Data Modeling

Man is a knot, a web, a mesh into which relationships are tied. Only those relationships matter

Saint-Exupéry

Data modeling

- A technique for modeling data
- A graphical representation of a database
- The goal is to identify the facts to be stored in the database
- Data modeling is a partnership between the client and analyst

The building blocks

- Entity
- Attribute
- Relationship
- Identifier

Data model quality

- A well-formed data model
- A high fidelity image

A well-formed data model

- Construction rules obeyed
- No ambiguity
  - All entities, attributes, relationships, and identifiers are defined
  - Names are meaningful to the client

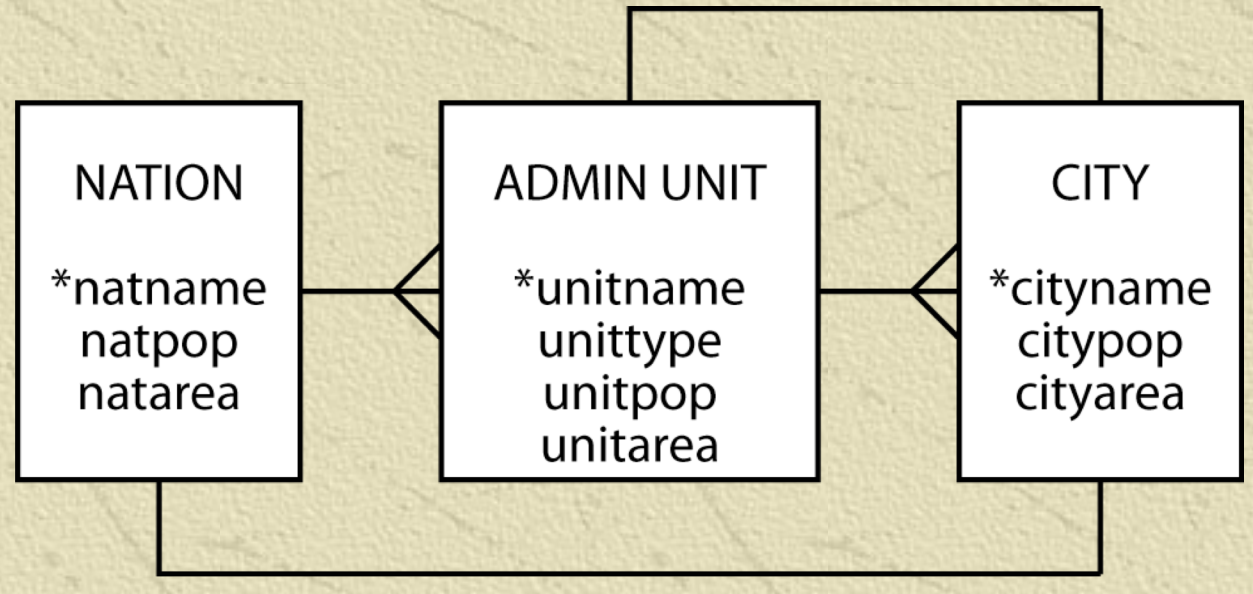
A high fidelity image

- Faithfully describes the world it is supposed to represent
- Relationships are of the correct degree
- Data model is complete, understandable, and accurate
- The data model makes sense to the client

Quality improvement

- Is the level of detail correct?
- Are all exceptions handled?
- Is the model accurate?

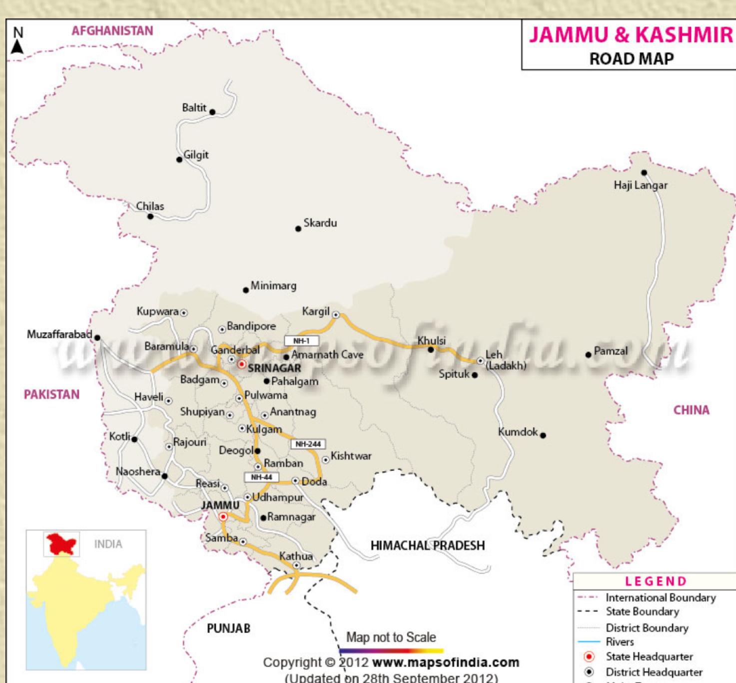
Pure geography



Can a nation have more than one capital?
Can a city be the capital of more than one state?

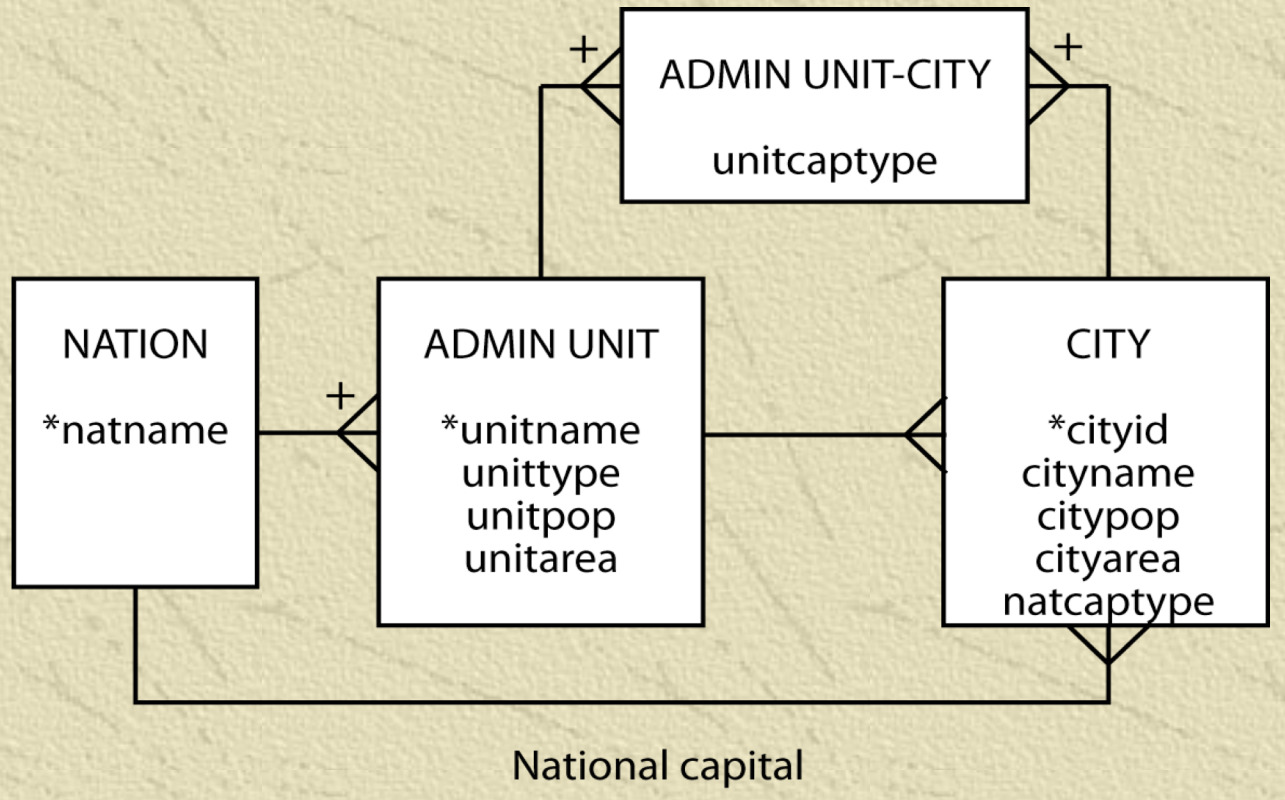

SOUTH AFRICA
BOTSWANA
MOZAMBIQUE
NAMIBIA
Gaborone
Kruger National Park
Kgalagadi Transfrontier Park
Pretoria
Johannesburg
Soweto
Nelspruit
Mbabane
Maputo
SWAZILAND
Greater St Lucia Wetland Park
Ladysmith
Upington
Bergville
Maseru
Drakensberg
Springbok
Bloemfontein
LESOTHO
Durban
Orange River
Umtata
Lusikisiki
Port St Johns
INDIAN OCEAN
The Shipwreck Coast
Grahamstown
East London
Garden Route
Port Elizabeth
Cape Town
Mossel Bay
George
Jeffrey's Bay
Cape of Good Hope
Cape Agulhas
0 250 km
0 150 miles


Bolivia
International Boundary
Department Boundary
Road
River
National Capital
Department Capital
City or Town
© 2007 Geology.com
PANDO
BRAZIL
PERU
LA PAZ
BENI
BOLIVIA
SANTA CRUZ
COCHABAMBA
ORURO
CHILE
POTOSI
CHUQUISACA
TARIJA
PARAGUAY
ARGENTINA
South Pacific Ocean
La Paz
Sucre
Lago Titicaca
Cochabamba
Oruro
Potosi
Santa Cruz
Trinidad
Cobija
Tarija
Uyuni
Camiri
Riberalta
Guajará-Mirim
Puerto Suarez
Corumba
Arica
Iquique
Antofagasta

