GEOG 413/613

LECTURE 4

Entity-Relationship Model

- The Entity-Relationship model (ER-model) is a diagrammatic representation of the *miniworld* into a set of entities and their relationships.
  - An entity is a unit with a real existence e.g. a company, a school, parcel
  - Attribute is a property of entity e.g. company name, school population, parcel ID
  - Attributes describe entities
  - An instance of an entity instance comprises the values assigned to its attributes

Entity-Relationship Model

- Attributes
  - An attribute can be simple or composite
    - Simple e.g. last name
    - Composite contains other attributes e.g. address contains multiple items
  - An attribute can be
    - single valued (e.g., the date of birth)
    - or multivalued (e.g., a person's telephone numbers)
  - An attribute can be
    - either stored in the database, e.g., a person's date of birth
    - or derived after processing the database content, e.g., the current person's age
- An attribute which identifies an entity instance is called a key attribute (e.g., a person's SIN, a parcel's id)

Key Attributes

- The *key* is an attribute or a group of attributes whose values can be used to uniquely identify an individual entity in an entity set.
  - Candidate key: each attribute or combination of attributes that identifies the row in a relation. For instance, the tuples in the relation of owners can be identified either through the SIN (candidate key 1) or the combination of attributes SURNAME-NAME-DoB (candidate key 2), assuming that there are no two owners in the database sharing the same combination of values in these three attributes.
  - A candidate key is called simple, when it consists of a single attribute (e.g., the candidate key 1) or composite, when it comprises more than one attributes (e.g., the candidate key 2)

Key Attributes

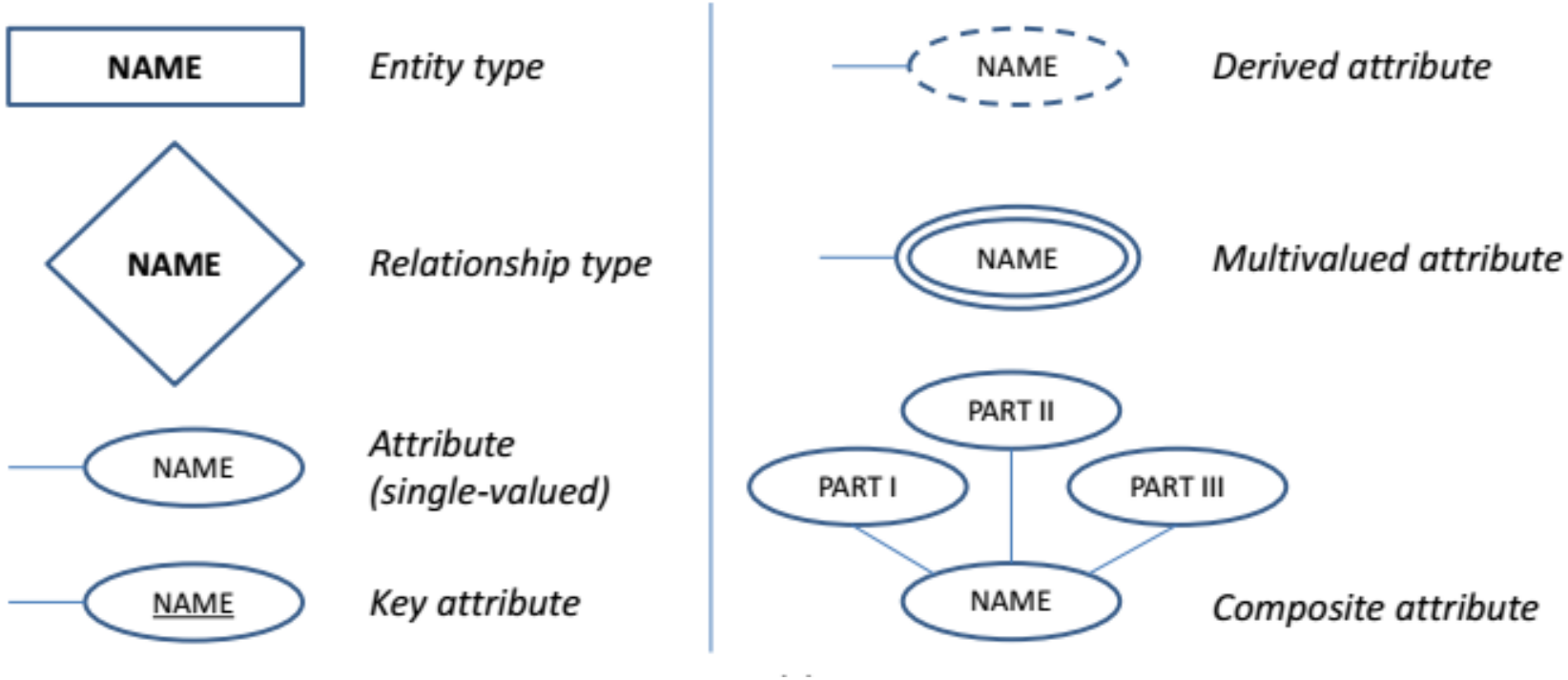
- The *key* is an attribute or a group of attributes whose values can be used to uniquely identify an individual entity in an entity set.
  - Primary key: the candidate key chosen to identify the rows in a relation. For practical reasons, it is the one with the fewest attributes. For example, the SIN (candidate key 1) is an ideal key for a table of owners.
  - The names of the primary key attributes are underlined in the relational schema.

