Data Management and Visualization

Dr. Ashish Kumar Jha

Session 2

Normalization and ER Diagrams

Agenda

Concepts or ER Modeling

How to use ER diagram to communicate about database

Diagrammatic technique for displaying an ER model

What is normalization

What are different normal forms

ER modeling

- Top-down approach to database design.
- Start by identifying the important data (called entities) and relationships between the data.
- Then add more details such as the information we want to hold about the entities and relationships (called attributes) and any constraints on the entities, relationships, and attributes.

Entities

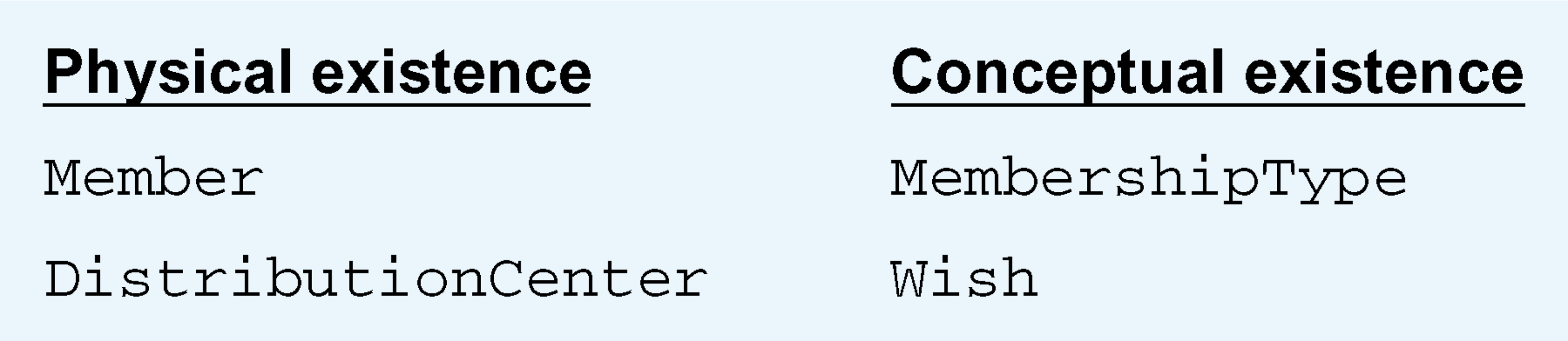
Entity

- A set of objects with the same properties, which are identified by a user or organization as having an independent existence.

Entity occurrence

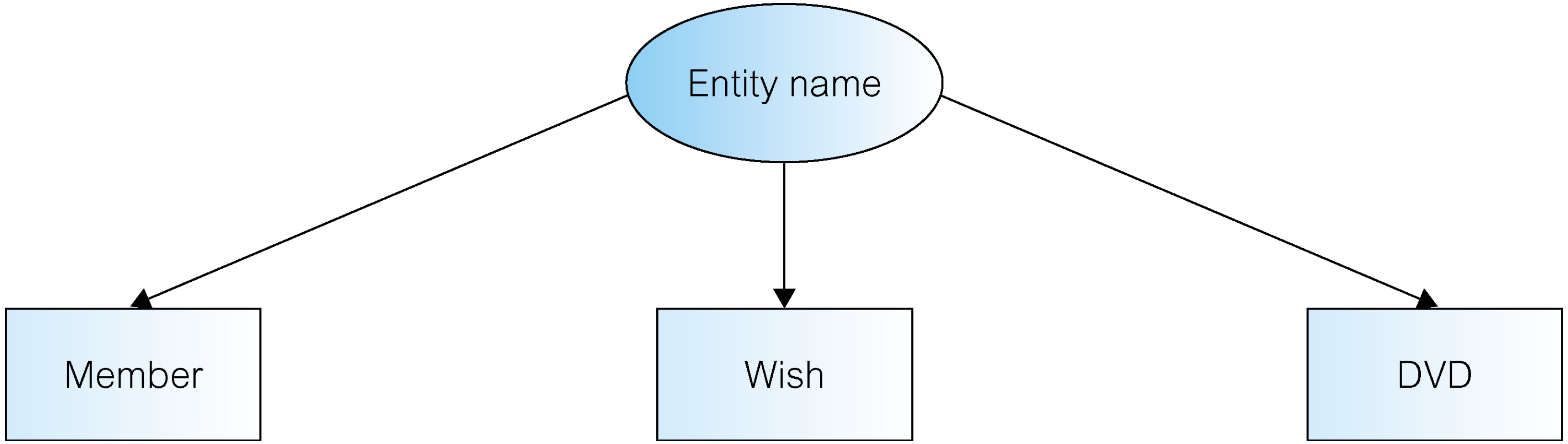
- Each uniquely identifiable object within a set.