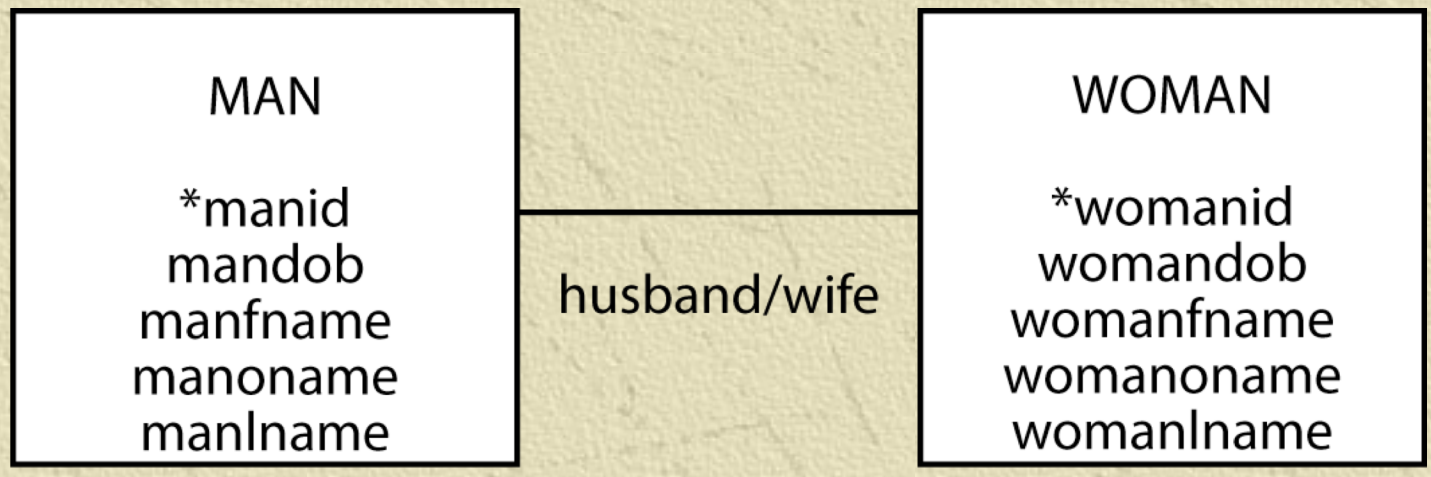

JAMMU & KASHMIR
ROAD MAP
AFGHANISTAN
PAKISTAN
CHINA
HIMACHAL PRADESH
PUNJAB
Baltit
Gilgit
Chilas
Skardu
Haji Langar
Minimarg
Kupwara
Kargil
Bandipore
Muzaffarabad
Baramula
Ganderbal
Amarnath Cave
SRINAGAR
Pahalgam
Khulsi
Leh (Ladakh)
Spituk
Pamzal
Badgam
Pulwama
Haveli
Shupiyan
Anantnag
Kulgam
Kotli
Rajouri
Deogol
Kishtwar
Kumdok
Ramban
Naoshera
Reasi
Doda
Udhampur
JAMMU
Ramnagar
Samba
Kathua
INDIA
NH-1
NH-244
NH-44
Map not to Scale
Copyright © 2012 www.mapsofindia.com
(Updated on 28th September 2012)
LEGEND
International Boundary
State Boundary
District Boundary
Rivers
State Headquarter
District Headquarter
Major Towns


ORISSA
MAHARASHTRA
CHHATTISGARH
KARNATAKA
HYDERABAD
Telangana
ANDHRA PRADESH

Geography revised

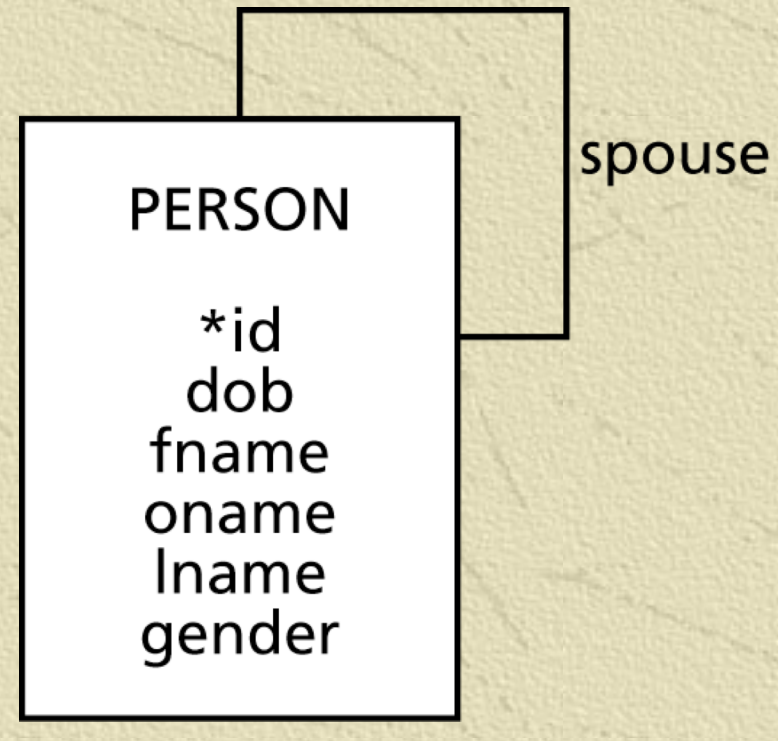


Family matters - take 1

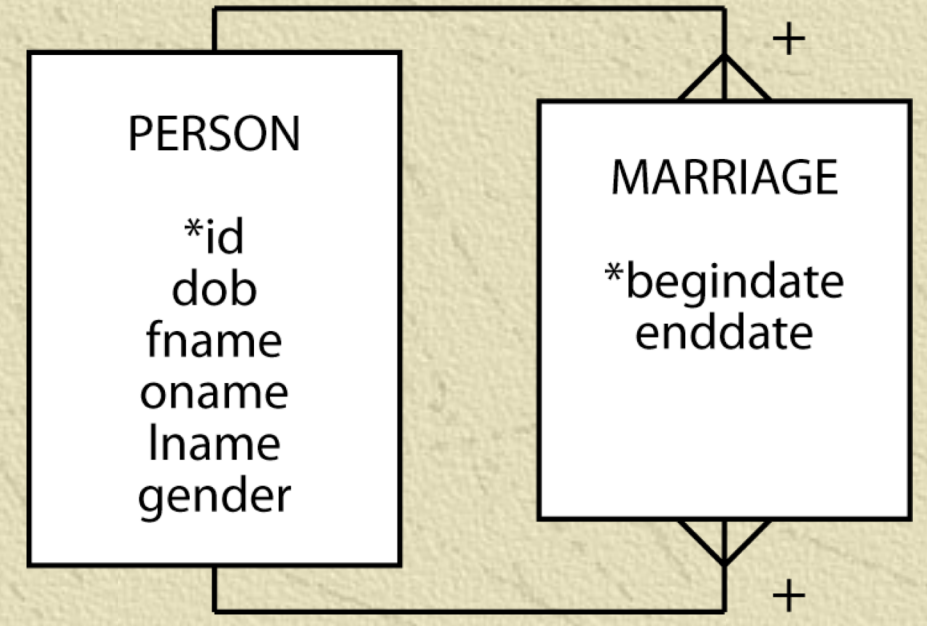


Can we generalize?

Family matters - take 2

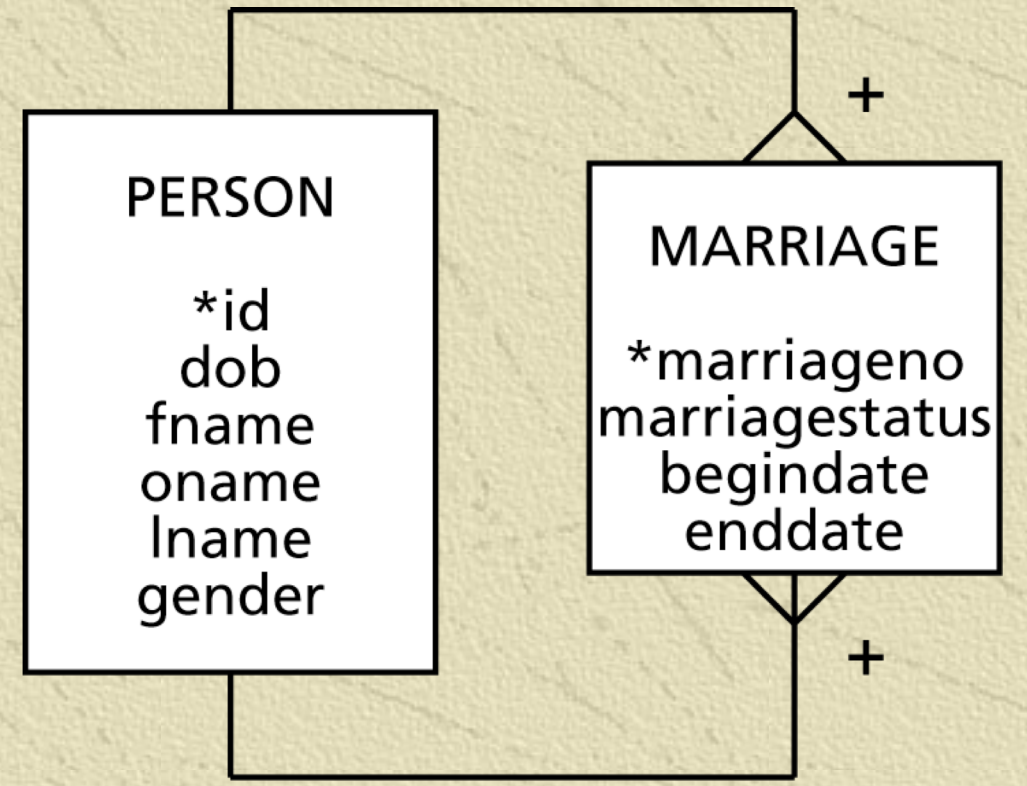


Family matters - take 3

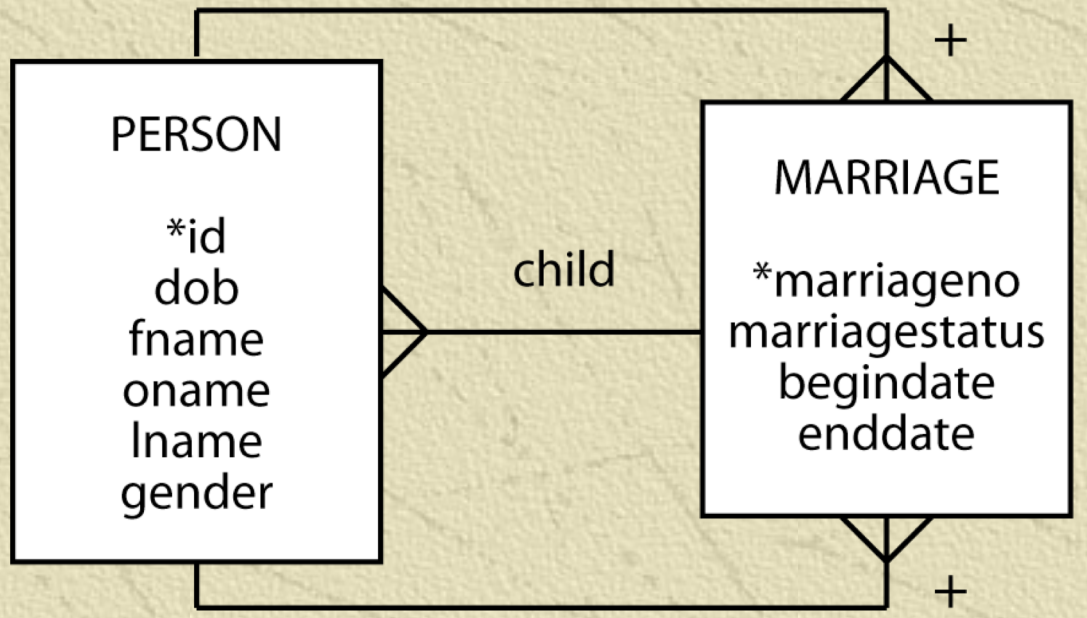


What about couples who are not officially married but have cohabited for an extended period?

