Database Concepts

Drs. Antonio A. Trani and D. Gallagher
Dept. of Civil and Environmental Engineering

Why Learn Databases?

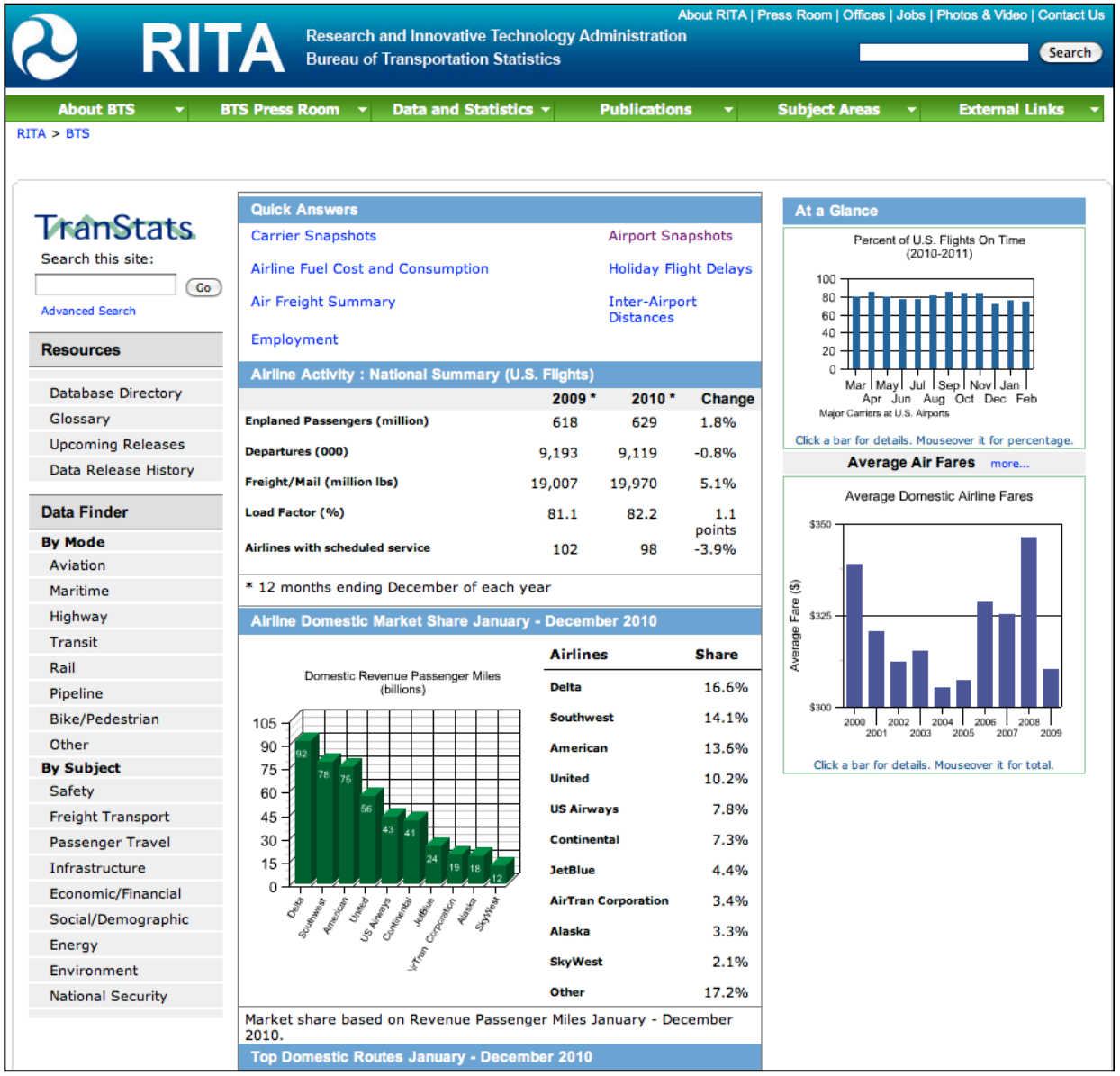
- Engineers need to manipulate large amounts of data
- Data sometimes comes in a variety of formats
- Data is both numeric and character or “string” data
- Databases simplify the handling of various data sets
  - Organizes information
  - Can create simple interfaces to enter and retrieve data

Database Management Systems

- Combination of computer hardware and software designed to (collect), organize, store, manipulate, and analyze data
- Database is collection of persistent data
- Purpose of database is to store info about certain types of objects termed entities or objects
- The Web is full of good examples of databases
  - Bureau of Transportation Statistics (BTS)
    - http://www.transtats.bts.gov/
  - National Geologic Map Databases (USGS)
    - http://ngmdb.usgs.gov/Other_Resources/rdb_es.html

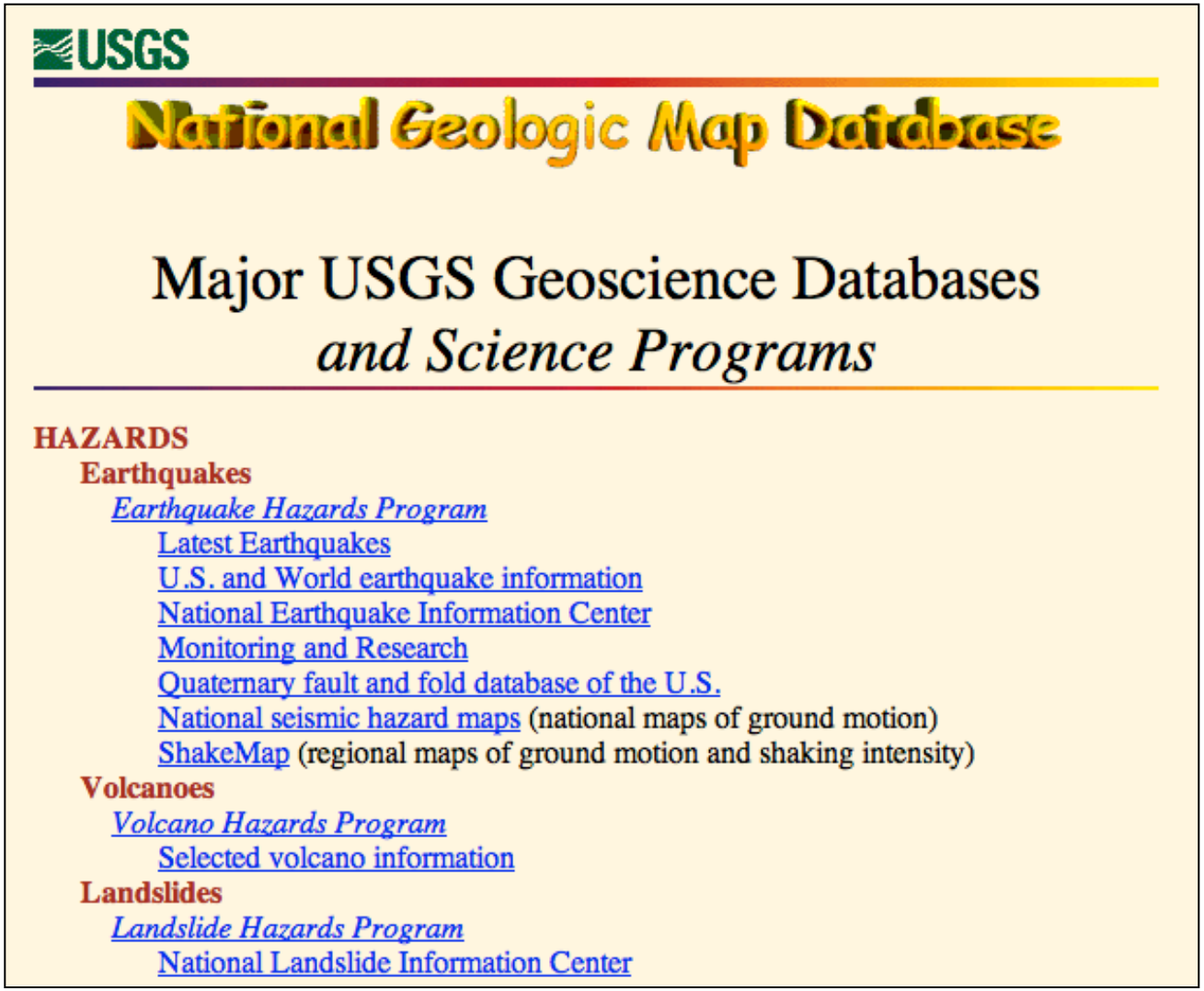
Bureau of Transportation Statistics (BTS)

