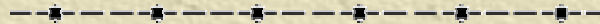


Data quality AND data structures



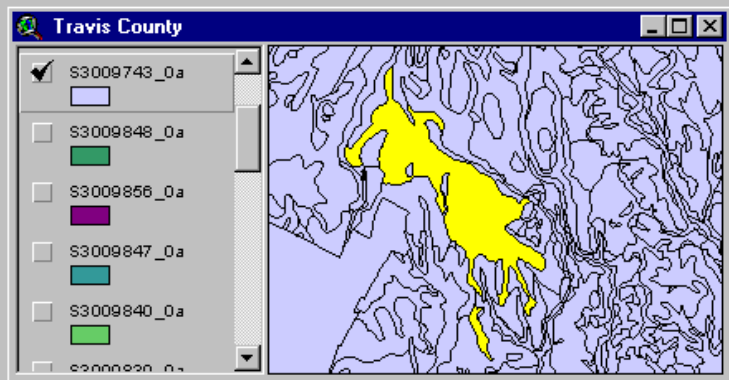
What about attribute data?

ArcView GIS 3.2

File Edit View Theme Graphics Window Help

Scale 1: 112,706

621,009.10
3,351,013.69



Attributes of S3009743_0a

Area	Perimeter	S3009743_0a	S3009743_0a-c	Mtugym	Sstsaic	Ssaic	Mtuid	Mtuc
4056750.09545	26566.16069	295	295	LeB	TX453	453	453LeB	LEWISVILL
250760.53580	3163.50490	296	296	UdD	TX453	453	453UdD	URBAN LAI
35729.62075	1071.81921	297	297	HsD	TX453	453	453HsD	HOUSTON
126391.15865	2425.79514	298	298	EuC	TX453	453	453EuC	EDDY SOIL
74018.01095	1353.27855	299	299	UdD	TX453	453	453UdD	URBAN LAI
180775.41385	3024.18519	300	300	UvE	TX453	453	453UvE	URBAN LAI

interp

Sstsaic	Mtuid	Segnum	Zipcode	Rating	Restof	Res
TX453	453LeB	1	23	3	37	
TX453	453LeB	1	5	4	20	3
TX453	453LeB	1	19	11	7	
TX453	453LeB	1	8	5	31	
TX453	453LeB	1	7	5	31	
TX453	453LeB	1	17	3	20	2
TX453	453LeB	1	18	5	23	
TX453	453LeB	1	22	11	11	
TX453	453LeB	1	14	9	41	
TX453	453LeB	1	20	11	33	
TX453	453LeB	1	11	5	37	
TX453	453LeB	1	10	5	22	3
TX453	453LeB	1	26	3	37	
TX453	453LeB	1	24	3	37	
TX453	453LeB	1	25	3	37	
TX453	453LeB	1	16	3	25	
TX453	453LeB	1	12	4	22	3
TX453	453LeB	1	13	9	41	
TX453	453LeB	1	4	6		
TX453	453LeB	1	3	5	37	
TX453	453LeB	1	1	3	26	
TX453	453LeB	1	2	3	25	
TX453	453LeB	1	6	3	37	
TX453	453LeB	1	9	5	31	
TX453	453LeB	1	21	11	11	
TX453	453LeB	1	15	4	37	
TX453	453LeB	2	23	11	54	
TX453	453LeB	2	5	11	54	
TX453	453LeB	2	19	11	54	
TX453	453LeB	2	8	11	54	
TX453	453LeB	2	7	11	54	
TX453	453LeB	2	17	11	54	
TX453	453LeB	2	18	11	54	
TX453	453LeB	2	22	11	54	
TX453	453LeB	2	14	11	54	

comp


Sstsaic	Mtuid	Segnum	Mtugym	Compname	Ssid	Compcc	Slope	Slopeh	Surftex	Otherph
TX453	453LeB	1	LeB	LEWISVILLE	TX0095	68	0	2	SIC	
TX453	453LeB	2	LeB	URBAN LAND	TX8002	25	0	2	VAR	
TX453	453Ln	1	Ln	LINCOLN (GADDY)	OK0034	100	0	1	LFS	
TX453	453Lu	1	Lu	LINCOLN (GADDY)	OK0034	85	0	1	LFS	
TX453	453Lu	2	Lu	URBAN LAND	TX8002	10	0	1	VAR	
TX453	453Mw	1	Mw	MISPELL AMPLIUS WATER	TX8037	100				

Capturing attribute data

- ✦ Attribute data is non-locational data that describes the spatial entities in your GIS.