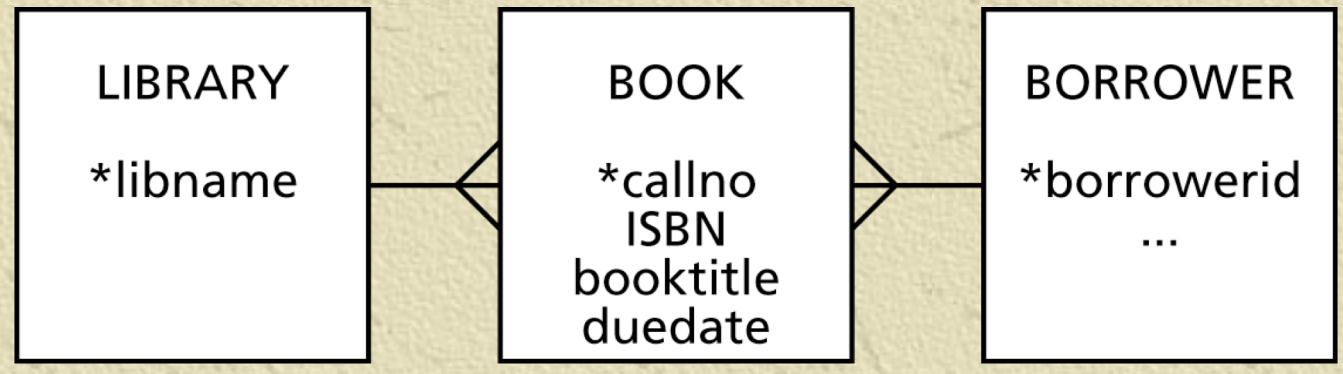
Family matters - take 4



Family matters - take 5

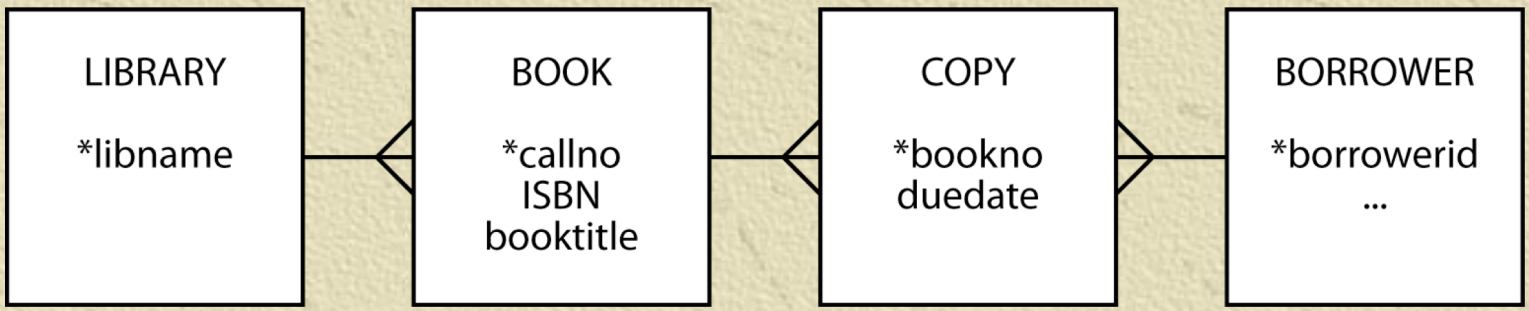


Bookish matters - take 1



Should copyno be an attribute of book?

Bookish matters - take 2

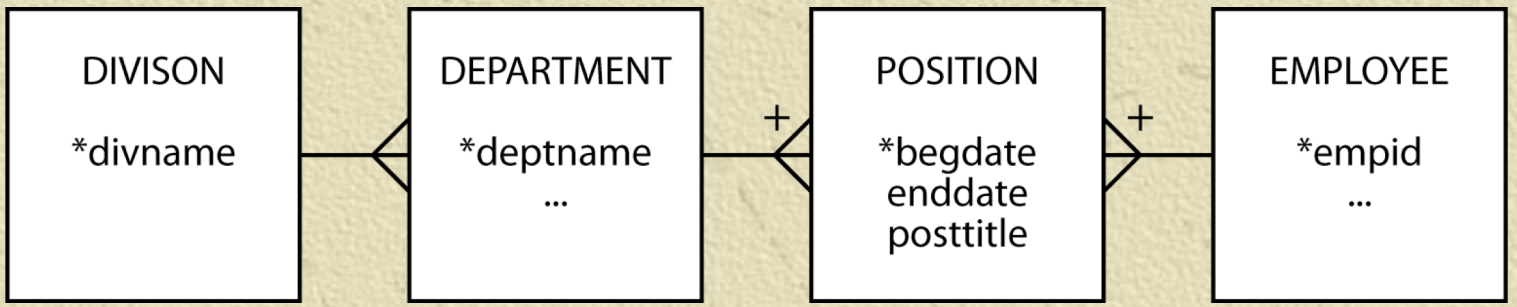


This model records only the current borrower of a copy of a book

History - take 1

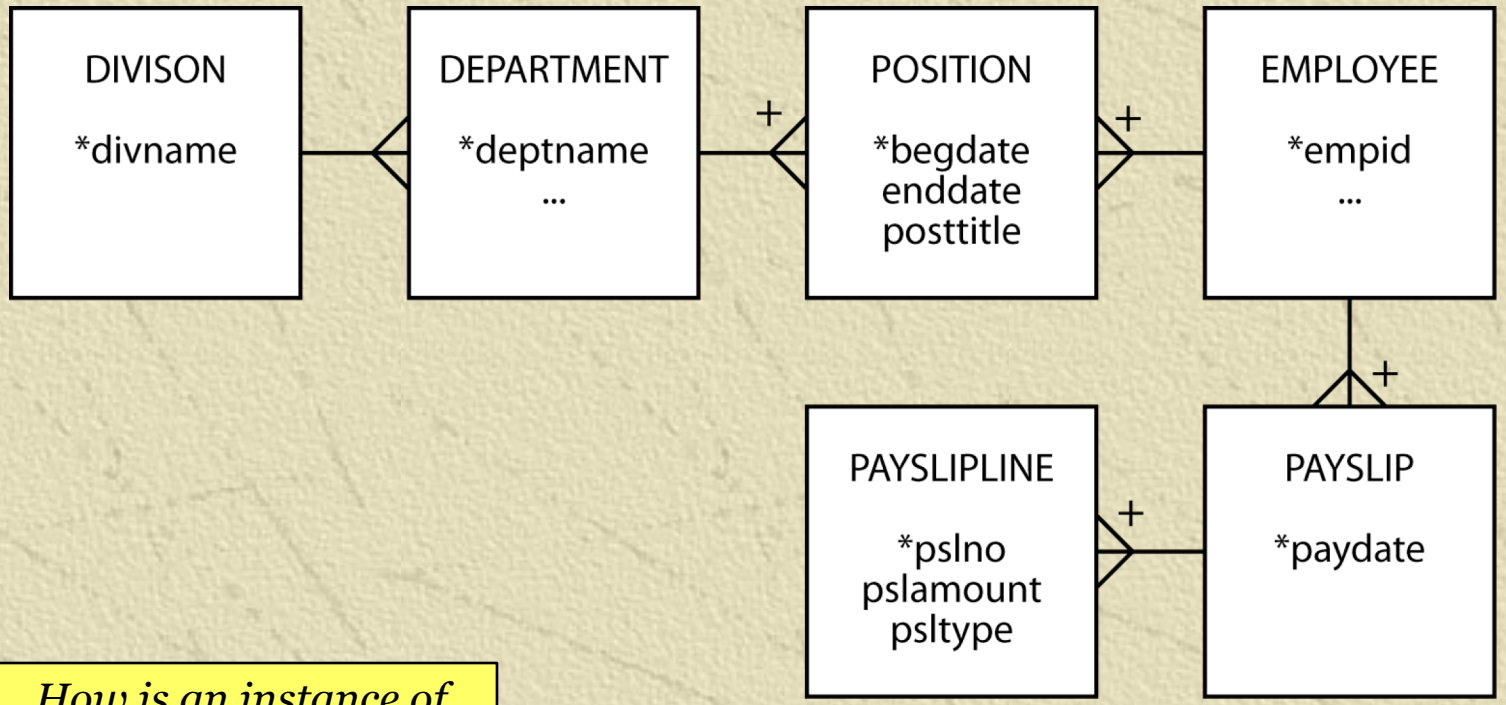
Can an employee work in multiple departments?