Key Attributes

- The *key* is an attribute or a group of attributes whose values can be used to uniquely identify an individual entity in an entity set.
  - Foreign key: is an attribute in a table that references the primary key in another table
  - Both foreign and primary keys must be of the same data type.

Entity-Relationship Model

- The representation of an entity type in an ER-diagram
  - rectangle denotes the name of the entity type
  - ellipses denotes the attributes assigned to an entity type

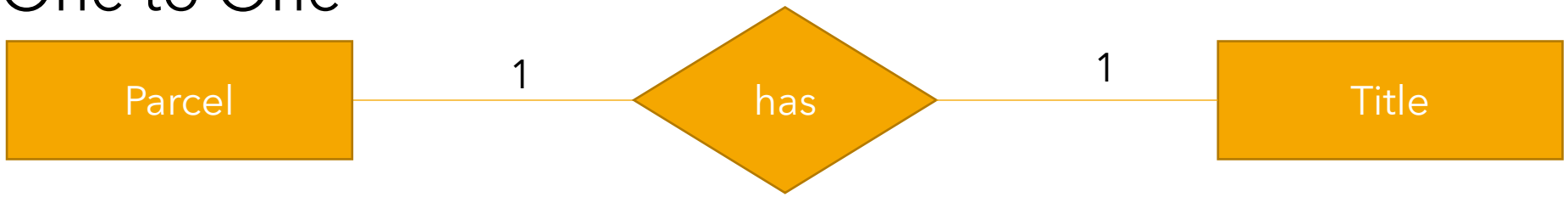


Entity and Relationship Types

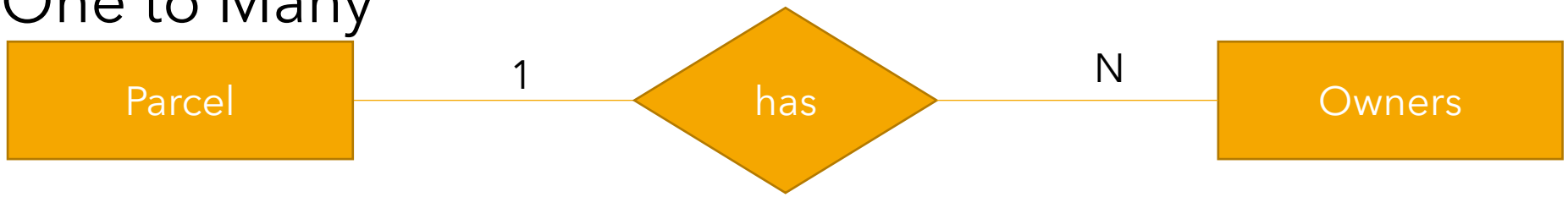
- The entities of one or more types may be related through relationships.
  - A relationship type *R* links entity types E_1, E_2, …, E_n in a structured manner
  - There are three types of relationship that may apply to a database
    - They are one-to-one relationships
    - one-to-many relationships
    - many-to-many relationships
  - These three types of relationship are also referred to the *cardinality*

Entity and Relationship Types

- One to One



- One to Many



- Many to Many

Relationship Types

- **One-to-one relationship**
  - One occurrence of an entity relates to exactly one occurrence of another entity.
  - One row in a specific table that relates to one row in another table
    - E.g. Each student is assigned one student ID

Relationship Types

- **One-to-many relationship**
  - One occurrence in an entity relates to many occurrences in another entity.
  - one row in a specific table that relates to multiple rows in a different table
    - E.g. One customer ID links to Many Order IDs

Relationship Types

- **Many-to-many relationship**
  - Multiple occurrences in one entity relate to multiple occurrences in another entity.
  - Several rows in a specific table that relate to several rows in another table
    - E.g. Book titles and Authors:
      - A book can be linked to more than one author and an author can be linked to more than one book title