Metadata


✠ Metadata are literally data about the data.

- 
- ✦ In the US, a standard for metadata has emerged through the National Geospatial Data Clearinghouse (NGDC).
 - ✦ The FGDC is also working with GeoConnections in Canada to develop international standards.
 - ✦ Search interfaces. The real power of structured metadata is in the ability to search for information in order to obtain appropriate data.
 - ✦ GIS scholars need to study ways to build search interfaces that make it easy for people to ask complex questions and get understandable answers.


Data Quality


✦ “Garbage in , garbage out.”

✦ The poorer the data quality, the poorer the decisions resulting from the analysis.



✦ Geographical entities are frequently not stable through time and space.

- 
- ✦ *Temporal* problems with data are related to change in the data itself.
 - ✦ We also experience error from *secondary* data. When we classify RS data, it is expected that a certain percentage of the ground cover will be misclassified.



✦ *Spatial* error is another dimension of data error.

✦ Data entry is often poorly controlled, and conducted far away from the location being represented.

Data quality in SDTS

- ✦ The FGDC was established to promote and ease dissemination of spatial data.
- ✦ It is specifically concerned with development of metadata standards.

- ✦ OpenGis works with the US SDTS and FGDC.
- ✦ In Canada, data standardization is overseen by GeoConnections which is responsible for creating a “national geospatial data infrastructure” (CGDI).

GeoConnections ...Canada's Geographic Information on the Internet

[Home](#)

[About GeoConnections](#)

[Help](#)

[Search](#)

[Contact Us](#)

[Français](#)



Canadian
Geospatial
Data
Infrastructure

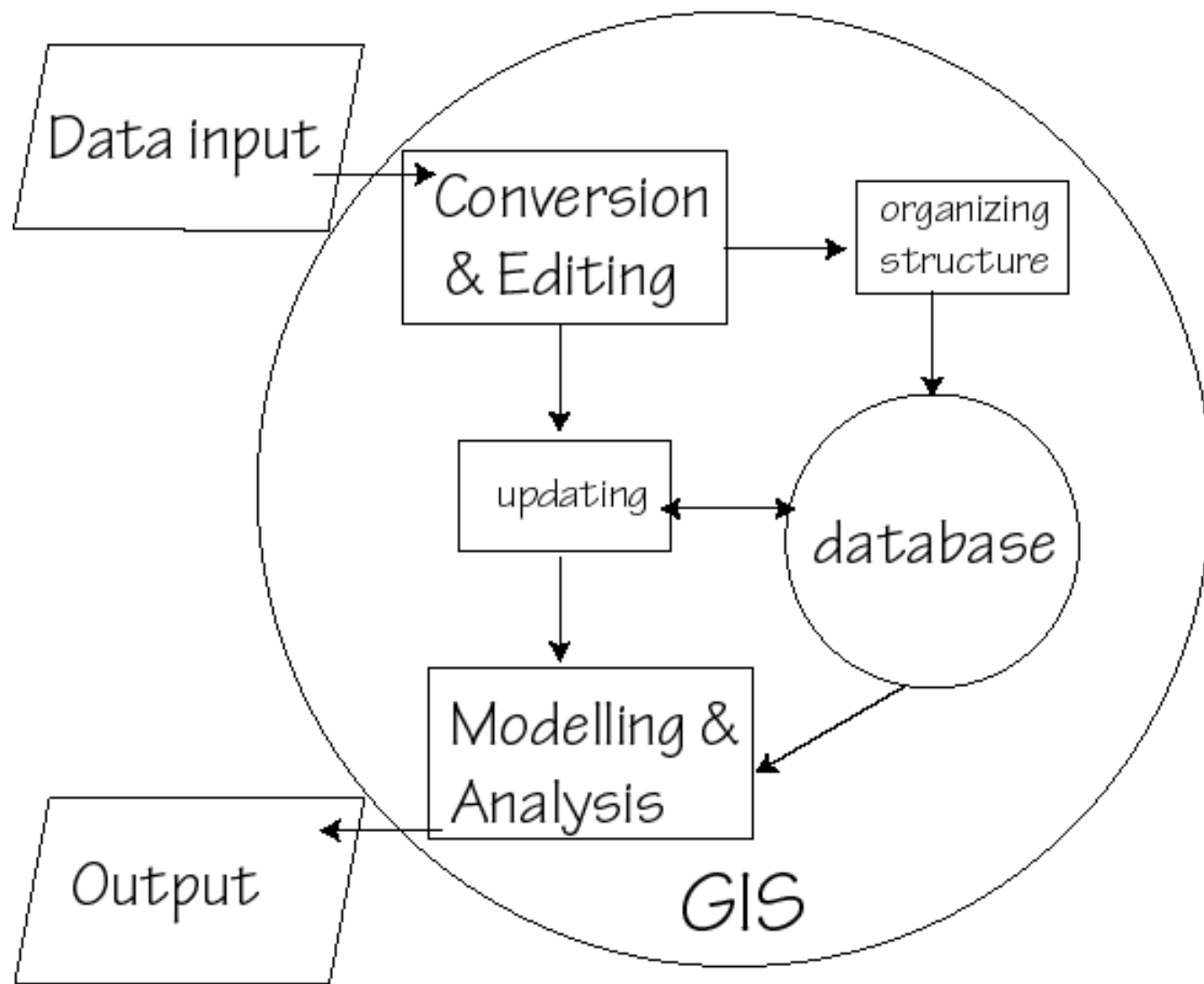
Components of Data Quality according to SDTS