History - take 2



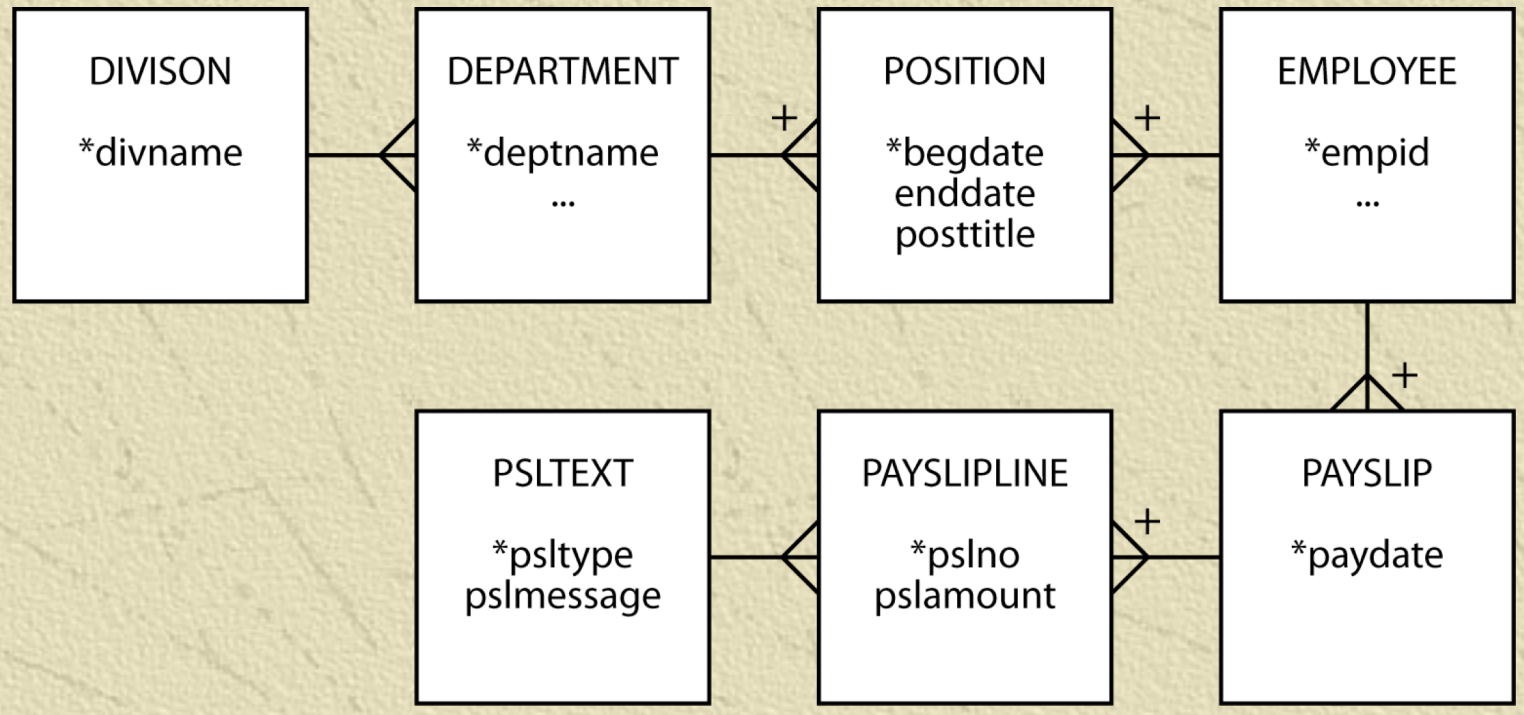
How do we keep track of an employee's pay checks?

History - take 3

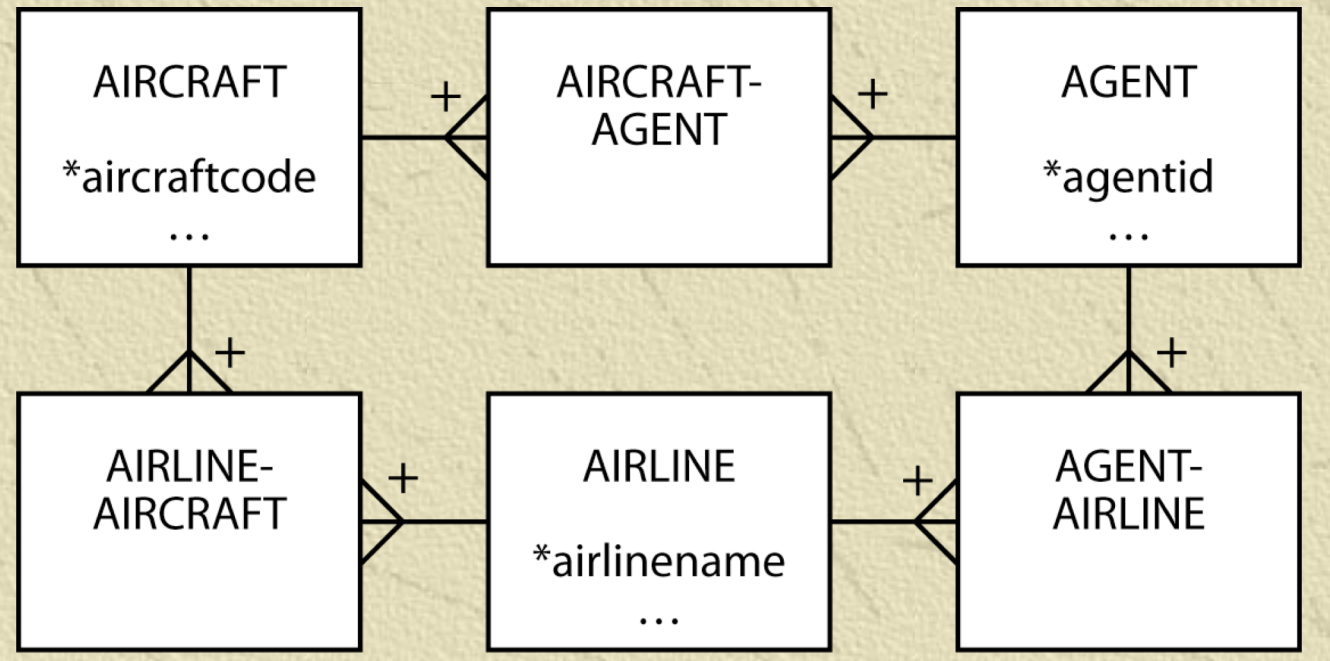


How is an instance of PAYSLIPLINE identified?