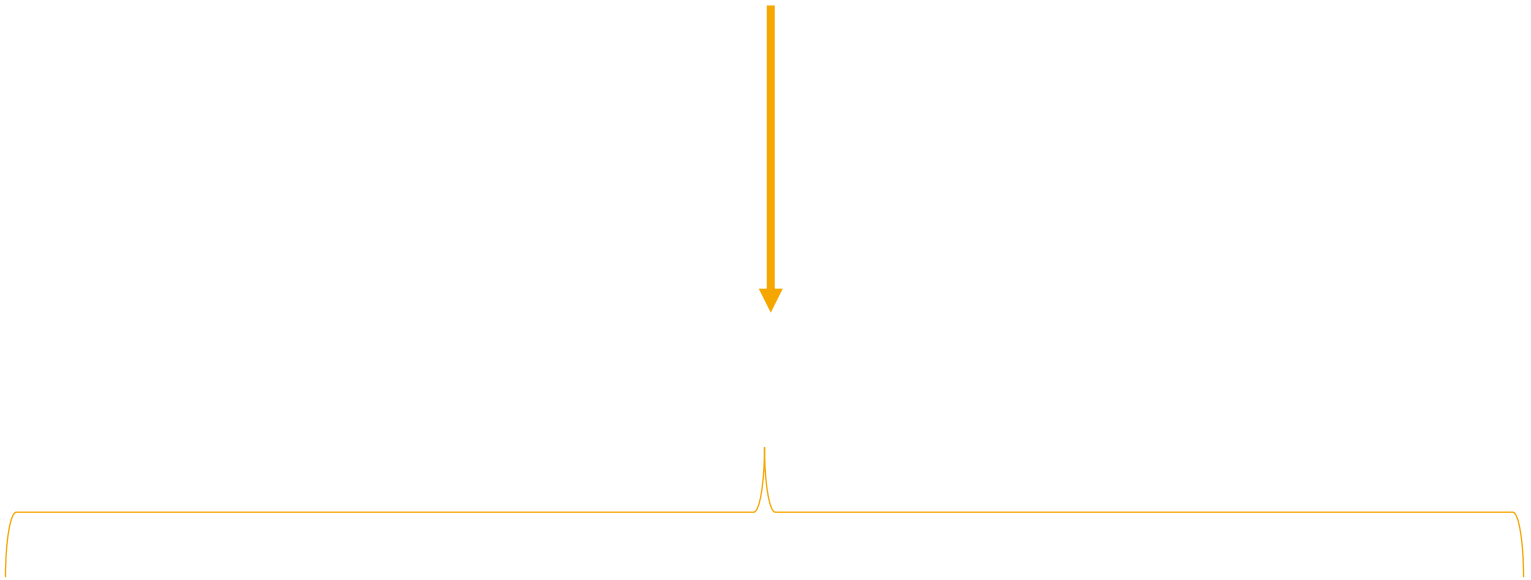
Normalization

- The process minimizing redundancy in a database
  - Characterizes the level of redundancy in a relational schema
  - Provides mechanisms for transforming schemas in order to remove redundancy
  - Data normalization follows certain rules which are categorized as "normal forms". There are 6 normal forms but we'll briefly look at only 3
- In principle, any information that can be applied to more than one record should be moved to its own table.
  - Each successive normal form applied must meet the rules of the previous form

Normalization

- First Normal Form (1NF)
  - Eliminates repeated data entries by giving a single value for each cell
  - To normalize a relation that contains a repeating group, remove the repeating group and form two new relations.
  - It creates unique records for each data set and uses a primary key to identify data sets.
  - These primary keys help to organize data that would otherwise need multiple fields.
    - An example of this process could be a student grade report.
    - The repeating group is the course information because a student can take many courses

Student Grade Report							
StudentNo	StudentName	Major	CourseNo	CourseName	InstructorNo	InstructorName	Grade



Student		
StudentNo	StudentName	Major

Student Course					
StudentNo	Course No	CourseName	InstructorNo	InstructorName	Grade