Entities with physical and conceptual existence



Physical existence	Conceptual existence
`Member`	`MembershipType`
`DistributionCenter`	`Wish`

ER diagram of entities



Relationships

Relationship

- A set of meaningful associations among entities.

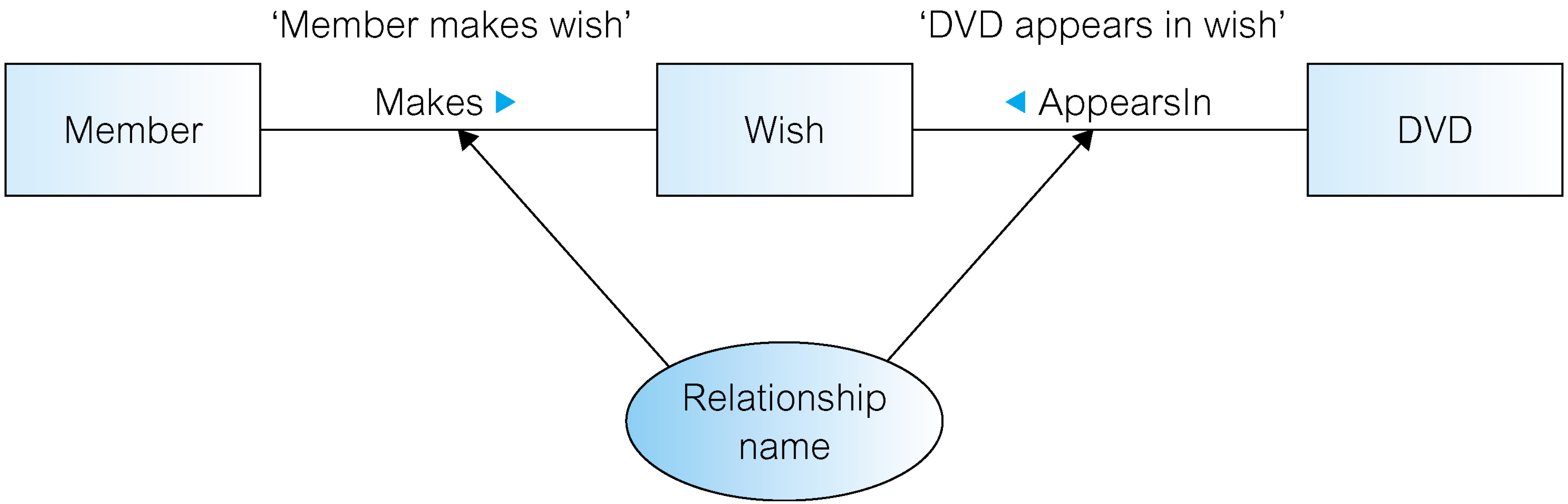
Relationship occurrence

- Each uniquely identifiable association within a set.

Degree of a relationship

- Number of participating entities in relationship.