Lineage	-refers to source materials, methods of derivation and transformations of data. Also date and whether some data were derived from different sources
Positional Accuracy	-accuracy of spatial fields, both horizontal and vertical as well as variations in accuracy in subsequent data or overlays.
Attribute Accuracy	-accuracy of thematic attributes. Collection method.
Logical Consistency	-refers to fidelity of relationships including topological inconsistencies, valid dates for attributes.
Completeness	-refers to the relationship between the attributes and all objects in their class, not included in the database. Includes selection criteria, definitions and mapping rules.

Organizing information in the computer: File and Data Access


- ✦ All of the information input into a GIS (as numbers) has to be stored in an orderly fashion in the computer. This is done by organizing numbers in lists, stacks, registers and arrays. These are referred to as file structures.
- ✦ File structures are used by data structures to organize data.

Data structures as grand central in GIS

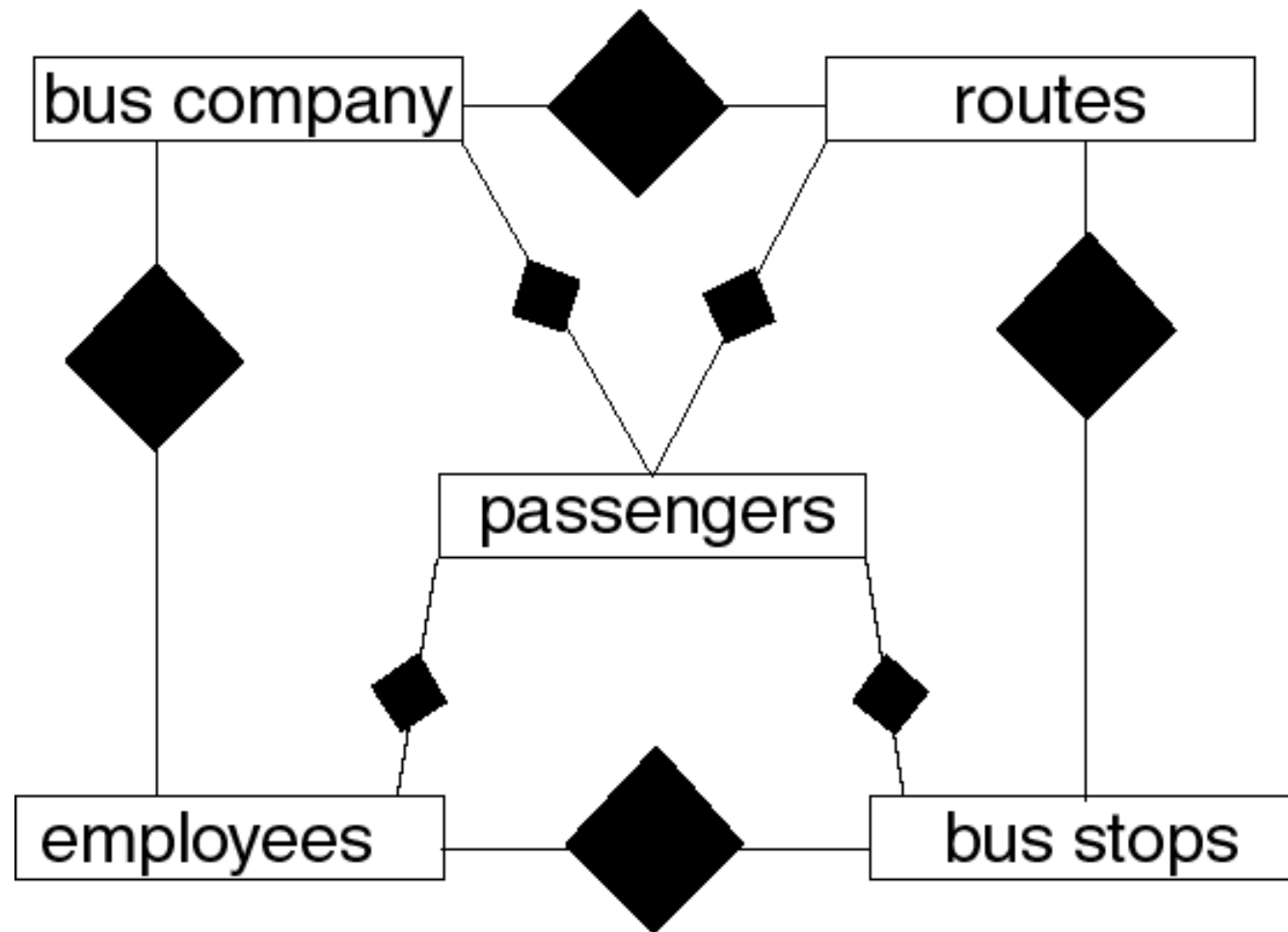


What data structures enable

- ✦ *conversion* --- transforming data from one format to another, from one unit of measurement to another, and/or from one feature classification to another.
- ✦ *organization* --- organizing or re-organizing data according to database management rules and procedures so that they can be accessed cost-effectively
- ✦ *structuring* --- formatting or re-formatting data so that they can be acceptable to a particular software application or information system
- ✦ *modelling* --- including statistical analysis and visualization of data that will improve user's knowledge base and intelligence in decision making

- 
- ✦ the concepts of "organization" and "structure" are crucial to the functioning of information systems
 - ✦ In the simplest sense, a data structure is a computerized record-keeping system.
 - ✦ Pared down to basics, we have *entities* and their *relationships* in a database. This is called the E/R model.