Normalization

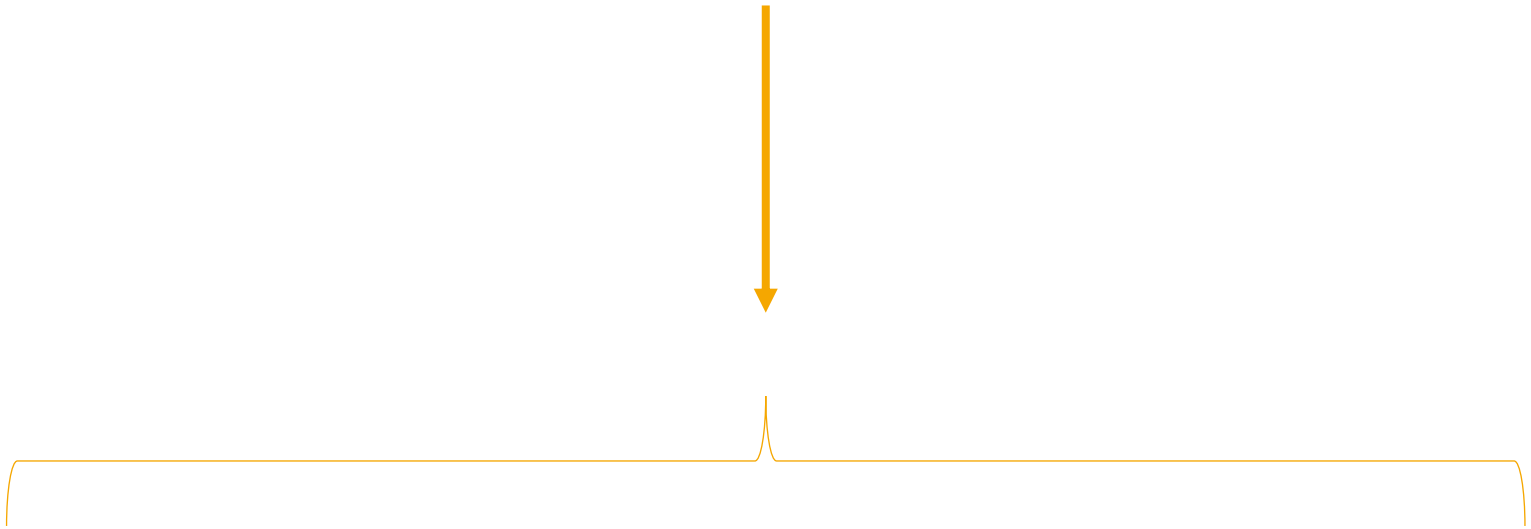
- Second Normal Form (2NF)
  - For 2NF the relation must first be in 1NF
    - Relation is automatically in 2NF if, and only if, the Primary Key comprises a single attribute.
  - Used to break data into multiple rows and separate tables
  - adds a distinct foreign key to a data set that corresponds with a value in the first normal groupings

  - If the relation has a composite PK, then each non-key attribute must be fully dependent on the entire PK

    - Student table is already 2NF
    - For the course inform not all attributes are filly depend on the Primary Key. The grade is fully dependent on the primary key

Student		
StudentNo	StudentName	Major

Student Course					
StudentNo	Course No	CourseName	InstructorNo	InstructorName	Grade



Student		
StudentNo	StudentName	Major

Student Course		
StudentNo	Course No	Grade

Course Instructor			
Course No	CourseName	InstructorNo	InstructorName

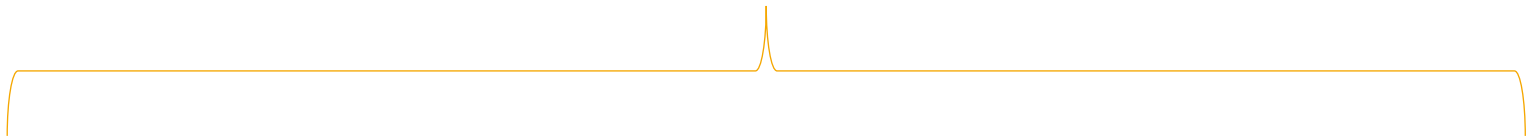
Normalization

- Third Normal Form (3NF)
  - focuses on eliminating any fields not dependent on the key. It is used most effectively for information that changes often.
  - If you change the primary key through this step, you must also move all related data into a different table.

Student		
StudentNo	StudentName	Major

Student Course		
StudentNo	Course No	Grade

Course Instructor			
Course No	CourseName	InstructorNo	InstructorName



Student		
StudentNo	StudentName	Major

Student Course		
StudentNo	Course No	Grade

Course	
Course No	CourseName

Instructor	
InstructorNo	InstructorName

Indexing Geographic Information

- Geographic databases tend to be very large and geographic queries computationally expensive
  - A database uses indexing to find data quickly
  - A database index is, conceptually speaking, an ordered list derived from the data in a table much like a book index
  - Using an index to find data reduces the number of computational tests that have to be performed to locate a given set of records
  - Full table scans are avoided when an index is created and stored in a table
- *A database index is a special representation of information about objects that improves searching*