ER diagram of relationships



Relationships

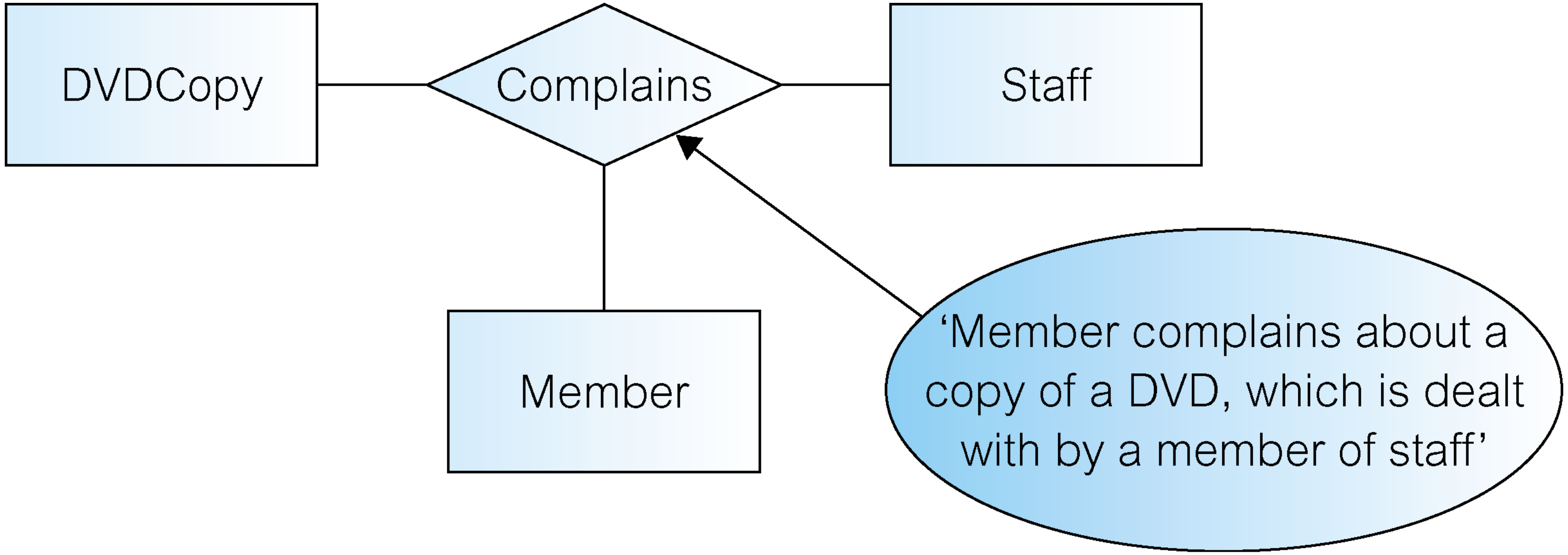
Relationship of degree :

- two is binary;
- three is ternary;
- four is quaternary.

Recursive relationships

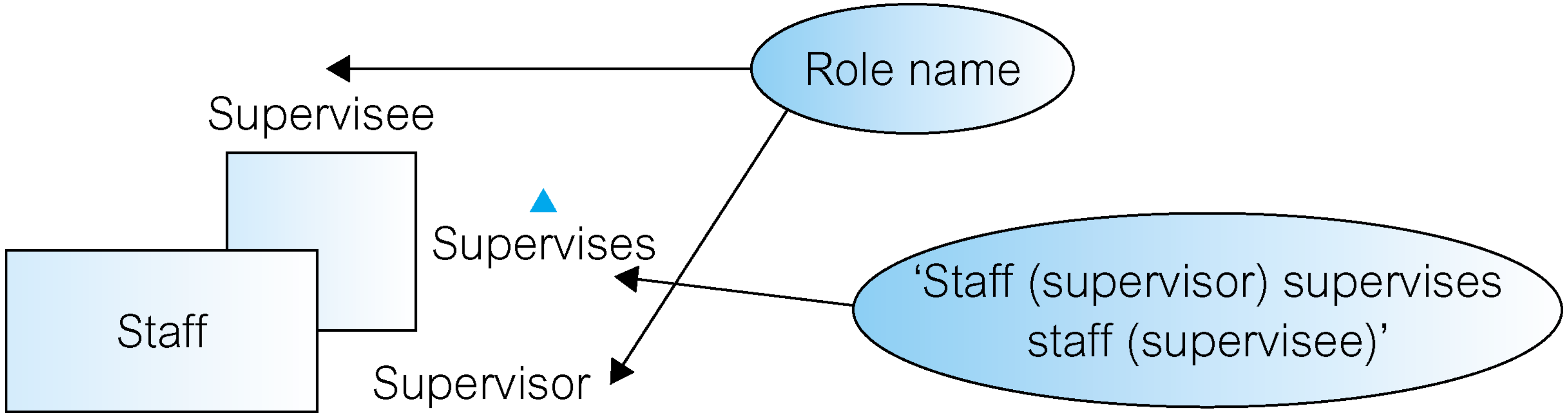
- Relationship where same entity participates more than once in different roles.
- Relationships may be given role names to indicate purpose that each participating entity plays in a relationship.

