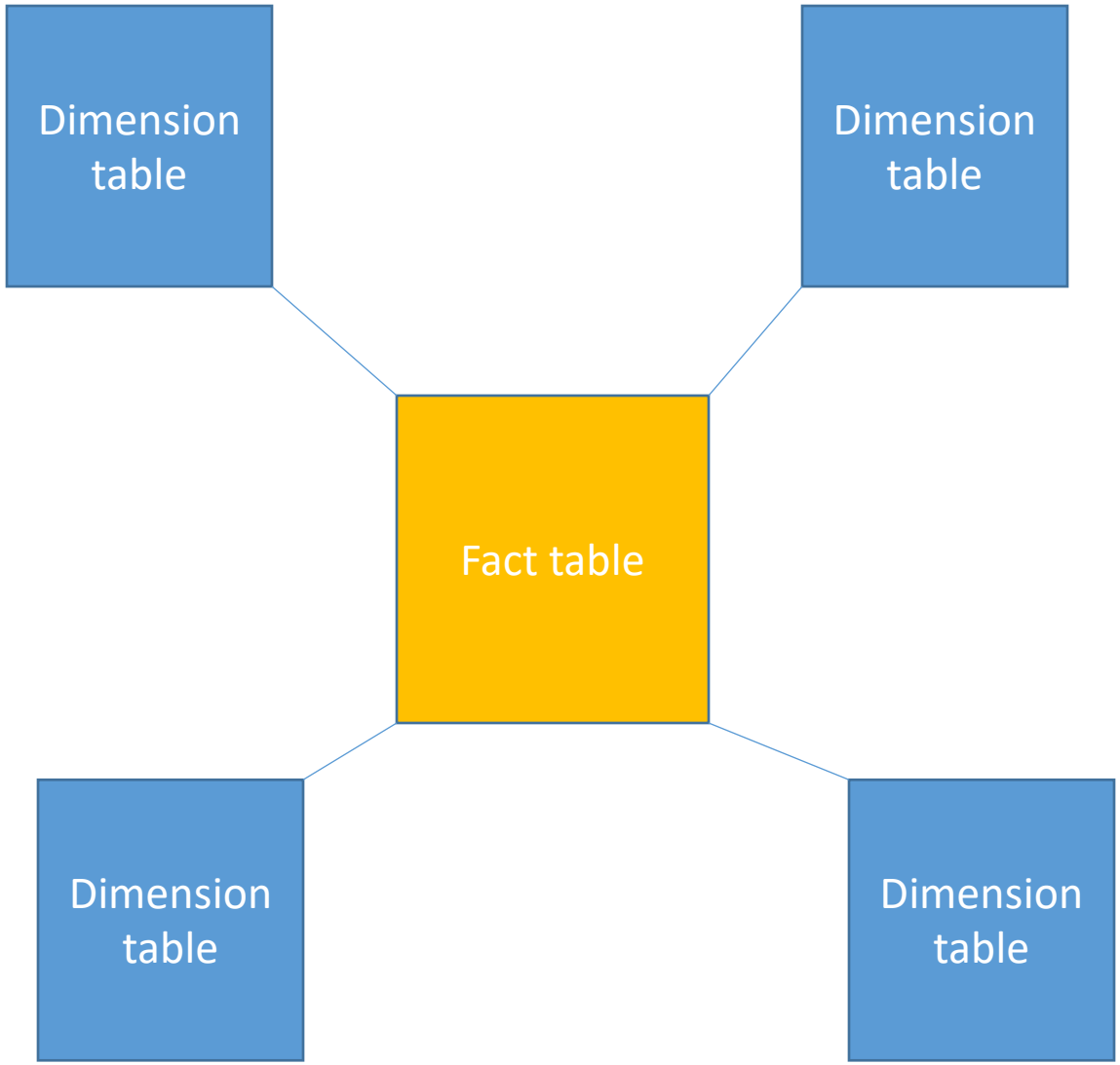
OLAP Example Queries

- Amazon analyzes purchases by its customers to identify products of likely interest to customers
- Analysts at Wal-Mart look for merchandise items with increasing sales in some region