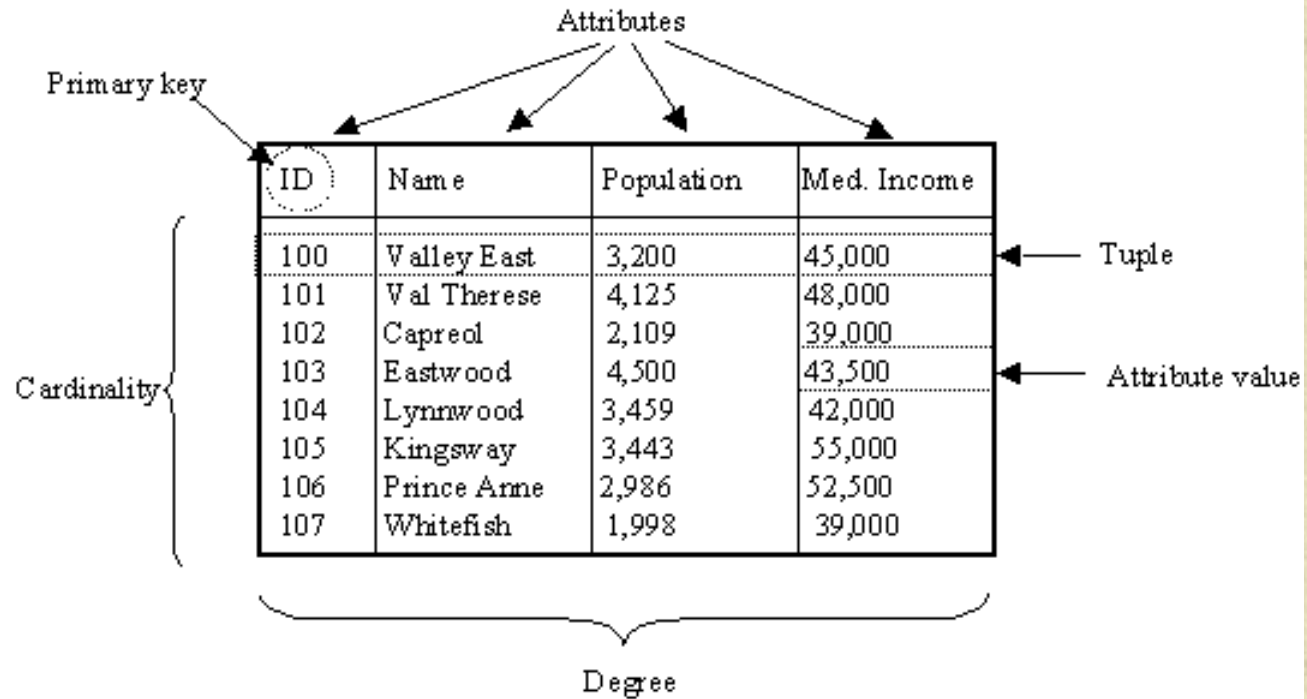
Entity/relationship (ER) diagram



Relational database structures

- ✦ The RDBMS underlies many GIS.
- ✦ The E/R model is applicable because databases don't just store information. They also store the relationships between bits of information.
- ✦ These relationships are what distinguish the RDBMS.
- ✦ The structure of the RDBS is based on collections of tabular relations called *tuples*.

Figure 13: Characteristics of a relational table



A.K. Yeung 1998-10-10 u51-13

In this example, there is a relationship between each city and its attributes (population and median income). The attribute titles are name, population and income. Each row or tuple has a list of values and each tuple is really an instance of the entity in the database.

Relations have the following properties:

1. The ordering of the tuples is not significant. (ex. If we sort by descending instead of ascending, the values won't change)
 2. Tuples in a relation are all distinct.
 3. Usually RBDS require that the data entities are *atomic*. Put more simply, this means one entry per field.
- ✠ The *degree* of the table is the number of its columns (fields) and the *cardinality* is the number of its tuple (entries).

Example of a student enrollment database:


✦ The following scheme might be used:

STUDENTS (STU_ID, SNAME, CLASS, TELEPHONE)

COURSES (CO_NO, CNAME, CROOM, TIME)

ENROLLMENT (CO_NO, STU_ID)

✦ Each of the above constitutes a table.


- 
- ✦ Note that the first field, in each table definition, is underlined. That is the *primary key*.
 - ✦ A key is an attribute that uniquely defines each tuple. Often, only one key is permitted. In that case, it is called the primary key.