History - take 4

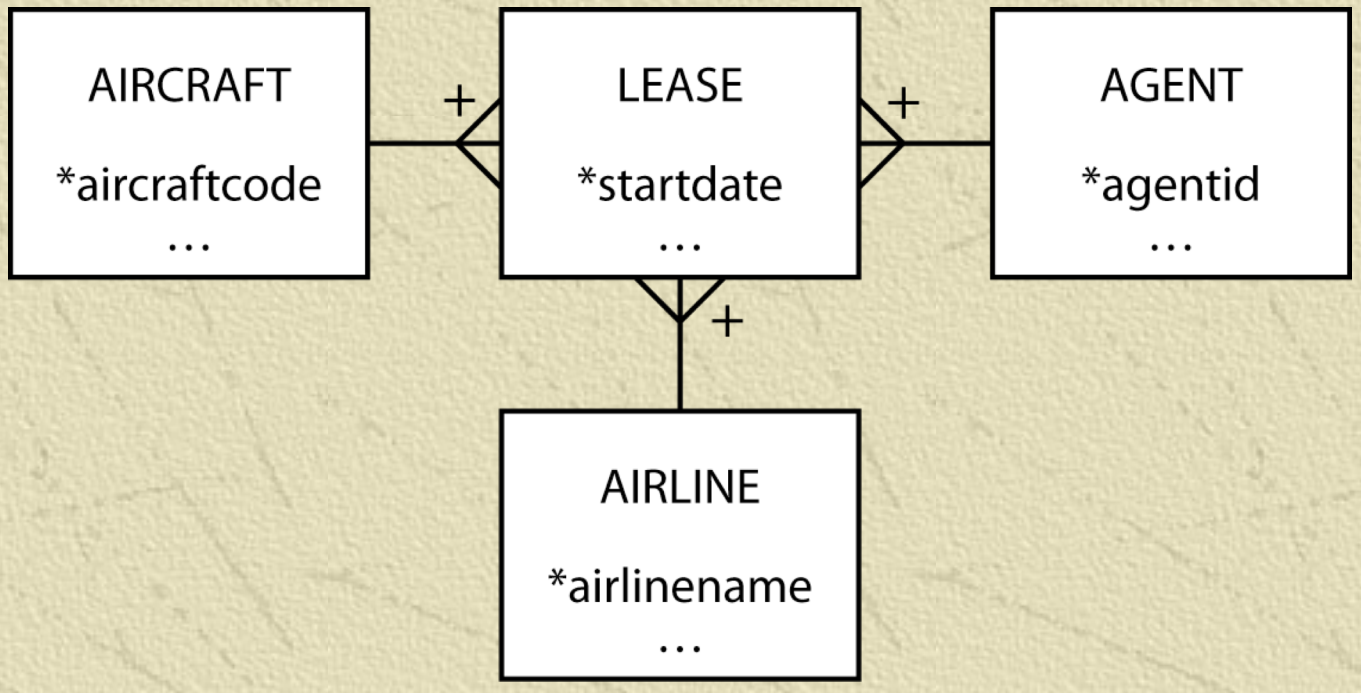


A ménage à trois for entities - take 1



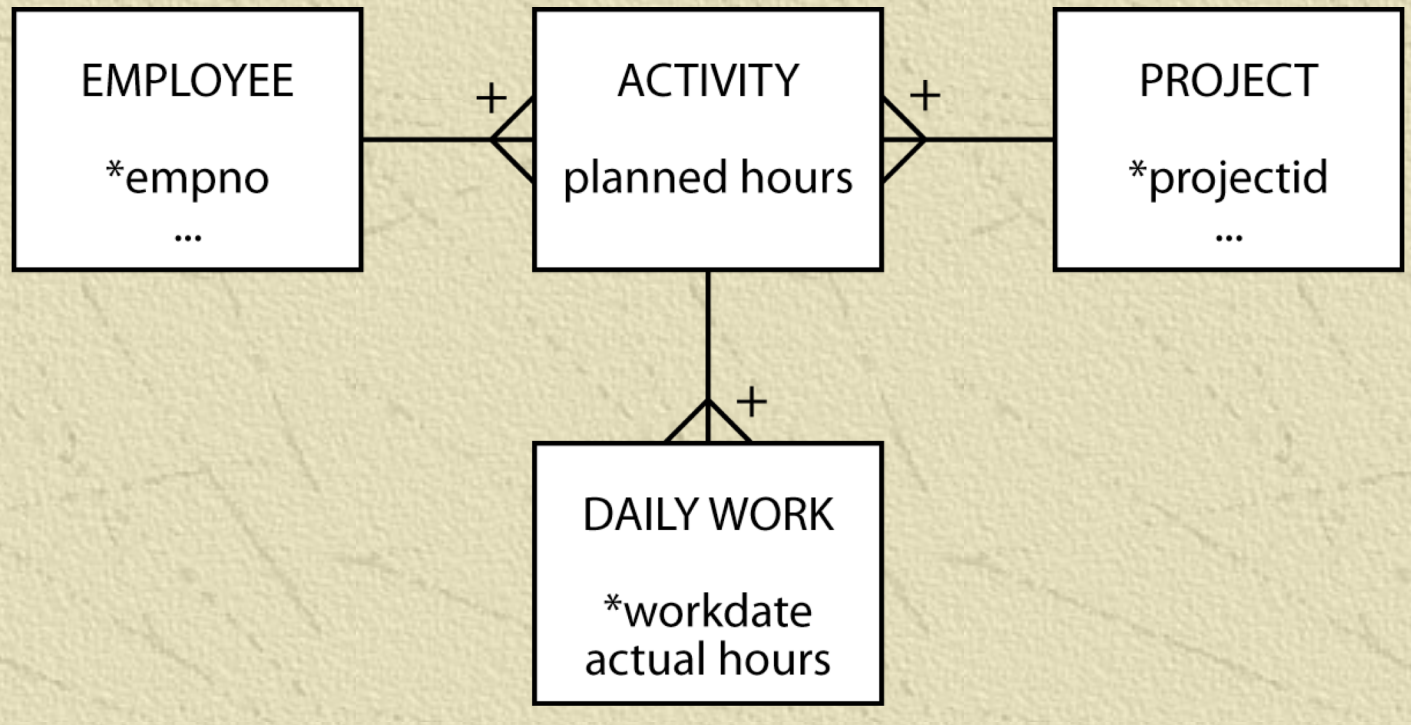
Where do we store information about the lease?

A ménage à trois for entities - take 2

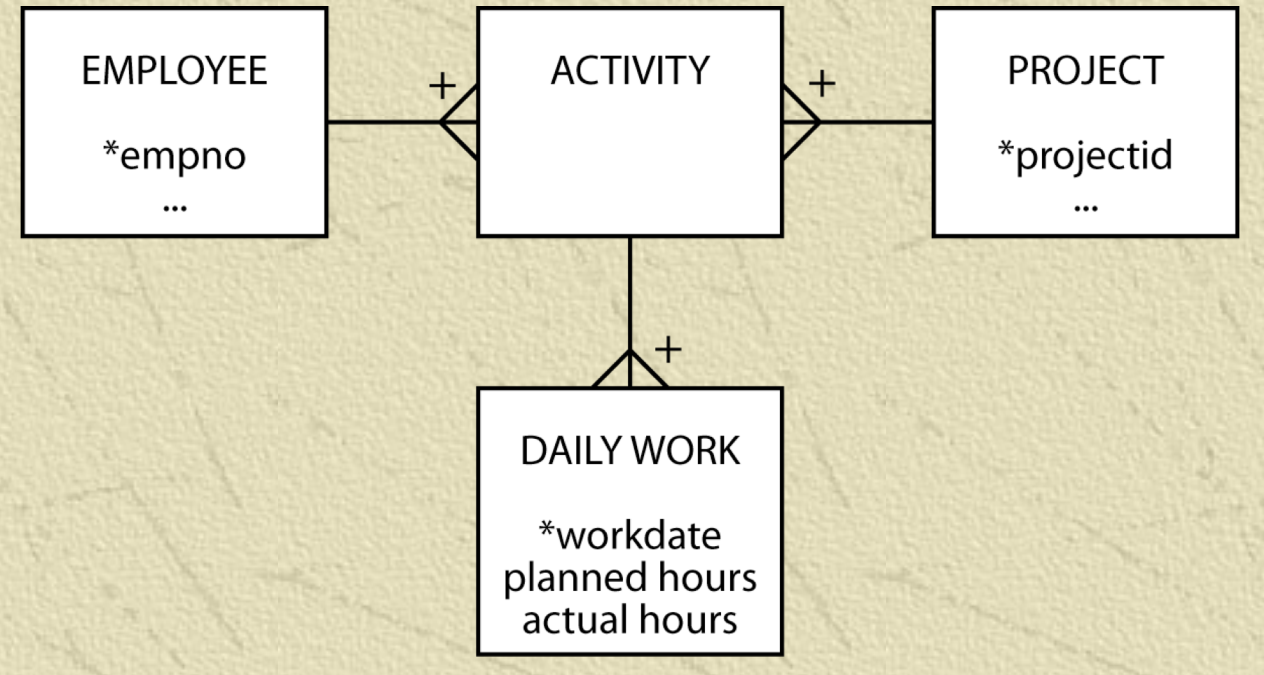


Why is start date part of the composite primary key?

Planning and doing - take 1



Planning and doing - take 2



Cardinality

Cardinality	Modality	Meaning
0,1	Optional	There can be zero or one instances of the entity relative to the other entity
0,n		There can be zero or many instances of the entity relative to the other entity
1,1	Mandatory	There is exactly one instance of the entity relative to the other entity
1,n		The entity must have at least one and can have many instances relative to the other entity

Minimalist approach

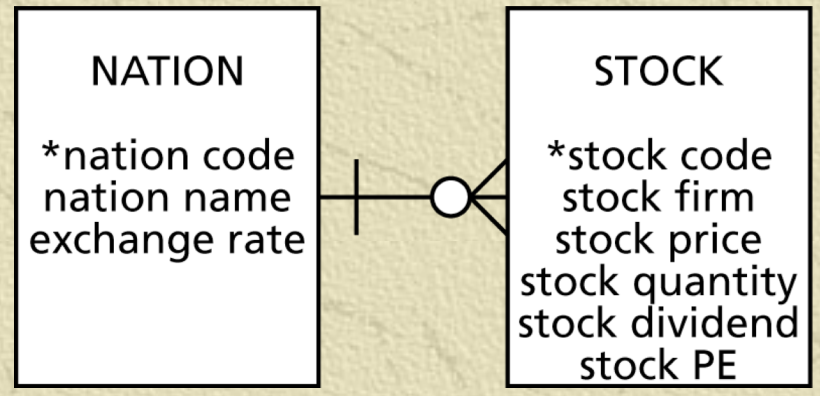
- Focus has been on identifying the basic cardinality (1:m or m:m?)
- Now add greater precision
  - There must be 1 instance
- Learn the basics and then add more detail

Modality

- Also known as optionality
- Cardinality indicates the range of instances in a relationship
- Modality defines the minimum number of instances
- Cardinality and modality are linked

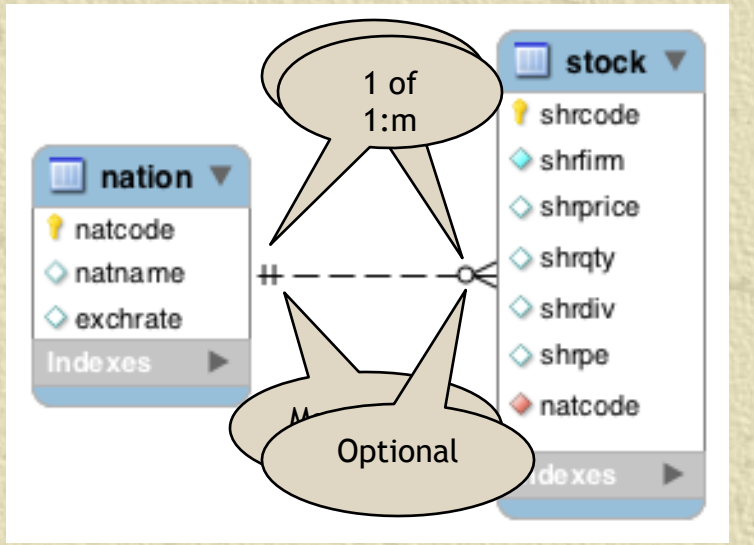
Modality and Cardinality

- Optional entity
  - Cardinality is 0
  - O
- Mandatory entity
  - Cardinality is 1
  - |