http://www.transtats.bts.gov/



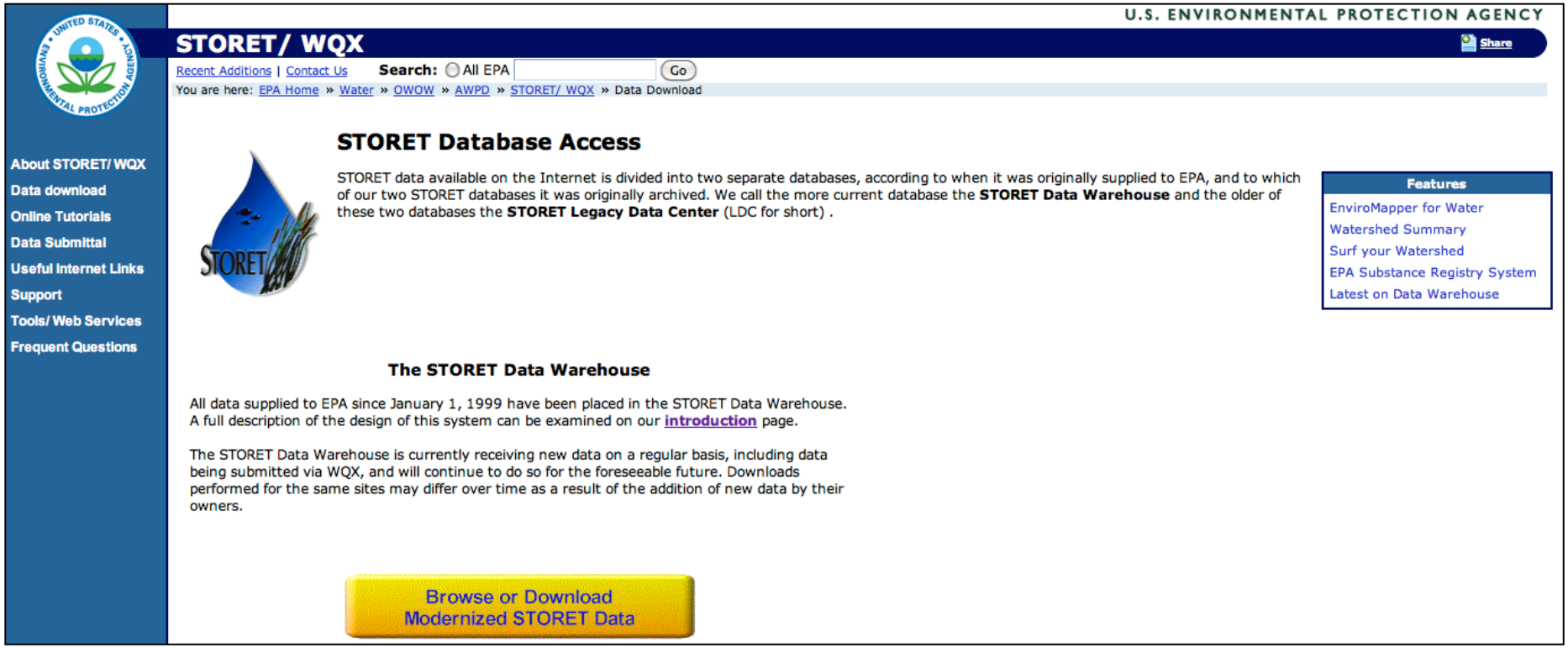
National Geologic Map Databases (USGS)

http://ngmdb.usgs.gov/Other_Resources/rdb_es.html



USEPA Water Quality Monitoring Data

http://www.epa.gov/storet/dbtop.html



Database Entities

- Entity
  - object about which database is designed to store information
  - for example, a water quality monitoring database, entities might include stations, pollutants, samples, analytical methods.
- Entity class or class
  - entire collection of all possible entities of a specific type that database might contain. An abstract description of an object
- Entity set
  - collection of specific entities that the database currently contains

Database Entity Attributes

- Properties of entities in entity class
  - Attributes usually stored as fields(columns) in table
  - Attributes used to uniquely identify individual entities within an entity class
  - Example: attributes of station may include station ID, River mile (location), date established, date removed, owner...
- Attributes also used to describe relationships to entities in other entity classes

Relational Database Programs

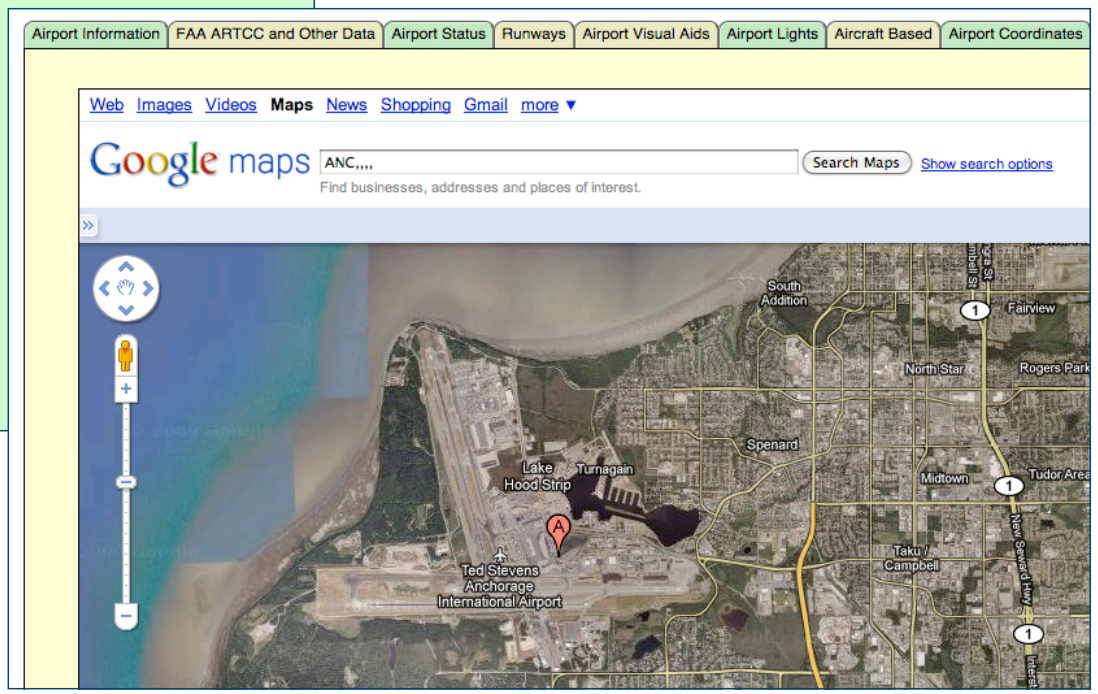
- **Oracle** (http://www.oracle.com/us/index.html)
- **Microsoft Access** (http://office.microsoft.com/en-gb/access/)
- **Filemaker** (http://www.filemaker.com/)
- **MySQL** (http://www.mysql.com/)
- **Microsoft SQL Azure** (http://www.microsoft.com/en-us/sqlazure/database.aspx)