STU_ID	SNAME	CLASS	TELEPHONE
1	Jones	2	555-1234
2	Smith	3	555-4321
3	Black	2	555-3344

Students

CO_NO	CNAME	CROOM	TIME	TERM
101	Geog1	5018	2-130	F
102	Geog2	5005	3-230	F
105	CS1	9000	5-930	S
108	Math2	7002	4-1030	S

Courses




✦ The **join** operation makes relational databases *relational*. The binary join takes two relations as input and returns a single relation. Syntax: Join ($rel_1, rel_2: att_1, att_2$) means that rel_1 and rel_2 are joined on attribute combinations of att_1 and rel_1 and att_2 and rel_2 .

✦ Using our Student Enrollment Example, we might **join** (ENROLLMENT, COURSES: CO_NO, CO_NO). This will theoretically join the database enrollment and courses.

✦ **Why will this join create a nonsense table?**

SQL Commands

- ✦ Relational database structures use Structured or Standardized Query Language (SQL) to allow users to setup their database scheme and then query the system.
- ✦ SQL is like a mini-programming language. So you can query the database and get returns



```
SELECT select-item-list
FROM table-reference-list
WHERE [condition]
```

```
SELECT SNAME
FROM STUDENTS
WHERE CLASS < 2
```

Will return the names of all students, from the students table, who are in year one.

```
SELECT STU_ID
FROM ENROLLMENT
WHERE COUNT (CO_NO)>2
```

This query will return the id #s of all students who have registered for more than one course.

SQL QUERIES (CON'T)

✦ Commands like SELECT, AVG, COUNT are included in all relational databases. This has the advantage of reducing the learning curve when moving from one RDBS to another.

✦ SQL also handles IFF statements, IF THEN statements.

```
SELECT STU_ID  
FROM STUDENTS  
IFF CLASS > 2
```

Will return the student ID IFF they are in year 3 or 4.