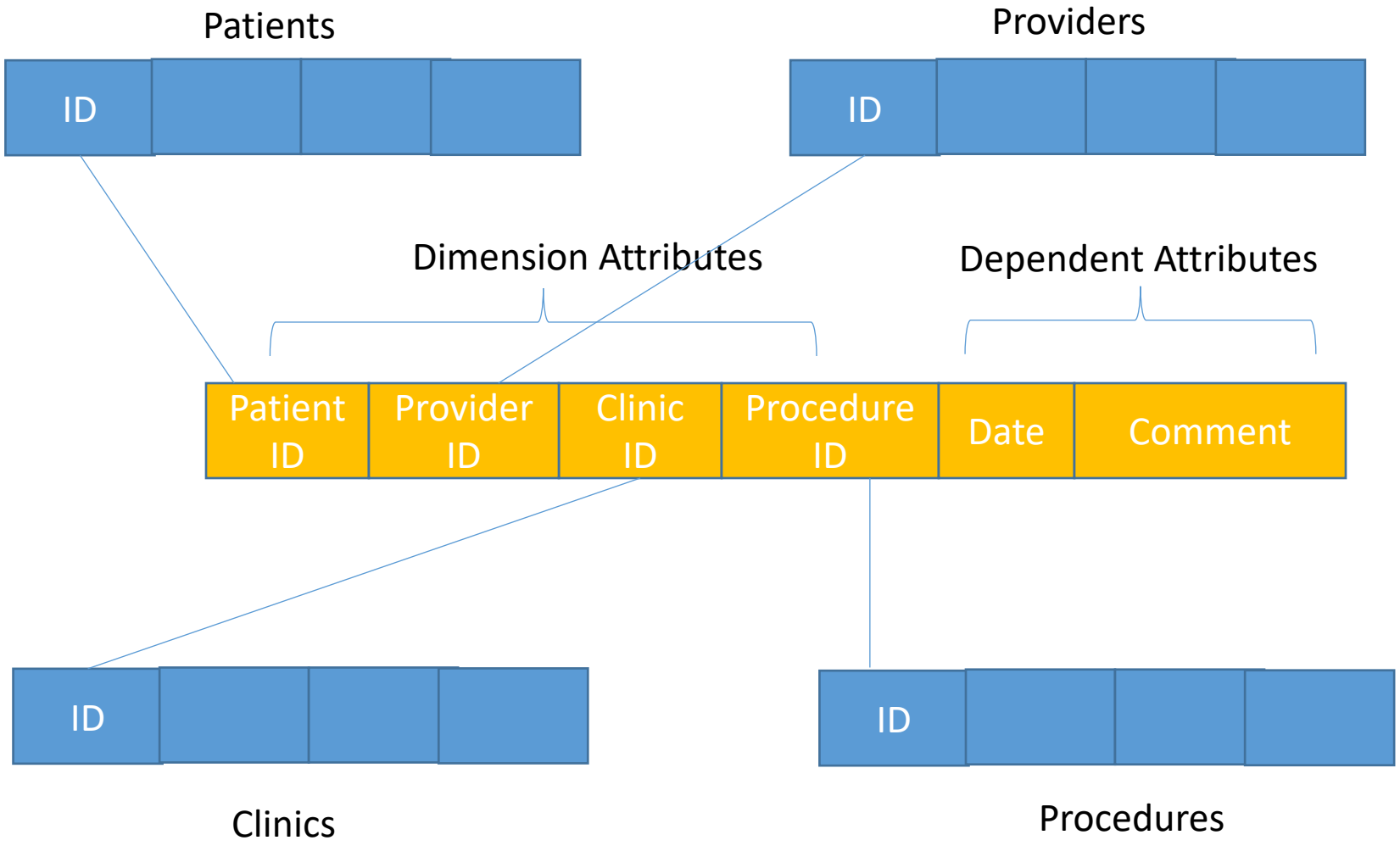
Data Warehouse

- The most common form of database integration
  - Copy source databases into a single database (data warehouse)
  - Update the data warehouse periodically (in batch mode)
  - Support analytic queries using a dimensional data model (vs. a normalized entity-relationship model)
- Example: VA CDW

Star Schema


Dimension table
Dimension table
Fact table
Dimension table
Dimension table

Star Schema Example



Example Queries

- Compare numbers of patient visits across different clinics for a given year
- Which are the top 10 most performed procedures among all clinics from 2010 to 2014

Beyond SQL

- NoSQL (graph databases like NEO4J, document databases like MongoDB)
- Semantic Web (standards for linked data and ontologies)

The End