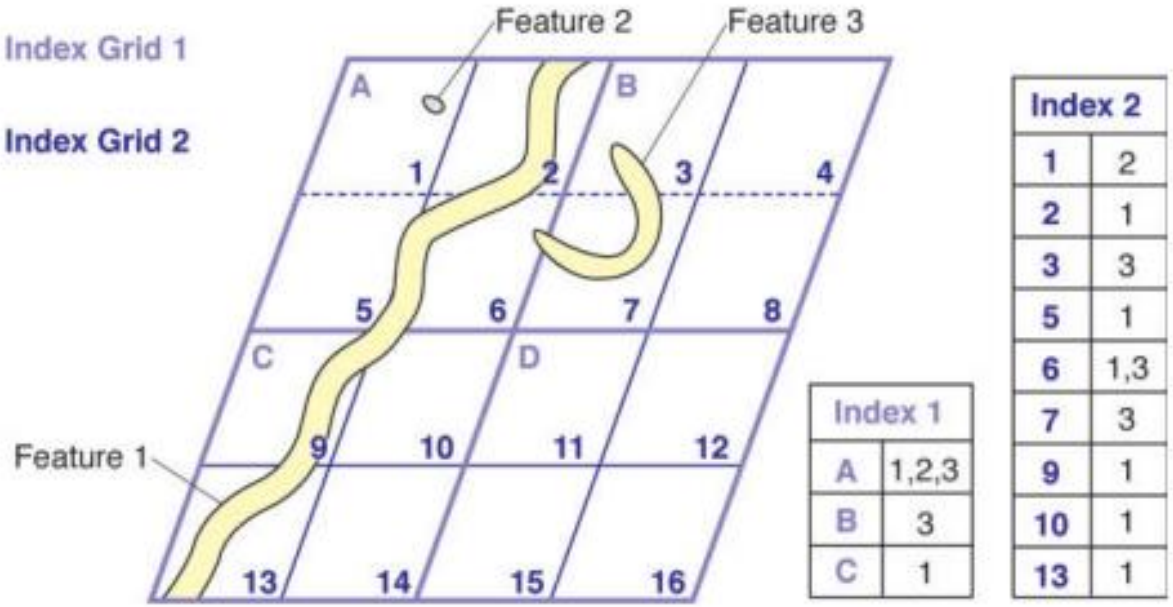
Indexing B-tree example

- Consider B-Tree indexing found in many DBMSs
  - Sort the original data into an ordered list
  - splits the ordered list into buckets of a given size (in this example it is four and then two)
  - the upper value for the bucket is stored
  - To find a specific value, such as 72, using the index involves a maximum of six tests: one at Level 1 (less than or greater than 36), one at Level 2 (less than or greater than 68), and a sequential read of four records at Level 3
  - Of course the larger the dataset, the more effective indexes are in retrieval performance

Longley, Goodchild, et al 2015. Geographic Information Science and Systems. Wiley.

Grid Indexing

- A grid index is similar to a mesh placed over a layer of geographic object
  - The highest (coarsest) grid (Index 1) splits the layer into four equal-sized cells.
  - Cell A includes parts of Features 1, 2, and 3;
  - Cell B includes a part of Feature 3; and Cell C has part of Feature 1.
  - There are no features on Cell D.
- The same process is repeated for the second-level index (Index 2).