Example of ternary relationship



‘Member complains about a copy of a DVD, which is dealt with by a member of staff’

Example of a recursive relationship



Attributes

Attributes

- Property of an entity or a relationship.
- Hold values that describe each occurrence of an entity or relationship, and represent the main source of data stored in the database.

Attribute can be classified as being:

- simple or composite;
- single-valued or multi-valued;
- or derived.

Attributes

Simple attribute

- Attribute composed of a single component.

Composite attribute

- Attribute composed of multiple components.

Attributes

Single-valued attribute

- Attribute that holds a single value for an entity occurrence.

Multi-valued attribute

- Attribute that holds multiple values for an entity occurrence.

Derived attribute

- Attribute that represents a value that is derivable from value of a related attribute, or set of attributes, not necessarily in the same entity.

Keys: Recap

Superkey

- An attribute, or set of attributes, that uniquely identifies each entity occurrence.

Candidate key

- A superkey that contains only the minimum number of attributes necessary for unique identification of each entity occurrence.

Keys: recap

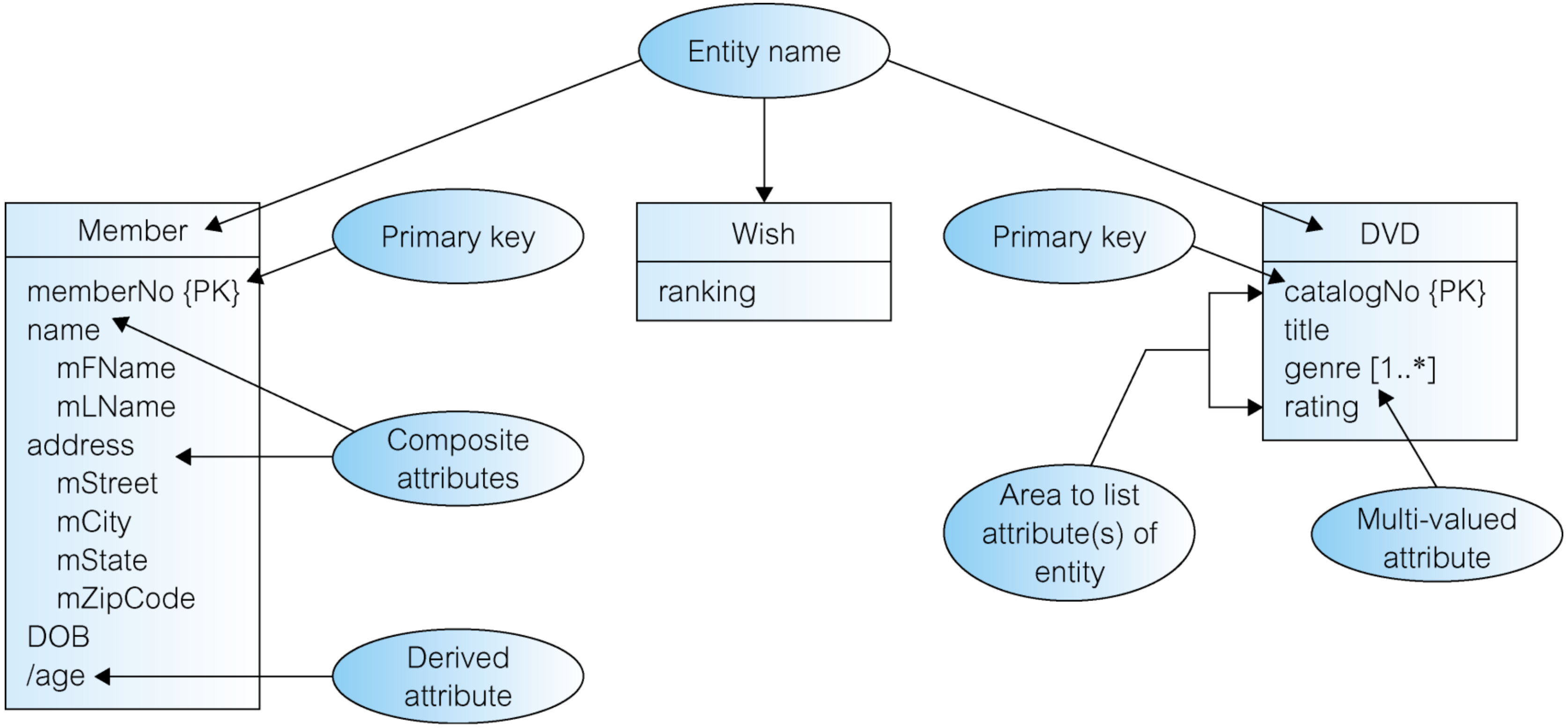
Primary key

- The candidate key that is selected to identify each entity occurrence.

Alternate key

- The candidate keys that are not selected as the primary key of the entity.

Diagrammatic representation of entities and attributes



More on Entities

Strong entity

- Entity that is not dependent on the existence of another entity for its primary key.

Weak entity

- Entity that is partially or wholly dependent on the existence of another entity, or entities, for its primary key.

Multiplicity constraints

Multiplicity constraints on relationships

– Represents the number of occurrences of one entity that may relate to a single occurrence of an associated entity.

- **Represents policies (called business rules) established by user or company.**

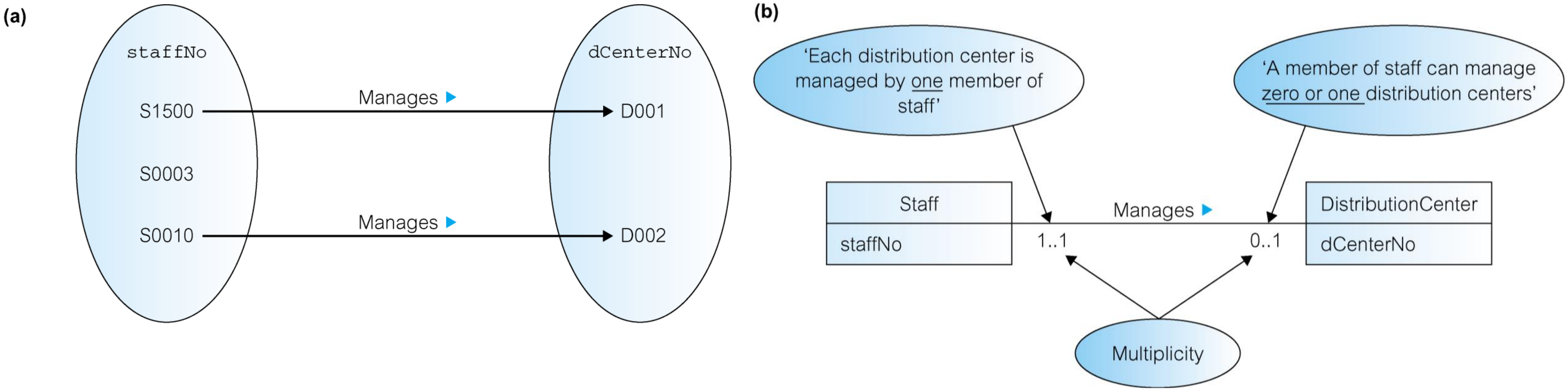
Multiplicity constraints

The most common degree for relationships is binary.