Sample Airport Database

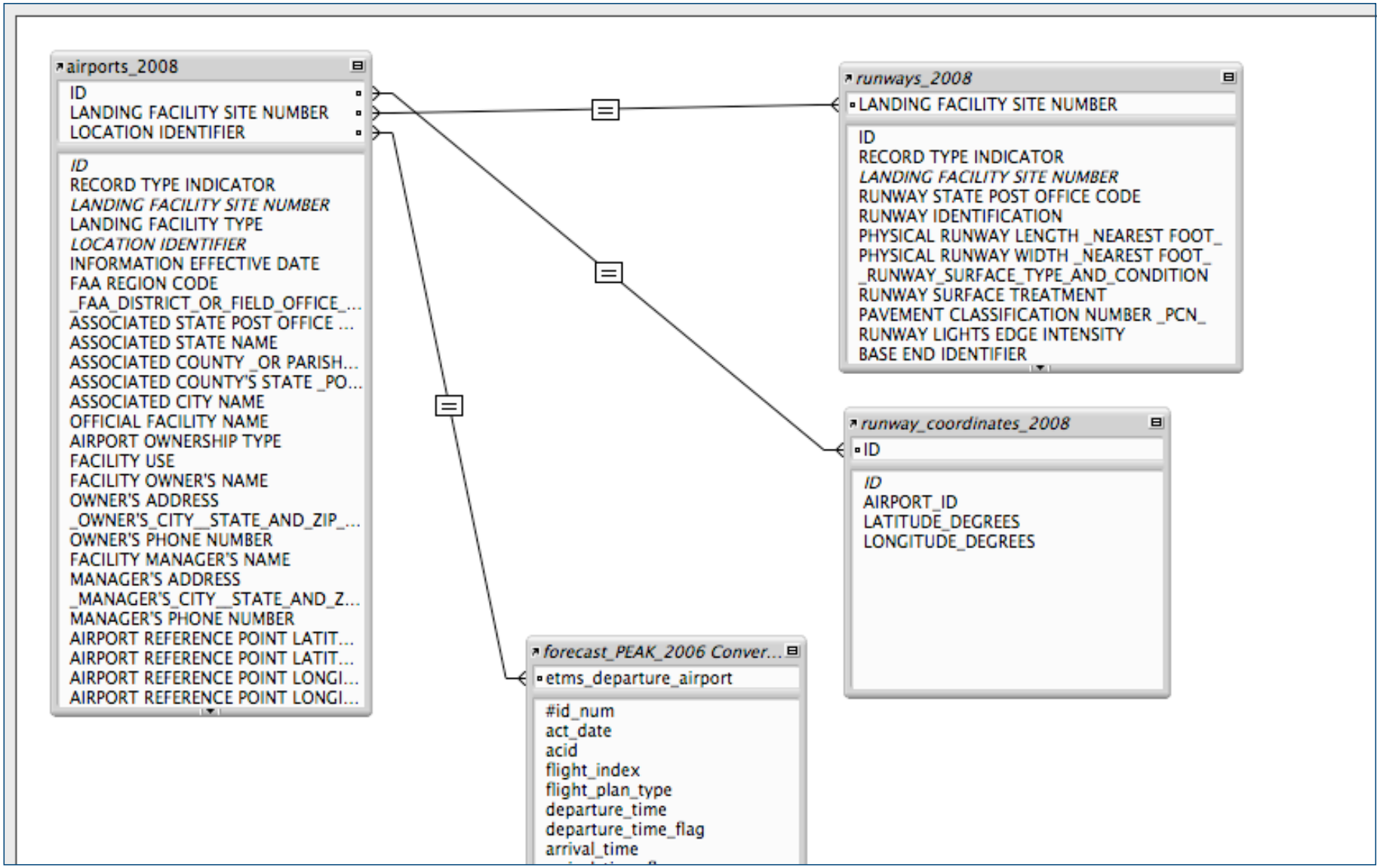
- Created in Filemaker Pro using publicly available data from the Federal Aviation Administration

Airport Information | FAA ARTCC and Other Data | Airport Status | Runways | Airport Visual Aids | Airport Lights | Aircraft Based | Airport Coordinates

LOCATION IDENTIFIER	ANC
OFFICIAL FACILITY NAME	TED STEVENS ANCHORAGE INTL
ID	26
RECORD TYPE INDICATOR	APT
LANDING FACILITY SITE NUMBER	50034.*A
LANDING FACILITY TYPE	AIRPORT
INFORMATION EFFECTIVE DATE	2/14/2008
FAA REGION CODE	AAL
_FAA_DISTRICT_OR_FIELD_OFFICE_CODE	NONE
ASSOCIATED STATE POST OFFICE CODE	AK
ASSOCIATED STATE NAME	ALASKA
ASSOCIATED COUNTY _OR PARISH_ NAME	ANCHORAGE
ASSOCIATED COUNTY'S STATE _POST OFFICE CODE_	AK
ASSOCIATED CITY NAME	ANCHORAGE
AIRPORT OWNERSHIP TYPE	PU
FACILITY USE	PU
FACILITY OWNER'S NAME	STATE OF ALASKA DOTPF
OWNER'S ADDRESS	BOX 196960
_OWNER'S_CITY__STATE_AND_ZIP_CODE	ANCHORAGE, AK 99519-6960
OWNER'S PHONE NUMBER	907-266-2525
FACILITY MANAGER'S NAME	MORTON PLUMB JR
MANAGER'S ADDRESS	BOX 196960
_MANAGER'S_CITY__STATE_AND_ZIP_CODE	ANCHORAGE, AK 99519-6960
MANAGER'S PHONE NUMBER	907-266-2525



Relationships in Airport Database



Entities in the Airport Database

- The database contains various datasets such as:
  - Airports
  - Runways
  - Latitude/Longitude coordinates, etc.

- All these “flat files” (simple two dimensional arrays) contain a common attribute or field

- For example: the airports 2008 and runways 2008 files have a common field called: “landing facility site number”

- This field establishes the relational structure between them

Database Functions in Excel (Simple Functions)

- Functions that provide pivot table functionality
- **Dfunction(database,field,criteria)**
  - **database:**
    - range of cells that make up the data
  - **field:**
    - name of field to perform operation
  - **criteria:**
    - range of cells that hold the criteria you want to hold with

Database Functions (List of Functions)

- The list functions include the following:
  - DAVERAGE(), DCOUNT(), DCOUNTA(), DGET(), DMAX(), DMIN(), DPRODUCT(), DSTDDEV(), DSTDDEVP(), DSUM(), DVAR(), and DVARP().
  - Where:
    - DCOUNTA(): Returns the count of nonblank records
    - DGET(): Returns the value of a specified field for a single matching record
- **Example:**
  - Cardata.xls

Database Example (Cardata.xls)

The syllabus has a file with car data
Name of the file Cardata.xls

	B	C	D	E	F	G	H
1	**Country**	**Type**	**Weight**	**Turning Circle**	**Displacement**	**Horsepower**	**Gas Tank Size**
2	Japan	Small	2700	37	112	130	13.2
3	Japan	Medium	3265	42	163	160	18
4	Other	Medium	2935	39	141	130	21.1
5	Other	Compact	2670	35	121	108	15.9
6	Other	Compact	2790	35	141	130	15.9
7	Other	Compact	2895	35	152	168	16.4
8	Other	Medium	3640	39	209	208	21.1
9	USA	Medium	2880	41	151	110	15.7
10	USA	Large	3350	43	231	165	18
11	USA	Large	3325	42	231	165	18
12	USA	Medium	3465	41	231	165	18.8
13	USA	Compact	2640	39	151	110	13.6
14	USA	Large	4285	44	307	140	25
15	USA	Large	3545	43	273	180	18
16	USA	Medium	3480	42	273	180	18.8
17	USA	Large	4025	42	262	150	27
18	USA	Compact	2655	38	133	95	15.6

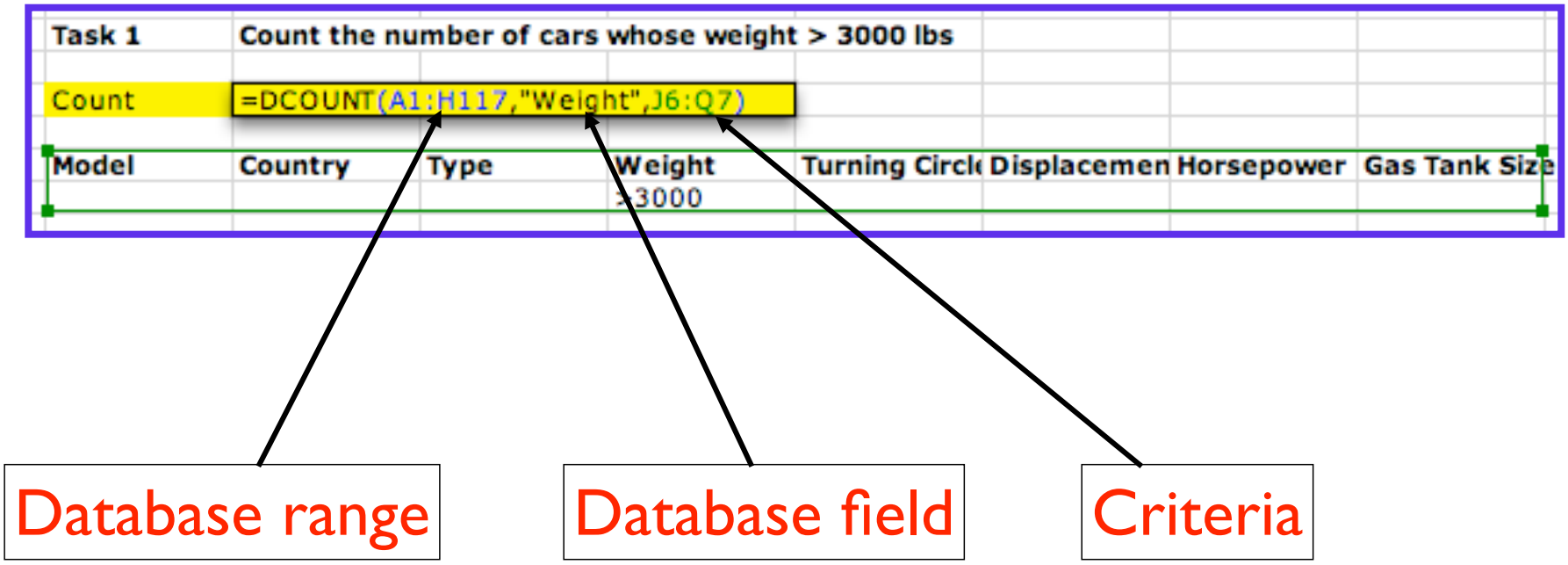
Database Example (Cardata.xls)