Index 1	
A	1,2,3
B	3
C	1

Index 2	
1	2
2	1
3	3
5	1
6	1,3
7	3
9	1
10	1
13	1

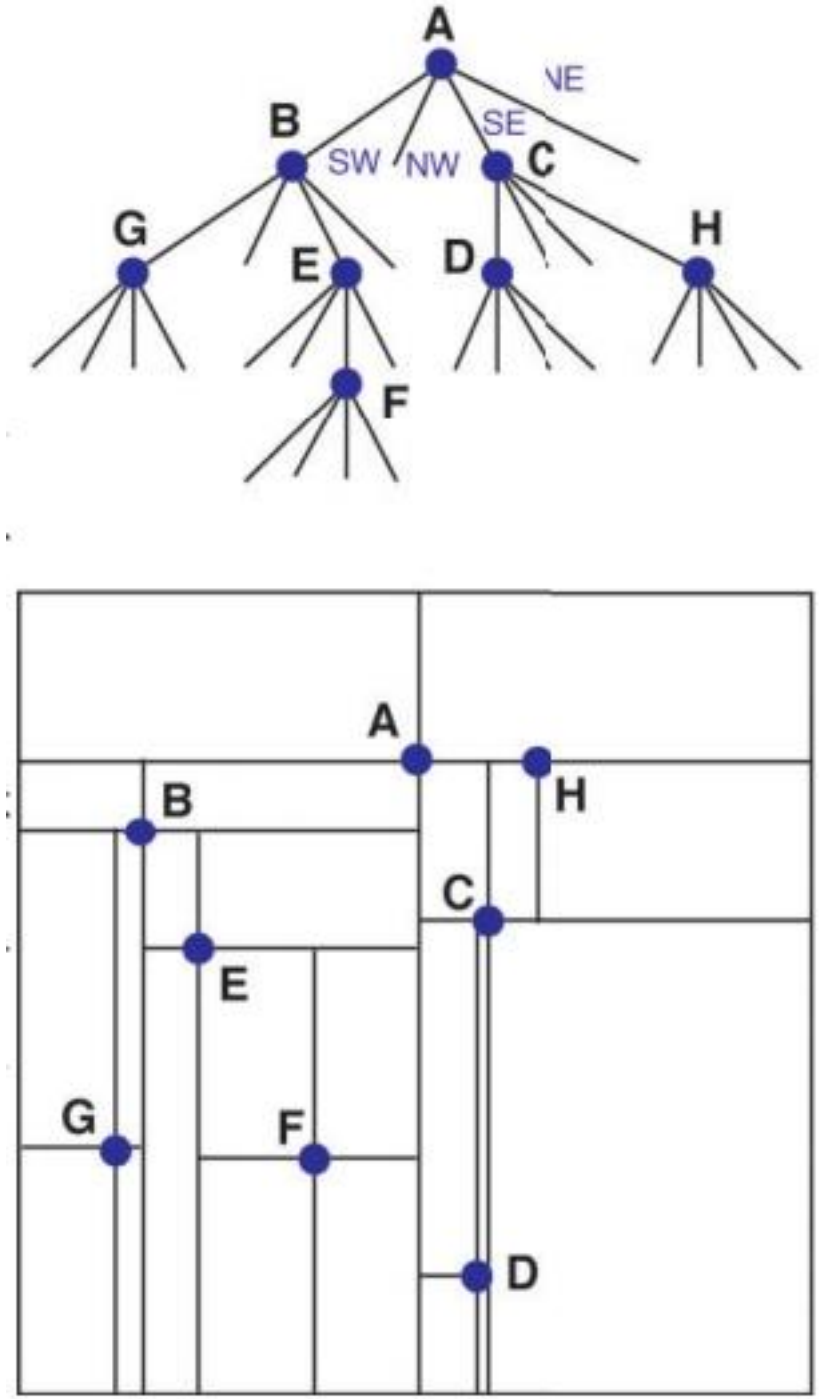
Longley, Goodchild, et al 2015. Geographic Information Science and Systems. Wiley.

A query to locate an object searches the indexed list first to find the object and then retrieves the object geometry or attributes for further analysis (e.g., tests for overlap, adjacency, or containment with other objects on the same or another layer). These two tests are often referred to as primary and secondary filters. Secondary filtering, which involves geometric processing, is much more computationally expensive.

Quadtree Indexes

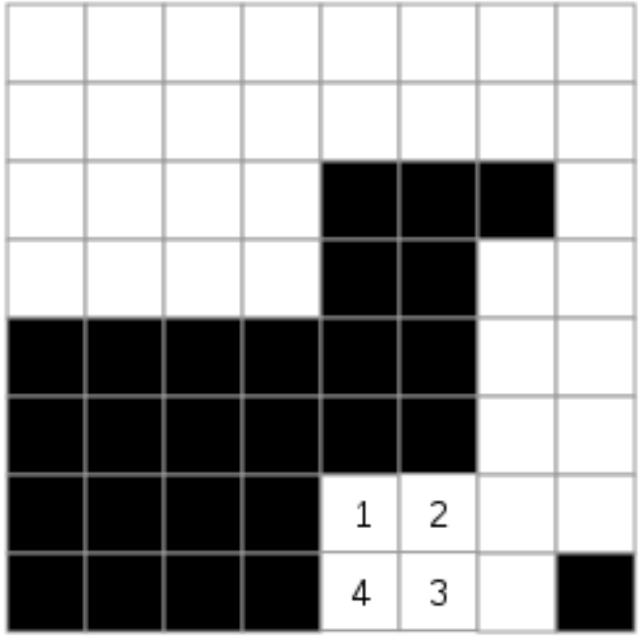
- In a point quadtree, space is divided successively into four rectangles based on the location of the points. The root of the tree corresponds to the region as a whole. The rectangular region is divided into four usually irregular parts based on the (x,y) coordinates of the first point. Successive points subdivide each new subregion into quadrants until all the points are indexed.