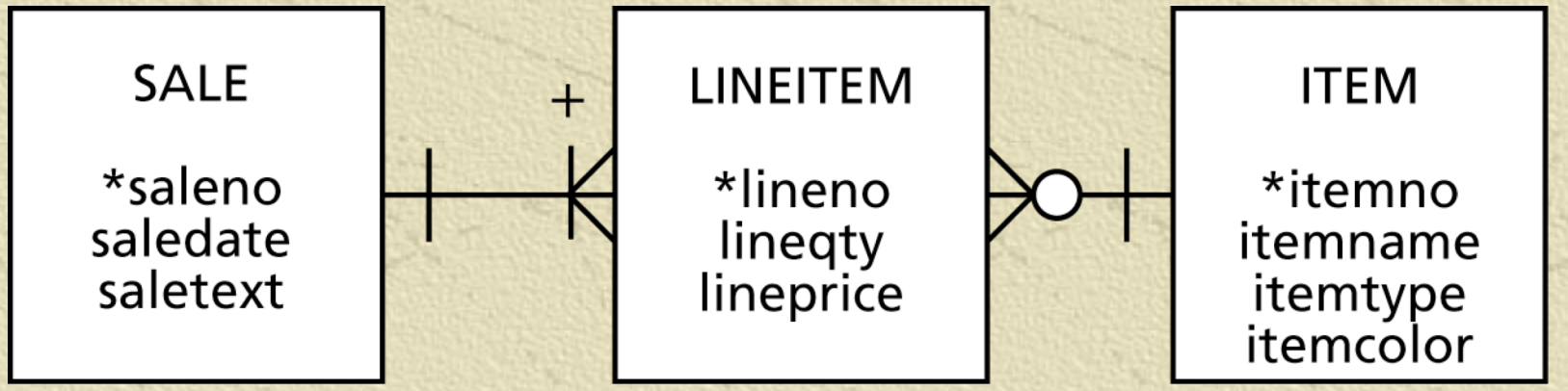


Cardinality

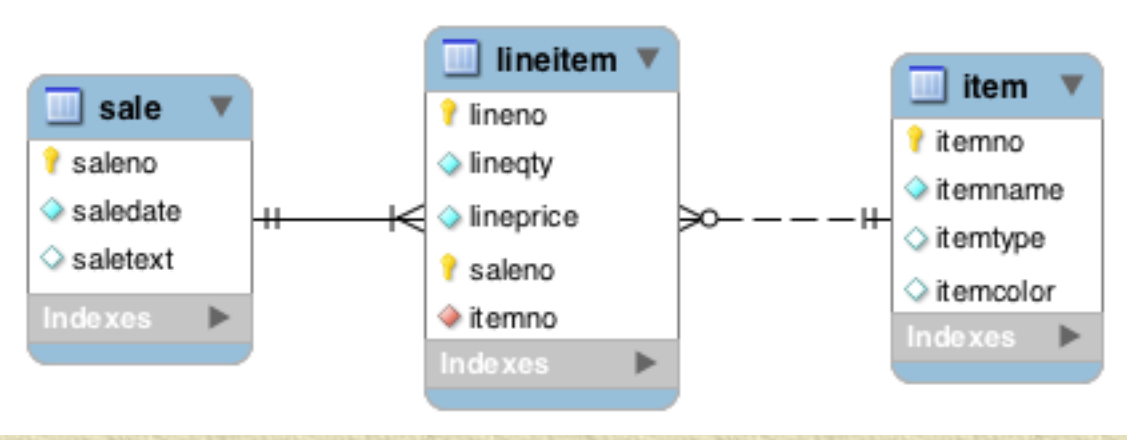
Modality



Modality

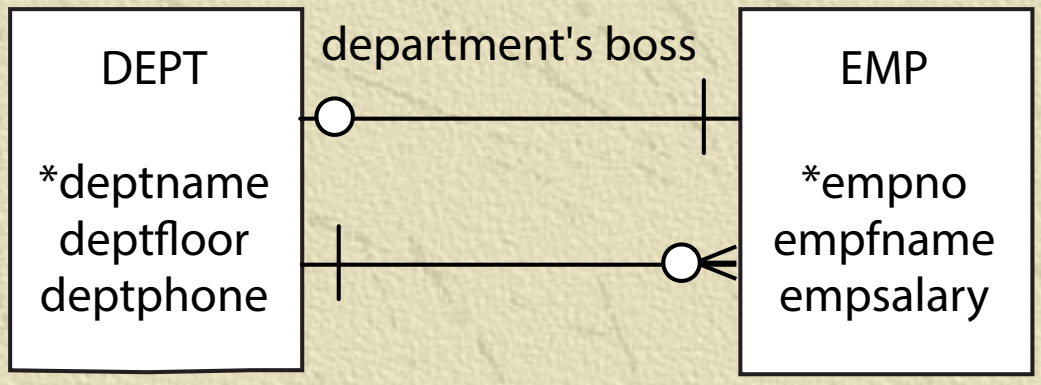


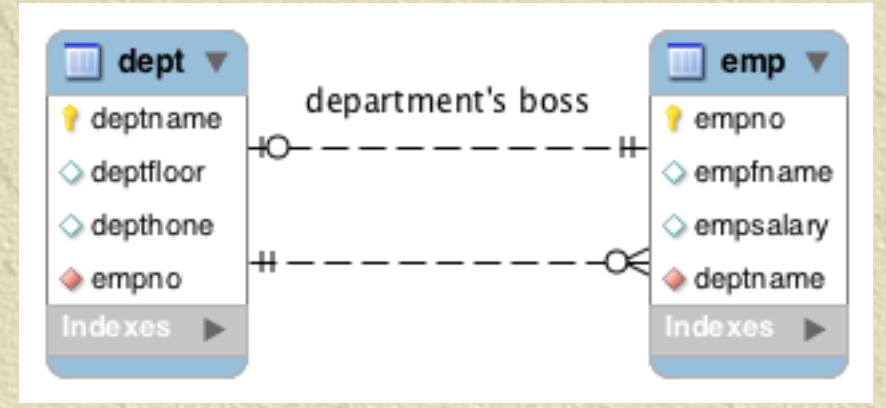
Can a lineitem exist without a sale?



Can an item exist without a lineitem?

Modality

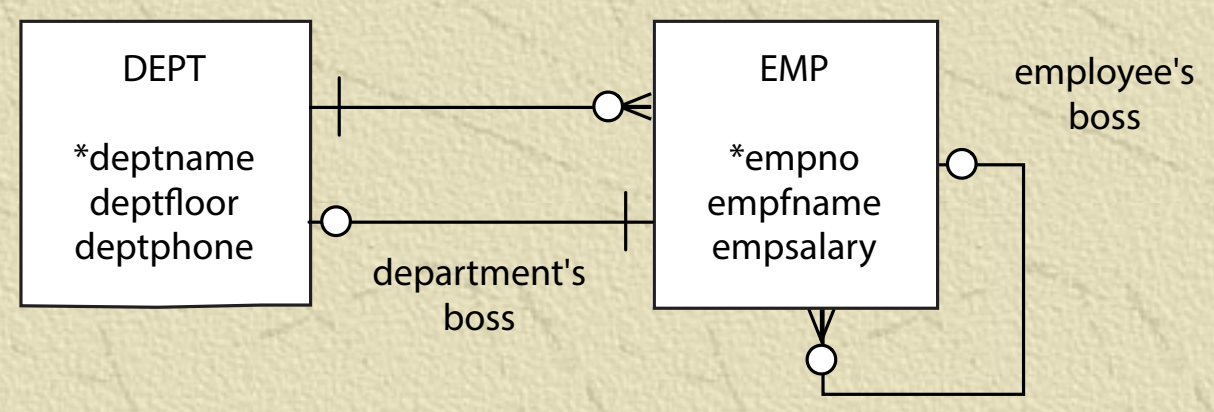


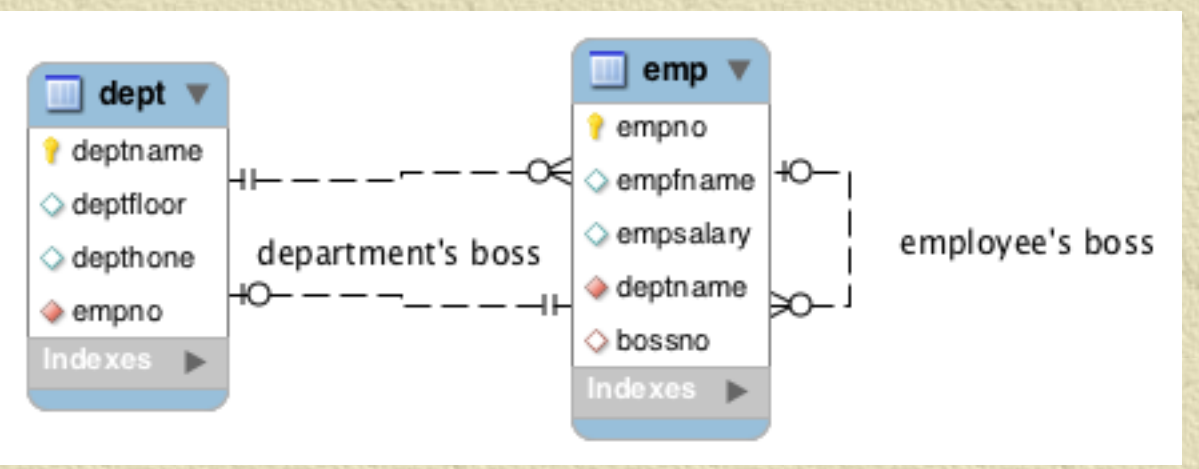


Is every employee a department boss?

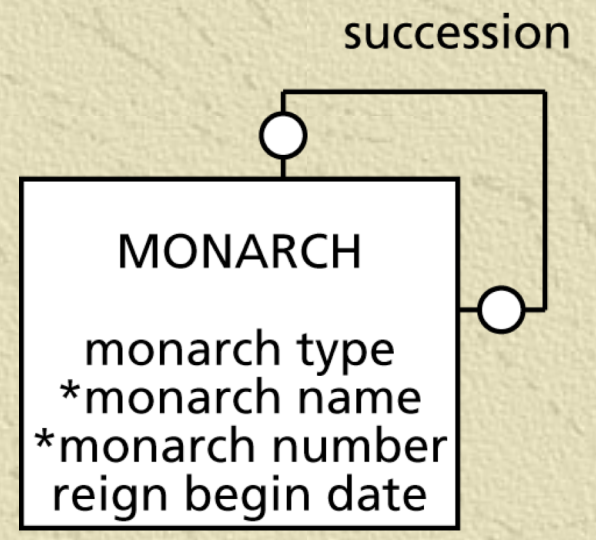
Does every department have a boss?

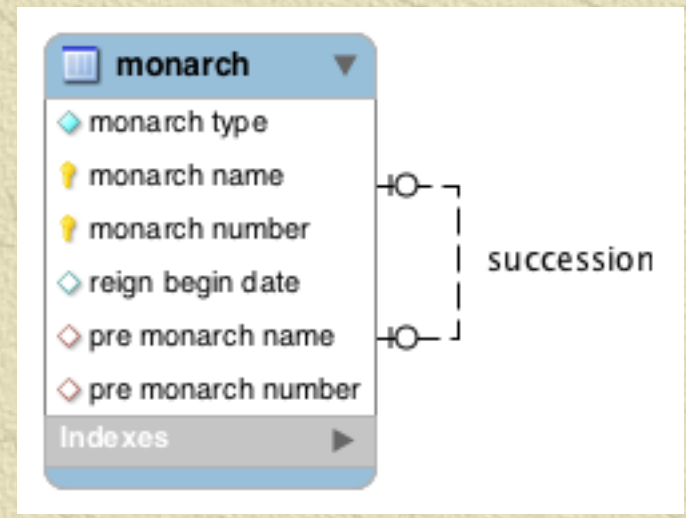
Modality


DEPT
*deptname
deptfloor
deptphone
EMP
*empno
empfname
empsalary
employee's
boss
department's
boss


dept
deptname
deptfloor
depthone
empno
Indexes
department's boss
emp
empno
empfname
empsalary
deptname
bossno
Indexes
employee's boss