- Create a separate section in the worksheet where the query will be done
- Copy the sequence of titles to help you guide the query (see below)

J	K	L	M	N	O	P	Q
Task 1	Count the number of cars		whose weight > 3000 lbs				
Count	50						
Model	Country	Type	Weight	Turning Circl	Displacemen	Horsepower	Gas Tank Size
			>3000				

Database Example (Cardata.xls)

- Suppose we want to count the number of cars whose weight > 3000 lbs
- Dcount(A1:H117,"Weight",J6:Q7)



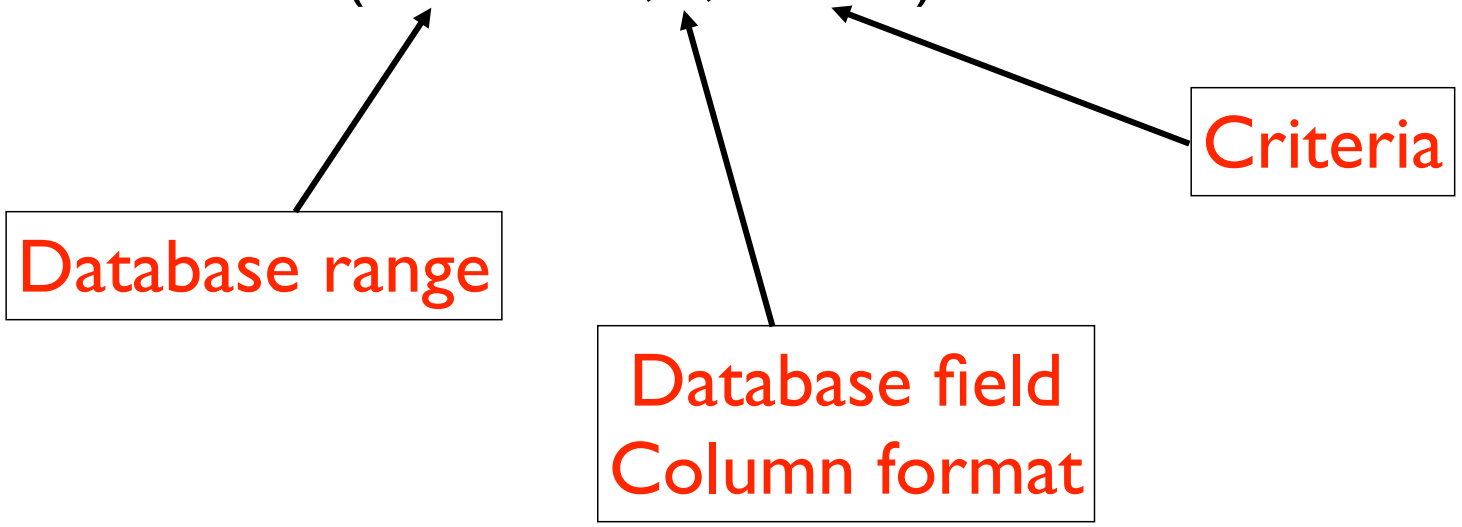
Database Example (Cardata.xls)

- Read the solution from the cell containing the database query
- Note that 50 cars weigh more than 3,000 lbs

J	K	L	M	N	O	P	Q
Task 1	**Count the number of cars**		**whose weight > 3000 lbs**				
Count	50						
Model	**Country**	**Type**	**Weight**	**Turning Circl**	**Displacemen**	**Horsepower**	**Gas Tank Size**
			>3000				

Database Example (Cardata.xls)

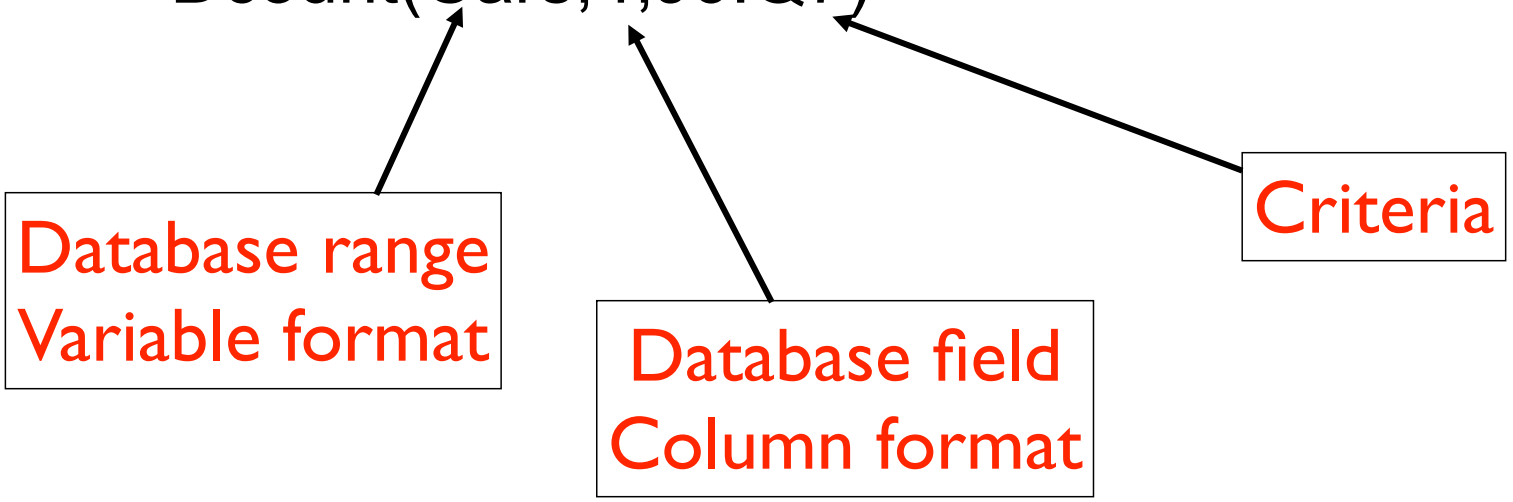
- A variation to specify the database field
- Dcount(A1:H117,4,J6:Q7)



J	K	L	M	N	O	P	Q
Task 1	Count the number of cars		whose weight > 3000 lbs				
Count	50						
Model	Country	Type	Weight	Turning Circl	Displacemen	Horsepower	Gas Tank Size
			>3000				

Database Example (Cardata.xls)

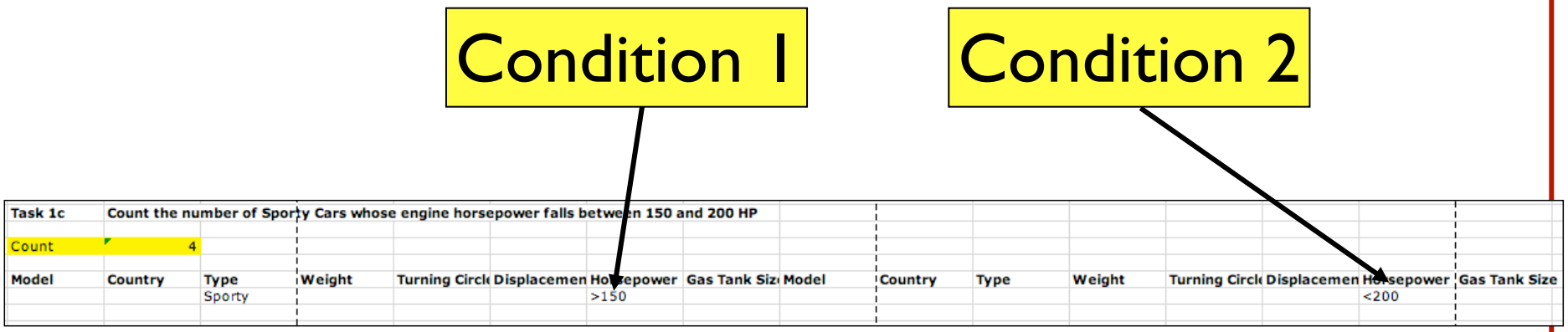
- A variation to specify the database range
- Dcount(Cars,4,J6:Q7)



J	K	L	M	N	O	P	Q
Task 1	Count the number of cars		whose weight > 3000 lbs				
Count	50						
Model	Country	Type	Weight	Turning Circl	Displacemen	Horsepower	Gas Tank Size
			>3000				

Database Example (Cardata.xls)

- Count the number of Sporty cars whose horsepower falls between 150 and 200 HP
- This requires two sets of conditions for one variable (Horsepower)



Database Example (Cardata.xls)
To Do in Class

- Count the number of cars made in Japan whose weight > 2,700 lbs
- Find the average horsepower of cars produced in the U.S.
- Find the average gas tank size of cars produced outside U.S. and Japan

DBMS Components

- Data structure
- Input method
- Editor
- Query ability
- Report generator
- Programming language
- Graphics generator

Data Structure

- Structural organization or model used to represent real world objects
- e.g: real world object = pipe
- Database data structure for “pipe”
  - location, material, diameter, date installed, ...