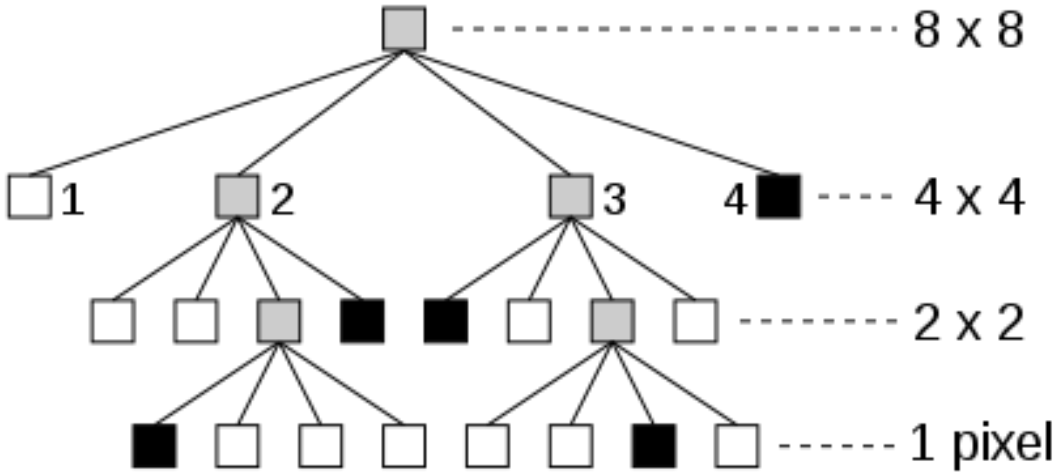
Quadtrees are used for both indexing and compressing geographic database layers. The many types of quadtrees can be classified according to the types of data that are indexed (points, lines, areas, surfaces, or rasters), the algorithm that is used to decompose (divide) the layer being indexed, and whether fixed or variable resolution decomposition is used.



Longley, Goodchild, et al 2015. Geographic Information Science and Systems. Wiley.

Quadtree Indexes





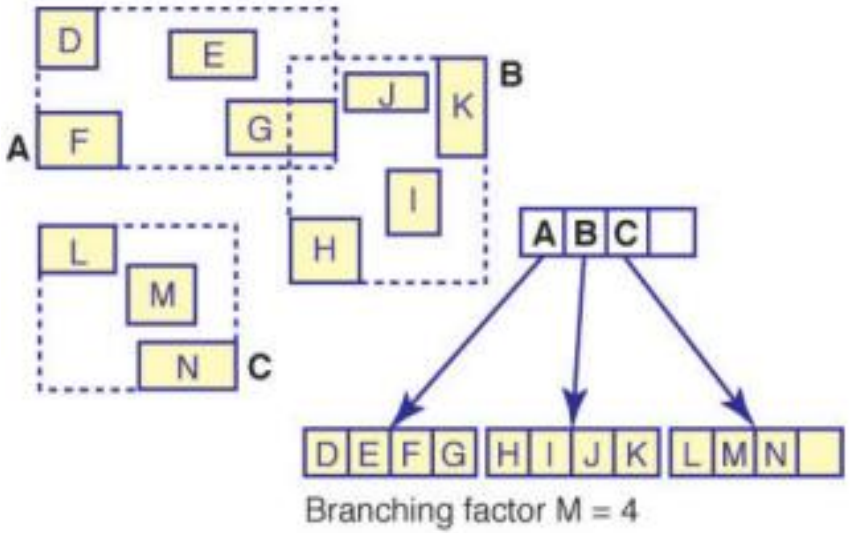
http://wiki.gis.com/wiki/index.php/Quadtree