Useless queries

AVG(TELEPHONE)

or

SELECT CLASS

FROM STUDENTS

WHERE TELEPHONE > 555-1213

✦ This type of mistake is easy to make. Beware also of false joins, links.

SQL and GIS

- ✦ The problem with SQL and spatial queries is that spatial data are complex (think of all the arc and chains that make up a single polygon) while SQL is best for uni-dimensional data.
- ✦ GIS researchers have tried to come up with extensions to SQL which incorporate spatial data but these have not made it into the mainstream of standards.
- ✦ One reason is that standards are controlled by industry and government and it is very difficult to have an effect on so large an infrastructure.

Set theory and its use in GIS

- ✦ Sets are collections of objects that are assumed to have the same characteristics, or belong to the same class.
- ✦ In order to enable many GIS operations, we assume that attributes constitute their own sets.
- ✦ *Set relations are fundamental to modelling spatial information.*

Construction of sets

- ✦ Elements or members of a set: the things to be modelled (this could include an attribute).
- ✦ Sets: collections of elements to be modelled.
Membership: The relationship between the elements and sets they belong to.
- ✦ Equality: relationship that holds between two sets when they both contain exactly the same members.
- ✦ Subset: A relationship between two sets where every member of one set is a member of the second. (Just like it sounds).

The number of elements in a set is its *cardinality*. (remember the cardinality of relational datatables)

Intersection: The operation that takes the two sets and returns the set of elements that are members of both the original sets. The intersection of sets S and T is written as $S \cap T$.

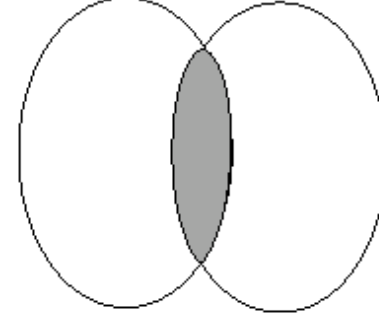
Union: The operation that takes the two sets and returns the set of elements that are members of at least one of the original sets. The union of sets S and T is written as $S \cup T$.

Difference: An operation that takes two sets and returns the set of elements that are members of the first set but not the second set. The difference of sets S and T is written $S \setminus T$.

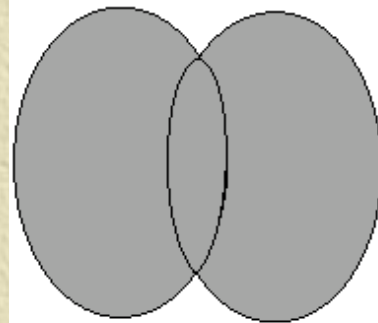
UNION = OR = \cup , INTERSECTION = AND = \cap ,
DIFFERENCE = NOT = \setminus .

✦ *Some sets are so well-used that they have their own (recognizable) name. These include the integers, Reals, the real plane (= ordered pairs in the real plane).*

Set intersection



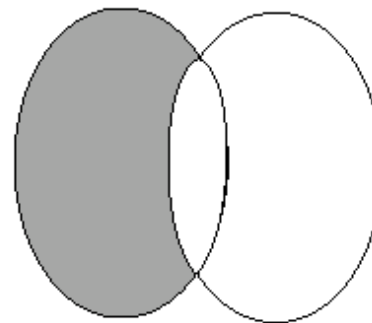
$$A \cap B$$



$$A \cup B$$

Set union

Set difference



$$A \setminus B$$

Overview of spatial analysis using set relations

A query might be structured to find number of hectares of (i) previously unforested land; (ii) with > 50% Douglas fir; and (iii) within 1000 meters of a four lane highway:

```
IF AREA = NPF AND TYPE = > .50 DOUGLAS  
AND HWY_DISTANCE < 1000 M THEN  
AREA = 1 ELSE AREA = 0.
```

✦ The area is assigned a Boolean value of 1 (digital for yes) if it meets the criteria and 0 if it fails.

Example

✠ A = all gas stations

✠ B = gas stations with diesel facilities

FIND:

$X = A \text{ AND } B$

$X = A \text{ OR } B$

$X = A \text{ XOR } B$

$X = A \text{ NOT } B$

Boolean operations:

- ✦ Are not commutative (the order of A and B matters)
- ✦ Require exact matches (no fuzzy definition)
- ✦ Implications of no fuzzy tolerance...