Database Structure Types

- Flat file (text file)
  - able to store/access 1 table of information
- Relational
  - able to relate data from multiple data tables simultaneously
- Other structure types exist (hierarchical, network,...) but are no longer commonly used

Limitations of flat file design

- Redundancy
  - flat file design usually has unnecessary repetition of data
  - Most people who collect data are not database experts

Date	Station No.	Agency	River Mile	Parameter	Conc
2/12/96	107062	USGS	111.2	Benzene	0.8
2/12/96	107062	USGS	111.2	Chloroform	1.2
2/13/97	LA66	USEPA	137.7	Toluene	2.7
2/13/97	LA66	USEPA	137.7	Benzene	1.1

Limitations, continued

- Multiple value problems
  - e.g. multiple pollutants measured from single sample
  - typical flat file fix is to assume all the pollutants that will be monitored and make column for each
  - Leads to wasted empty space and difficulty in adding/deleting pollutants.

Example commercial available microcomputer relational DBMS

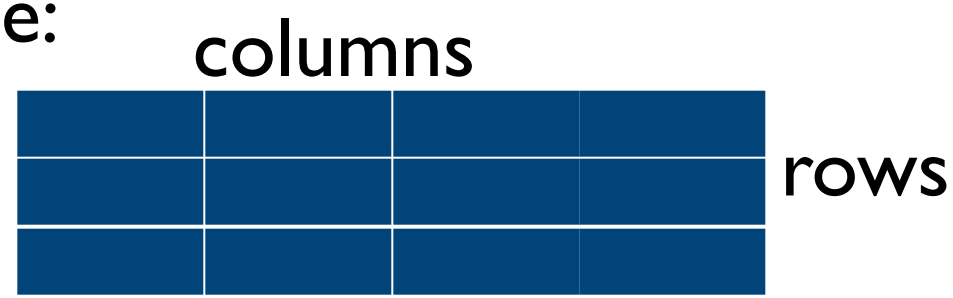
- Microsoft Access
- Lotus
- Borland Paradox
- Borland dBase
- Oracle
- FileMaker Pro

Relational Structure

- Data is organized into independent 2-dimensional arrays
- No formal linkages required between arrays
- Most easily modified structure
- Theoretically least complex and most intuitive structure for user

Basic Elements of Relational Database

Table:



Database: collection of tables

Alternative Relational Database Terms

table = relation = array

row = record = tuple

column = field = attribute

Table Concepts

- Table may have a name intended to convey the meaning of the table as a whole.
- Size of table refers to the number of rows
  - Degree of table refers to the number of columns
- The order of the rows or columns is not important (as long as keep headings with columns). i.e. a sorted table is considered the same table.
- Often denoted by table name with attributes BOOK(ISBN,Title,Price)

Relationship Between Multiple Tables in the Relational DBMS

Date	Parameter	Conc
2/1/96	Pb	2.4
2/2/96	tol	1.7
2/3/96	tol	0.3
2/14/96	Cu	0.5

Name	Parameter	Conc
copper	Cu	mg/L
toluene	tol	ug/L
lead	Pb	mg/L

Design Principles for Relational Database

- Required key fields
- Normalization
  - first normal form
  - second normal form
  - third normal form

Key Field

- A key is a field or combination of fields that uniquely identify each record
- Duplicate entries in key fields are not allowed within a table
- It is good practice, and often required, for each table to have key field(s).
- examples
  - single key field: unique ID sample number
  - multiple key field: date + time + location + pollutant

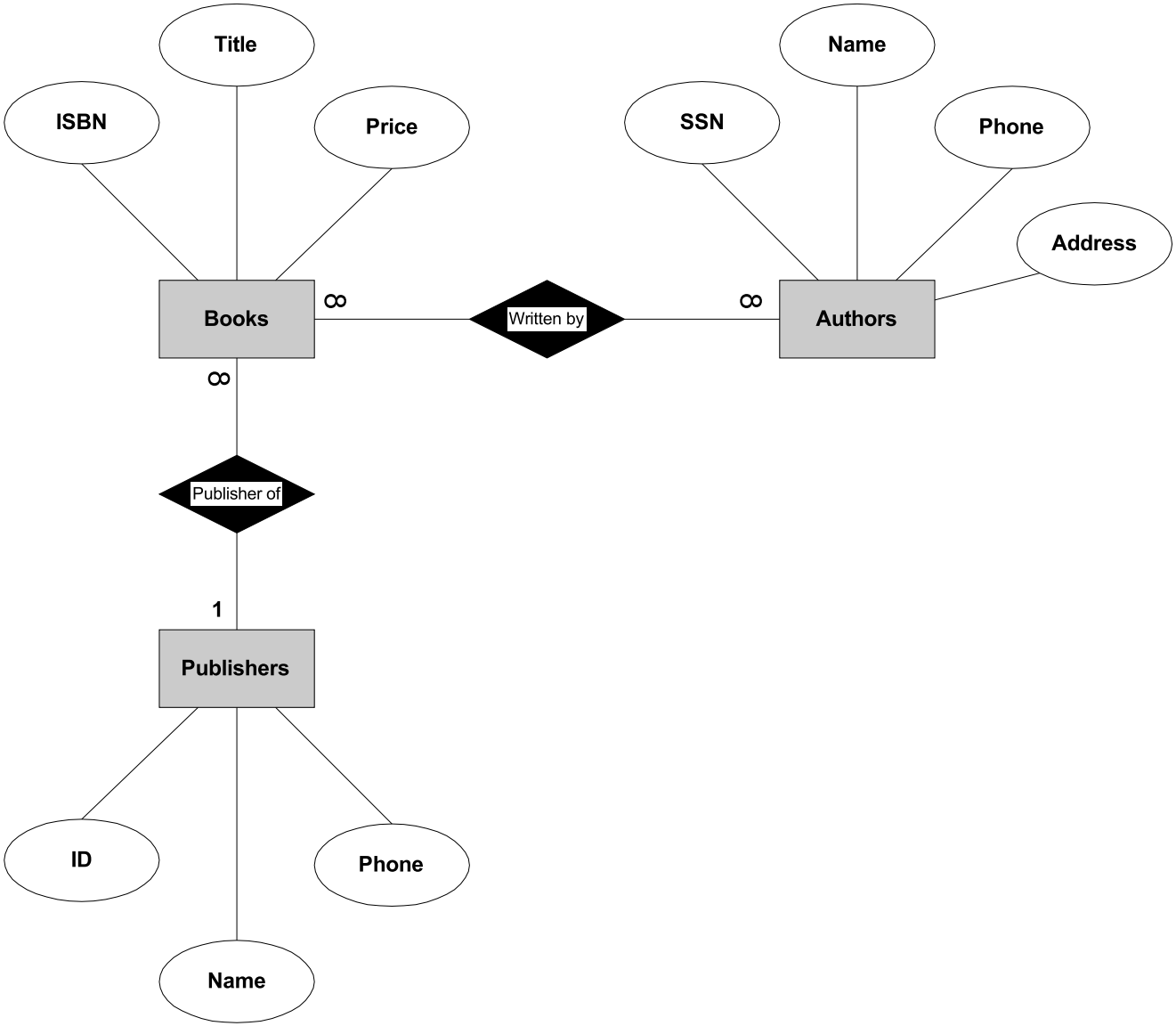
Keys

- Set of attributes that uniquely defines any entity from among all possible entities in the entity class that may appear in the database is called a superkey. Ex: ISBN
- Superkeys can contain more attributes than absolutely necessary, e.g. SSN and LastName for USCitizen class. Key is minimal superkey, e.g SSN