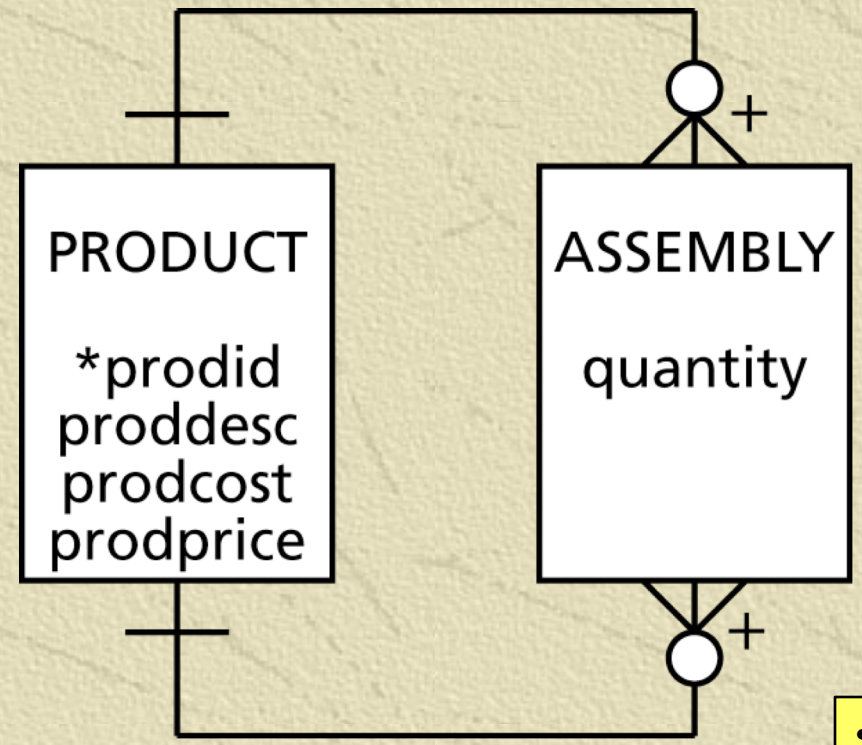
Modality

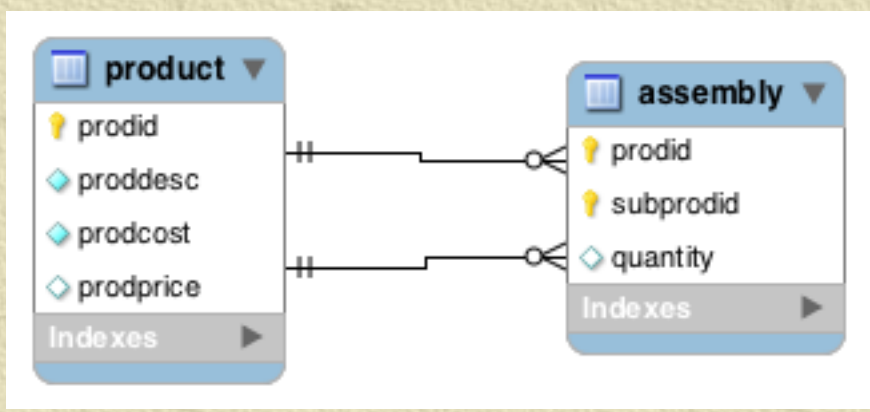




Why is it optional for a monarch to have a successor?

Modality





- *Optional for a product to have components*
- *Optional for a product to be a component*
- *Every assembly must have products*

Modality

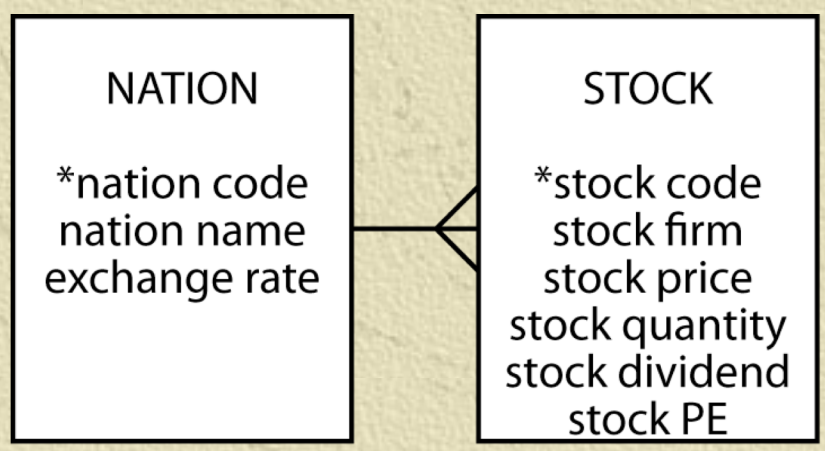
- Adds additional information to a data model
- If a relationship is mandatory then add a constraint
  - Could be
    - Referential integrity constraint
    - Application logic

Entity types

- Independent
- Dependent
- Associative
- Aggregate
- Subordinate

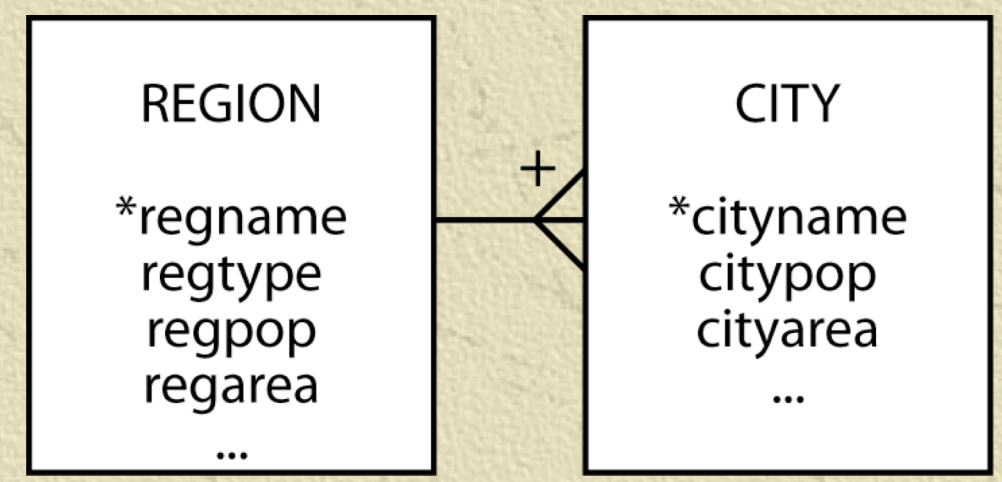
Independent

- Often a starting point
- Prominent in the client's mind
- Often related to other independent entities



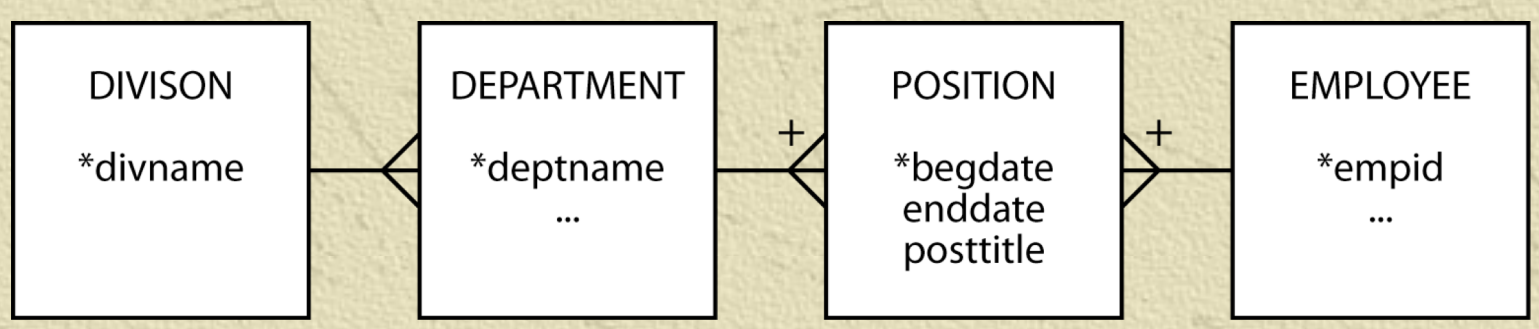
Dependent

- Relies on another entity for its existence and identification
- Can become independent if given an arbitrary identifier



Associative

- A by-product of an m:m relationship
- Typically between independent entities
- Can store current or historical data
- Can become independent if given an arbitrary identifier

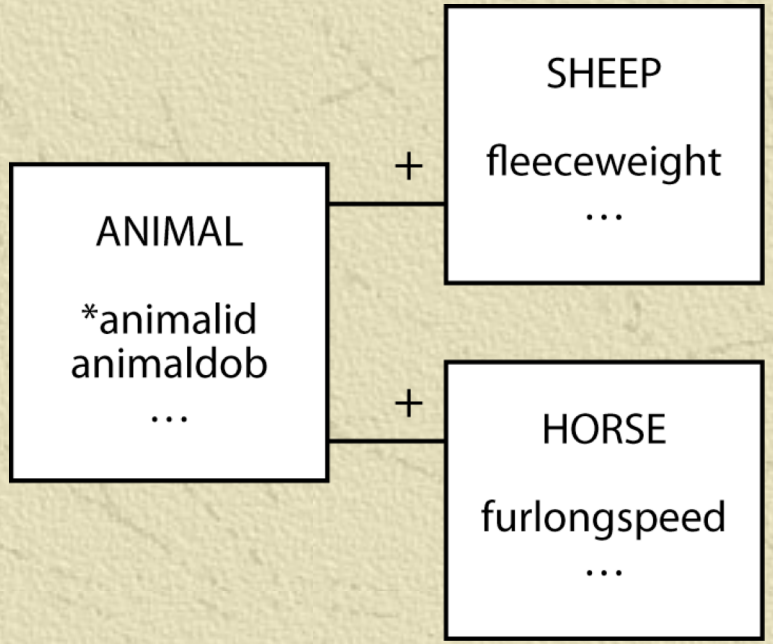


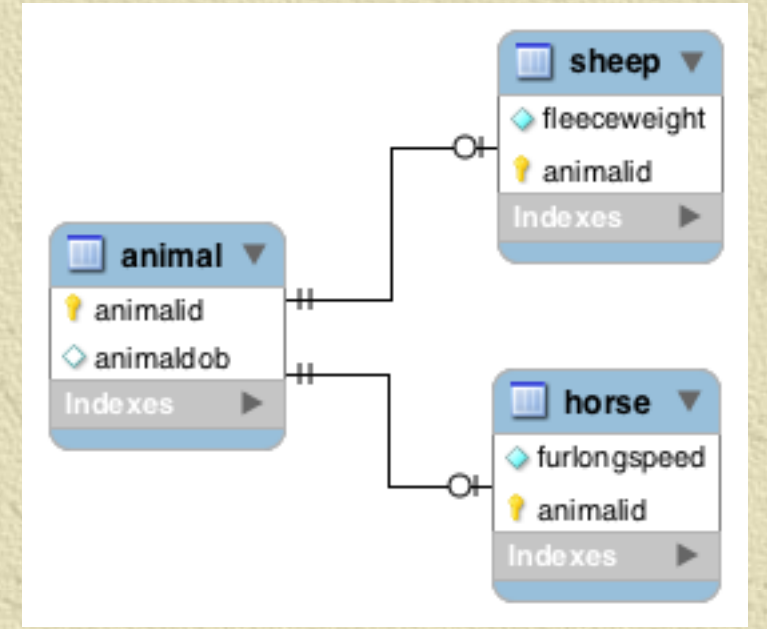
Aggregate

- Created from several different entities that have a common prefix or suffix
- Commonly used with addresses or names