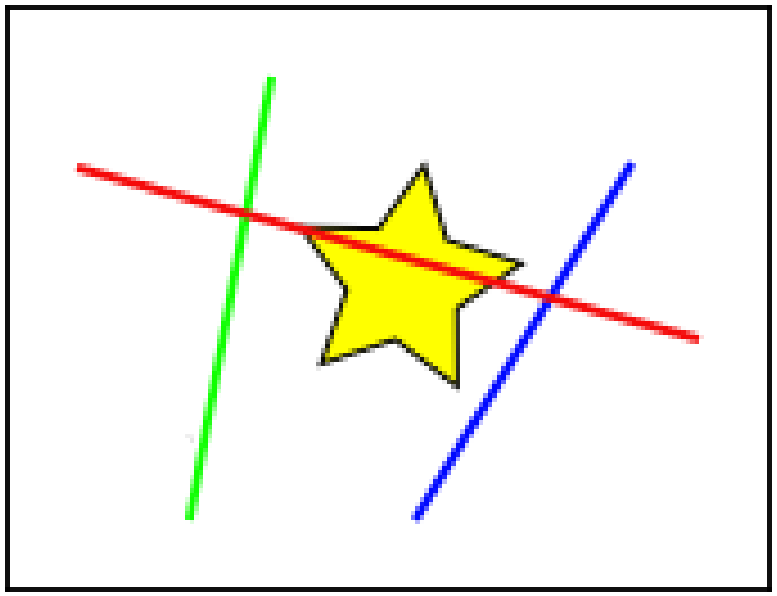
R-Tree Indexes

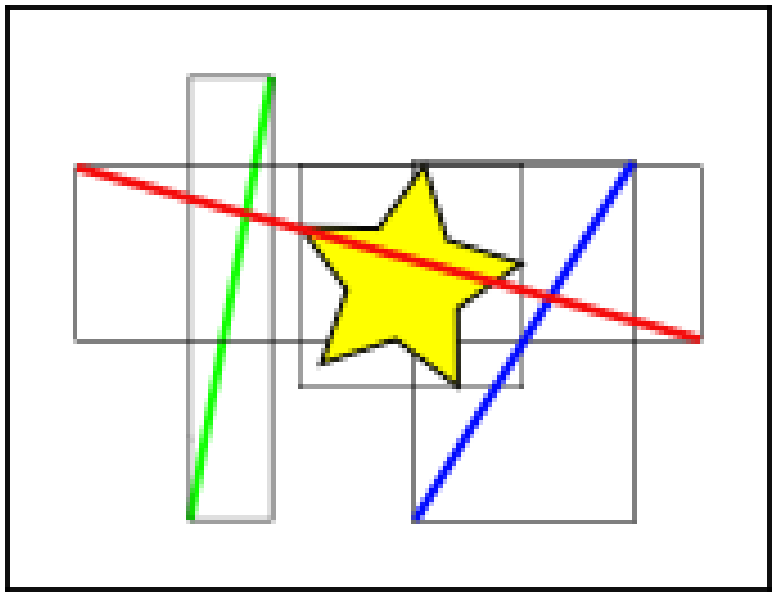
- R-trees group objects using a rectangular approximation of their location called a minimum bounding rectangle (MBR)
  - Groups of point, line, or area objects are indexed based on their MBR
- Objects are added to the index by choosing the MBR that would require the least expansion to accommodate each new object.
- If the object causes the MBR to be expanded beyond some preset parameter, then the MBR is split into two new MBRs.

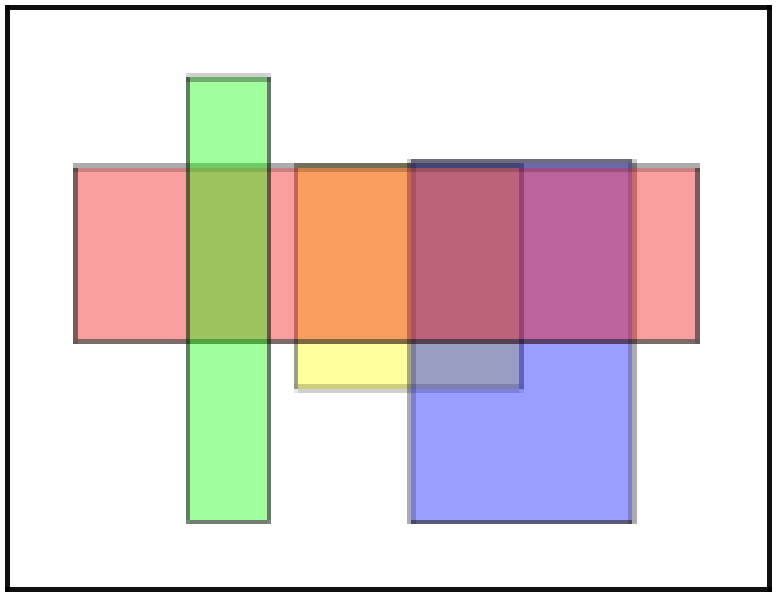
R-Tree Indexes

- The lowest level contains three “leaf nodes”;
- the highest has one node with pointers to the MBR of the leaf nodes.
- The MBR is used to reduce the number of objects that need to be examined in order to satisfy a query.
- R-trees are popular methods of indexing geographic data because of their flexibility and excellent performance.









R-Tree Indexes

- The number of lines that intersect the yellow star is **one**, the red line. But the bounding boxes of features that intersect the yellow box is **two**, the red and blue ones.
- The way the database efficiently answers the question "what lines intersect the yellow star" is to first answer the question "what boxes intersect the yellow box" using the index (which is very fast) and then do an exact calculation of "what lines intersect the yellow star" **only for those features returned by the first test**.