Entity-Relationship Diagram

- Used to display the entity classes in a database model with their attributes and relationships
- entity class - rectangles
- attributes - ellipses
- relations - diamonds

Example Entity-Relationship Diagram



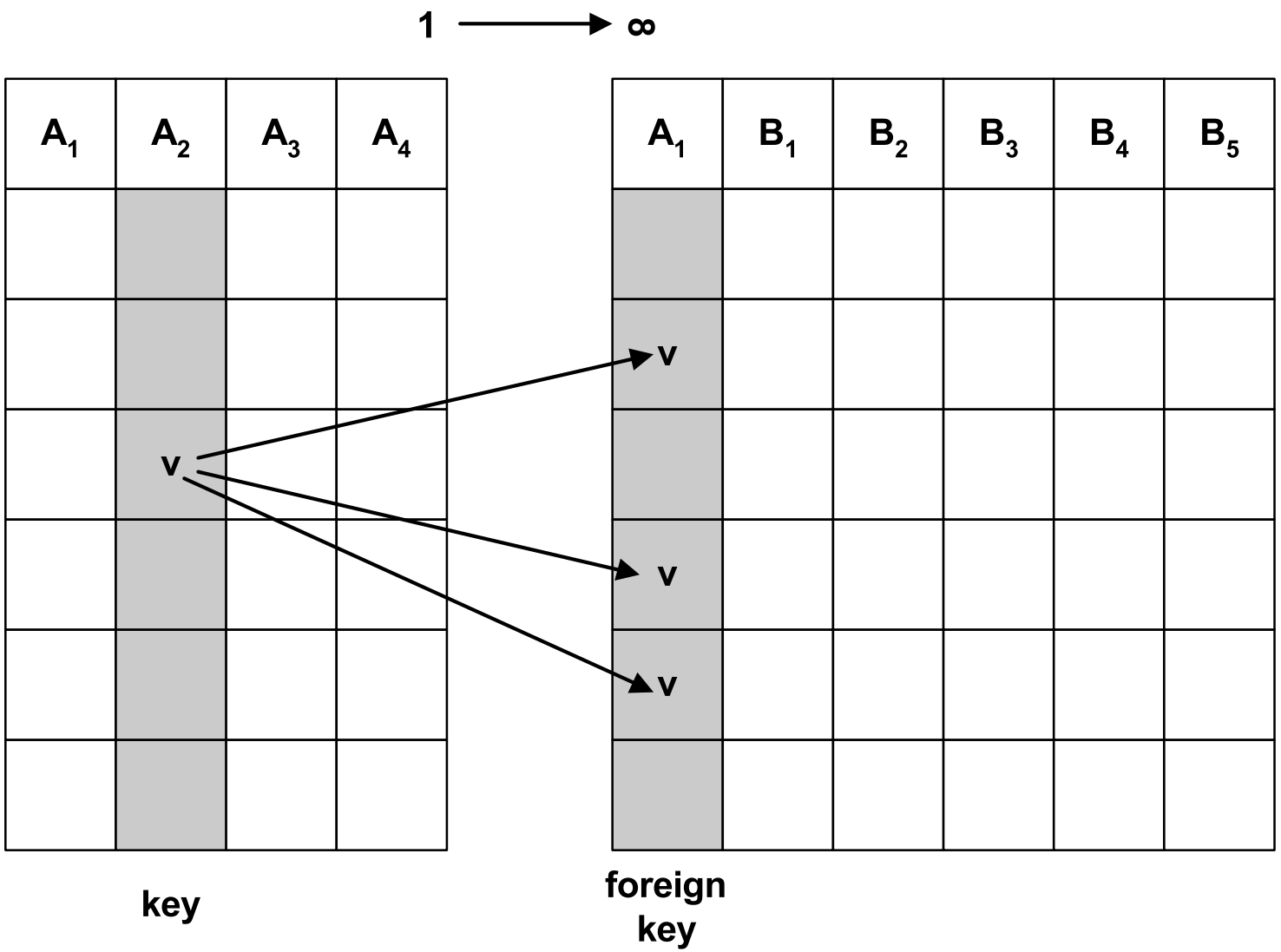
Relationship Types

- One-to-one
  - each entity in one table related to at most one entry in related table. Fairly rare.
- One-to-many
  - each entity in one table related to many entities in related table
- Many-to-many

Implementing One-to-Many Relationships

- Add key attribute from related table. Thus, to implement ***Publisher Of*** relationship
  - BOOK(ISBN, Title, Price) becomes
  - BOOK(ISBN, Title, Price, PublisherID)
  - Publisher ID in Book entity class is referred to as ***foreign key***, since is key for foreign entity class

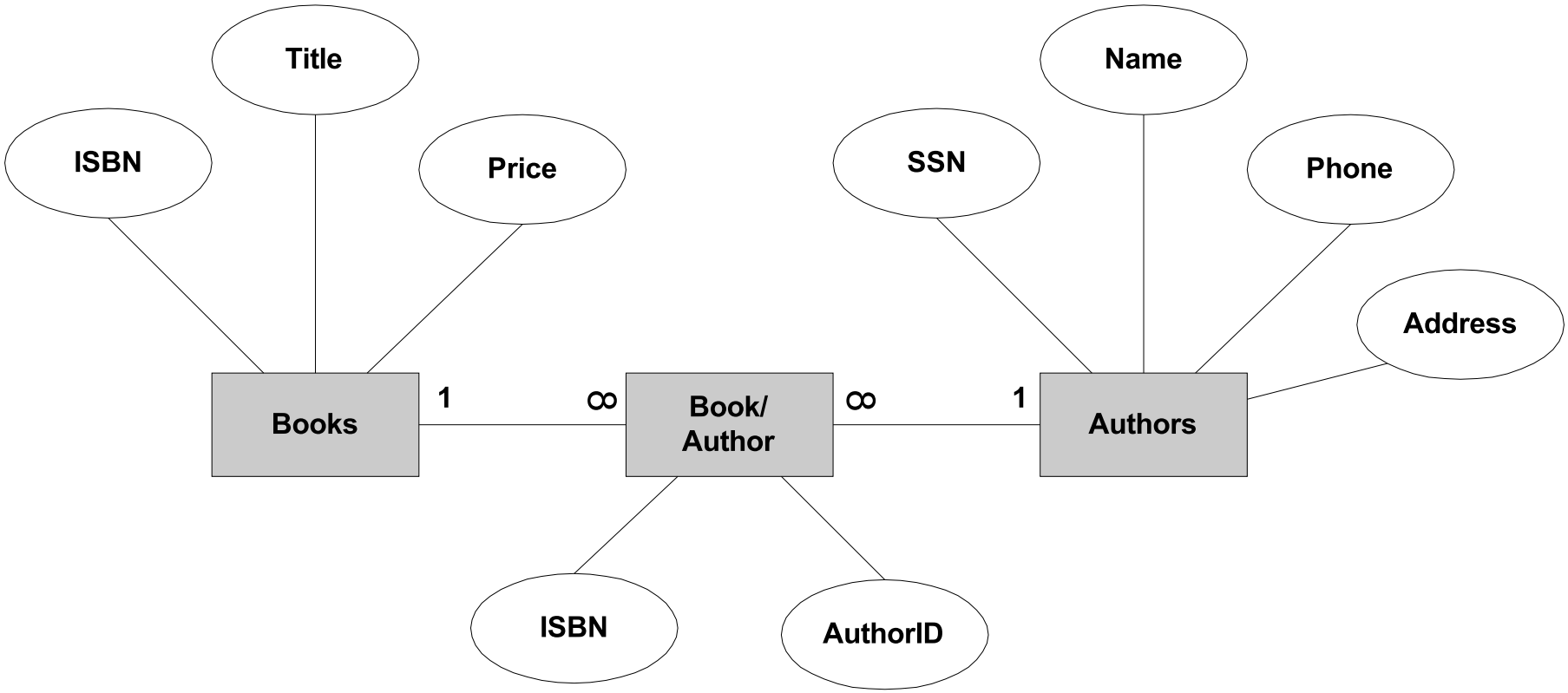
One to Many Relationship



Implementing Many-to-Many Relationships

- More involved than one-to-many
- Cannot simply treat as 2 one-to-many, would result in redundancy
- Need to add new table then treat as 2 one-to-many relationships
  - BOOK/AUTHOR(ISDN,SSN)
- Usually not shown on initial ERD.