Subordinate

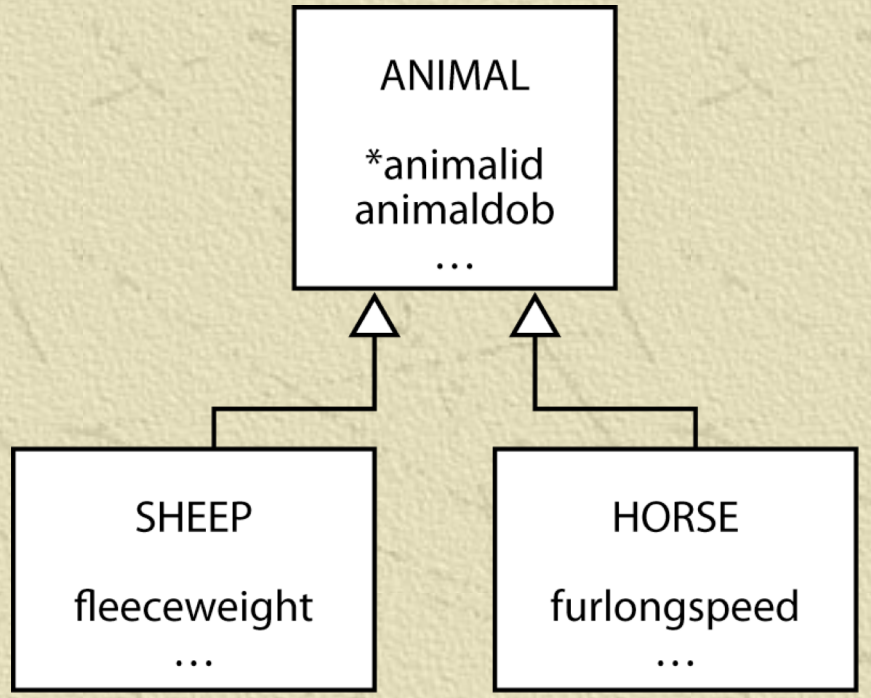
- An entity with data that can vary among instances





Generalization

- A relationship between a more general element and a more specific element



Generalization

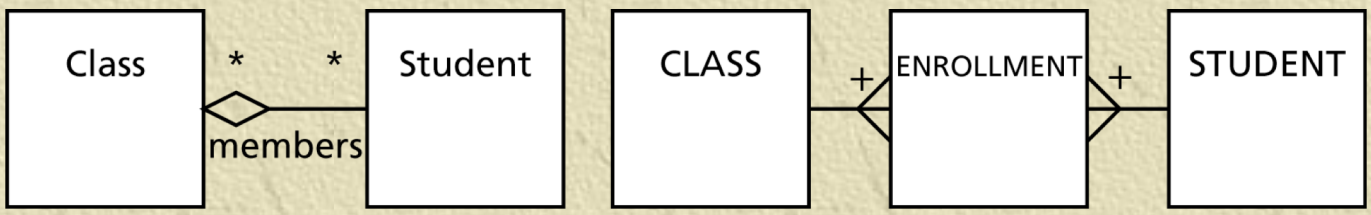
- Map with one table for each entity
- For each of the subtype entities the primary key is that of the supertype entity
- You must also make this column a foreign key so that a subtype cannot be inserted without the presence of the matching supertype

UML aggregation

- Aggregation is a part-whole relationship between two entities

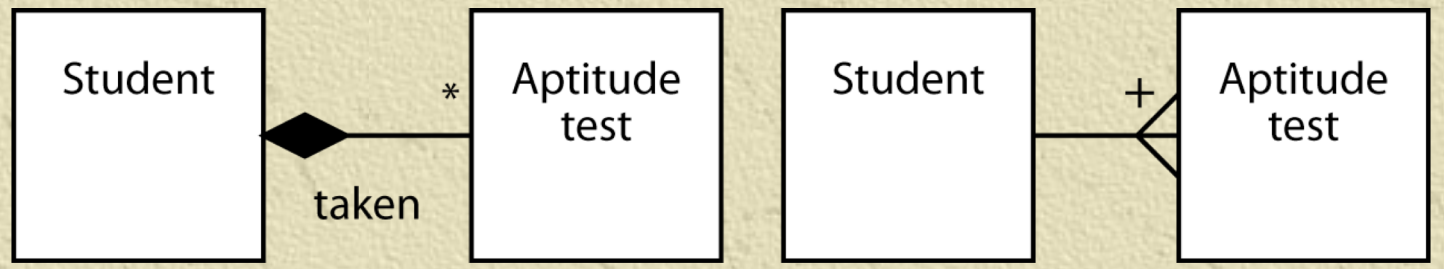
UML shared aggregation

- One entity owns another entity, but other entities can own that entity as well

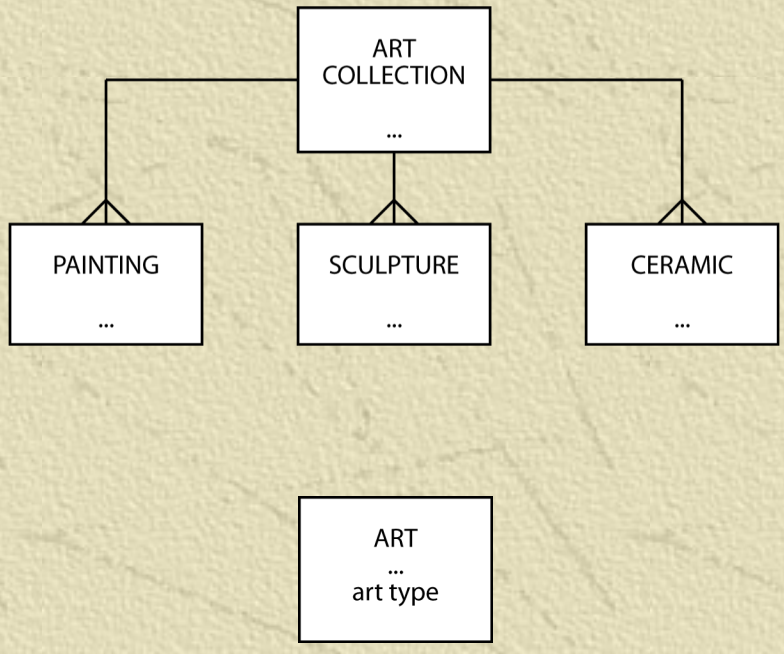


UML composite aggregation

- One entity exclusively owns the other entity



Data model contraction

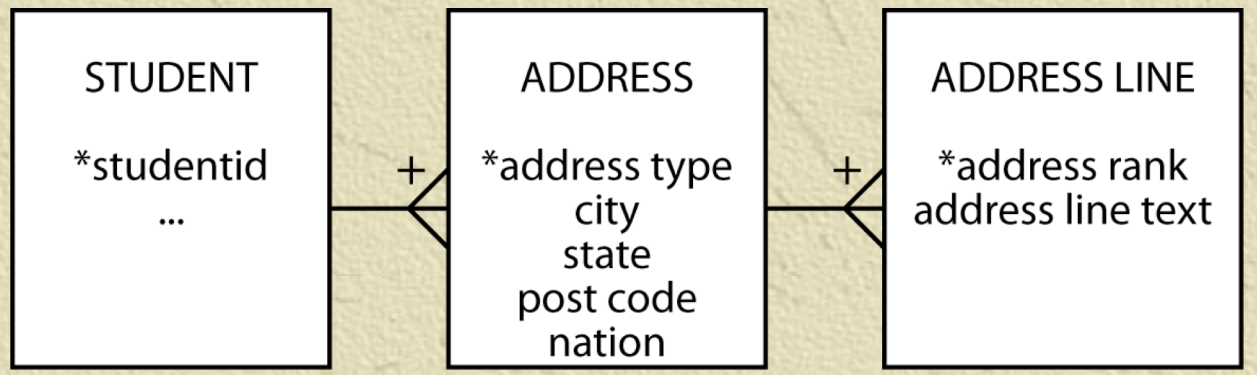


Hints on data modeling

- The model will expand and contract
- Invent identifiers where necessary
- Identifiers should have only one purpose – identification
- A data model does not imply ordering
- Create an attribute if ordering of instances is required
- An attribute's meaning must be consistent

Names and addresses

- The query test
  - If an attribute has parts, are any of the parts ever likely to appear in a query?
- Have an understanding on representing names and addresses in a data model



Post code

- A US zip code is CHAR(5) because leading zeroes are displayed
  - Boston MA 02201
- Full US zip is CHAR(10)
  - 30602-6273
- VARCHAR(20) probably covers all countries

Hints on data modeling

- Single instance entities are OK
- Select names carefully
- Synonyms—different words have the same meaning
  - Get clients to settle on a common word or use views
- Homonyms—same word has different meanings
  - Clarify to avoid confusion
- Naming associative entities
  - Concatenate entity names if there is no obvious real world name

Hints on data modeling

- Uncover all exceptions
- Label relationships to avoid ambiguity
- Keep the data model well-formed and accurate

Meaningful identifiers

- An identifier is meaningful when some attributes of the entity can be inferred from the identifier's value

Advantages	Disadvantages
Recognizable and rememberable	Identifier exhaustion
Administrative simplicity	Reality changes
	Loss of meaningfulness

Recommendation

- Nothing, however, is lost and much is gained by using **non-meaningful** identifiers
- Non-meaningful identifiers serve their sole purpose well
  - To uniquely identify an entity
- Attributes are used to describe the characteristics of the entity
- A clear distinction between the role of identifiers and attributes creates fewer data management problems

The seven habits of highly effective data modelers

- Immerse
- Challenge
- Generalize
- Test
- Limit
- Integrate
- Complete

Key points

- A high-fidelity data model handles all exceptions
- Identifiers need identify only an instance
- Data modeling skills take time to develop