Example many-to-many relationship



Referential Integrity

- When using foreign keys to implement relationships, each value of foreign key must have matching value in related table, otherwise ***dangling reference***
- This restriction is termed referential constraint
- Ensuring referential restraint is termed ensuring referential integrity

Referential Integrity, continued

- Referential integrity can be lost by adding new records with foreign key that does not exist in related table.
- Referential integrity can also be lost if value of key is changed or deleted, e.g. delete a publisher in Publisher database. Many records in Book database now have dangling references.

Cascading Updates and Deletions

- Options in many database programs
- Cascading update: If value in referenced key is changed, then all matching entries in the foreign key are automatically changed.
- Cascading delete: If value in referenced key is deleted, then all matching records with same value in foreign key are deleted
- Maintains referential integrity, but should be used with extreme caution.

First Normal Form

No more than one value may be contained in each field

The following table is not in first normal form because there are multiple values in the Parameter and Conc fields

Date	Station No.	Agency	River Mile	Parameter	Conc
2/12/96	107062	USGS	111.2	Benzene Chloro-form	0.8 1.2
2/13/97	LA66	USEPA	137.7	Toluene Benzene	2.7 1.1

Conversion to First Normal Form

- Split records containing multiple entries in a field into multiple records instead.
- The follow table shows the results of converting previous table to 1st normal form.

Date	Station No.	Agency	River Mile	Parameter	Conc
2/12/96	107062	USGS	111.2	Benzene	0.8
2/12/96	107062	USGS	111.2	Chloroform	1.2
2/13/97	LA66	USEPA	137.7	Toluene	2.7
2/13/97	LA66	USEPA	137.7	Benzene	1.1

Second Normal Form

- All non-key fields must be a fact about the entire key
- Following table is not in 2nd normal form because 2 of the fields (agency, river mile) relate only to station number, not to the rest of the key (date, parameter)

Date	Station No.	Agency	River Mile	Parameter	Conc
2/12/96	107062	USGS	111.2	Benzene	0.8
2/12/96	107062	USGS	111.2	Chloroform	1.2
2/13/97	LA66	USEPA	137.7	Toluene	2.7
2/13/97	LA66	USEPA	137.7	Benzene	1.1
key	**key**	**non key**	**non key**	**key**	**non key**

Conversion to 2nd normal form

Table 1

Date	Station No.	Parameter	Conc
2/12/96	107062	Benzene	0.8
2/12/96	107062	Chloroform	1.2
2/13/97	LA66	Toluene	2.7
2/13/97	LA66	Benzene	1.1
key	**key**	**key**	**non key**

Table 2

Station No.	Agency	River Mile
107062	USGS	111.2
LA66	USEPA	137.7
key	**no key**	**no key**

- To convert table to 2nd normal form, non-key fields should be moved to a new table

Third Normal Form

- A non-key field may not contain a fact about another non key field

Sample ID	Station No.	Date Station Establ.	Sample Date	Parameter	Conc
101	107062	1964	1/14/96	Benzene	0.8
102	107062	1964	1/15/96	Chloroform	1.2
103	108935	1979	1/17/96	Toluene	2.7
key	**non key**	**non key**	**non key**	**non key**	**non key**

Conversion to 3rd Normal Form

- Non-key fields that refer to other non-key fields should be moved to new table

Table 1

Sample ID	Station No.	Sample Date	Parameter	Conc
101	107062	1/14/96	Benzene	0.8
102	107062	1/15/96	Chloroform	1.2
103	108935	1/17/96	Toluene	2.7
key	**non key**	**non key**	**non key**	**non key**

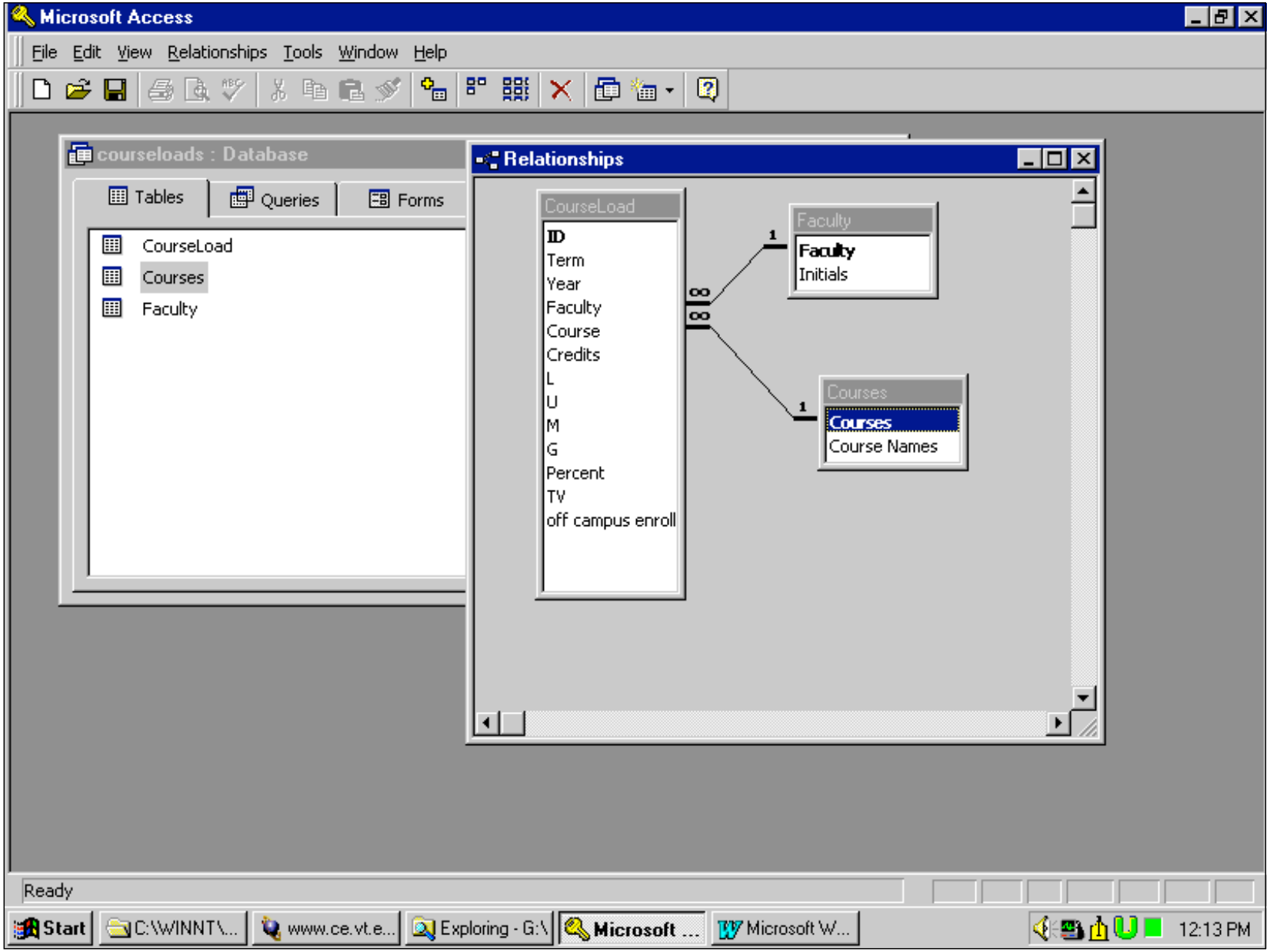
Table 2

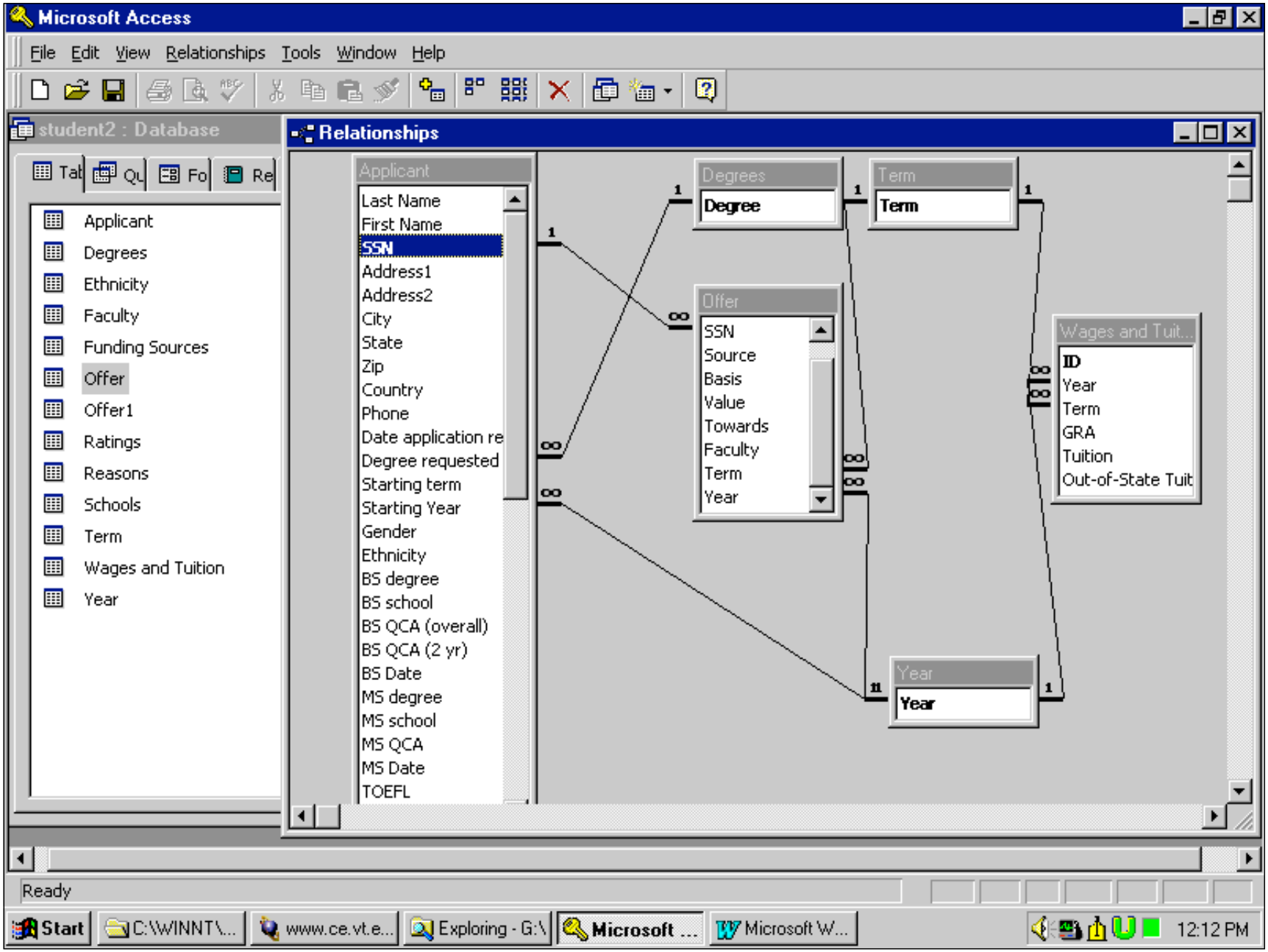
Station No.	Date Station Establ.
107062	1964
108935	1979
key	**non key**

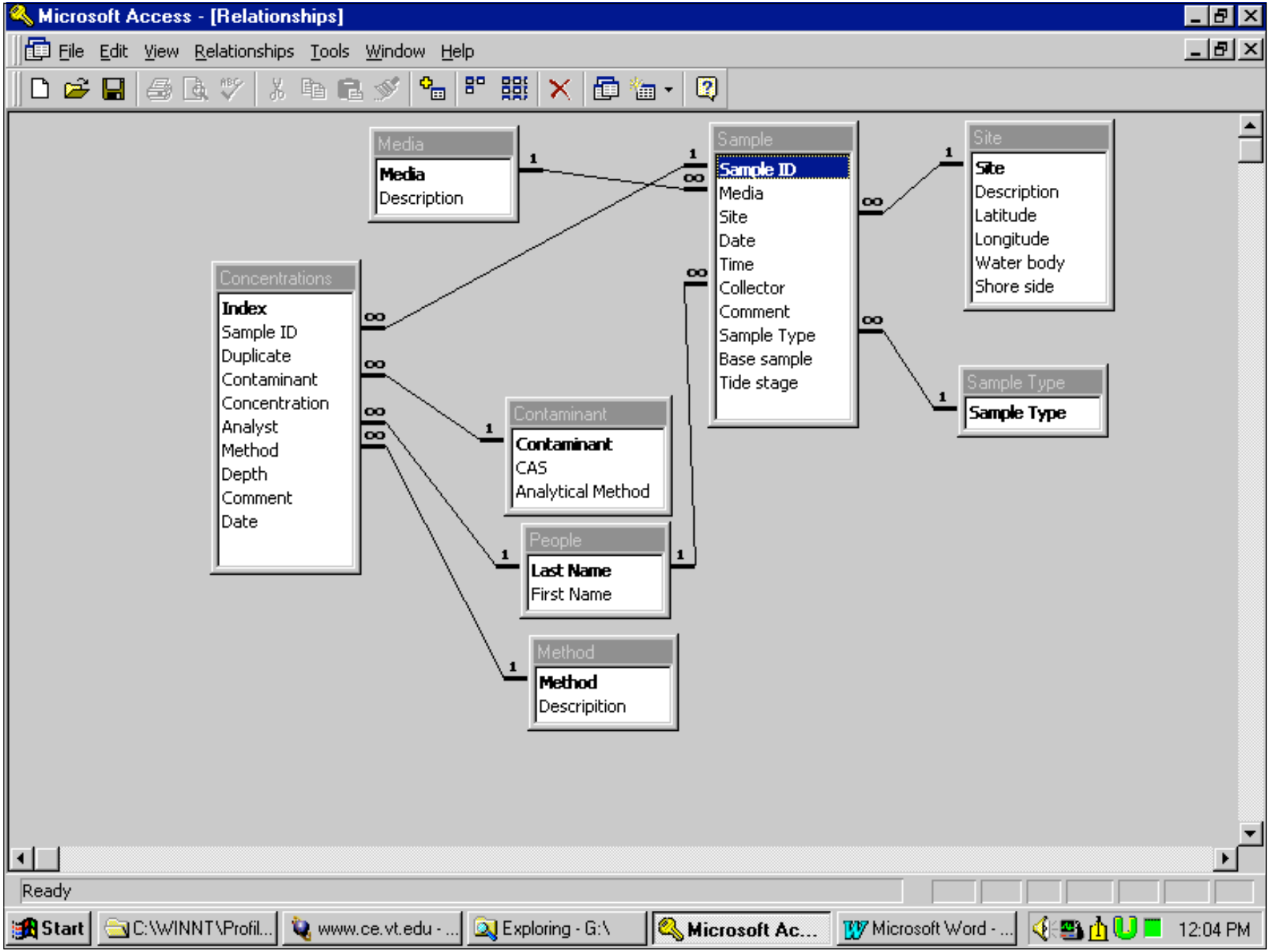
Example Relationships

- Faculty Course Loads
- Graduate Student Applications
- Pesticide Sampling Database

Example Relationships (MS Access)




Microsoft Access
File Edit View Relationships Tools Window Help
student2 : Database
Tab Qu Fo Re
Applicant
Degrees
Ethnicity
Faculty
Funding Sources
Offer
Offer1
Ratings
Reasons
Schools
Term
Wages and Tuition
Year
Relationships
Applicant
Last Name
First Name
SSN
Address1
Address2
City
State
Zip
Country
Phone
Date application re
Degree requested
Starting term
Starting Year
Gender
Ethnicity
BS degree
BS school
BS QCA (overall)
BS QCA (2 yr)
BS Date
MS degree
MS school
MS QCA
MS Date
TOEFL
Degrees
Degree
Term
Term
Offer
SSN
Source
Basis
Value
Towards
Faculty
Term
Year
Wages and Tuit...
ID
Year
Term
GRA
Tuition
Out-of-State Tuit
Year
Year
Ready
Start
C:\WINNT\...
www.ce.vt.e...
Exploring - G:\
Microsoft ...
Microsoft W...
12:12 PM


Microsoft Access - [Relationships]
File Edit View Relationships Tools Window Help
Media
Media
Description
Sample
Sample ID
Media
Site
Date
Time
Collector
Comment
Sample Type
Base sample
Tide stage
Site
Site
Description
Latitude
Longitude
Water body
Shore side
Concentrations
Index
Sample ID
Duplicate
Contaminant
Concentration
Analyst
Method
Depth
Comment
Date
Contaminant
Contaminant
CAS
Analytical Method
Sample Type
Sample Type
People
Last Name
First Name
Method
Method
Descripition
1
∞
Ready
Start
C:\WINNT\Profil...
www.ce.vt.edu - ...
Exploring - G:\
Microsoft Ac...
Microsoft Word - ...
12:04 PM