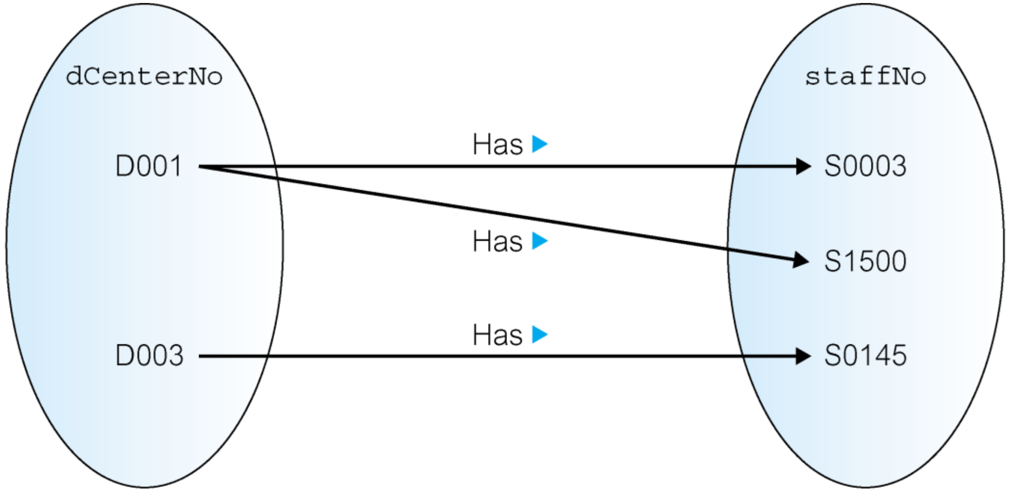
Binary relationships are generally referred to as being:

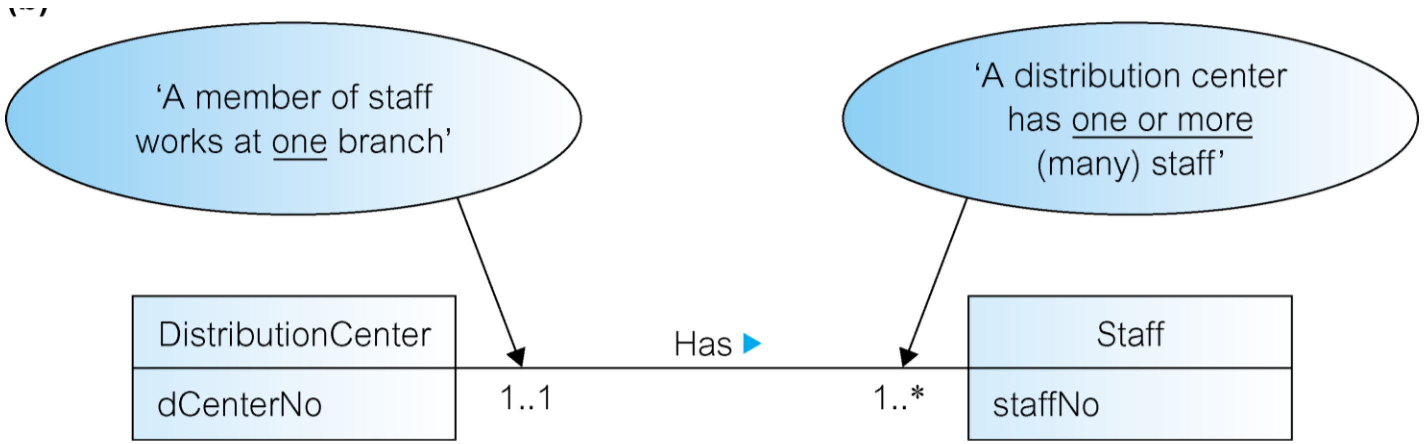
- one-to-one (1:1)
- one-to-many (1:*)
- many-to-many (*:*)

1:1 relationship (a) semantic net and (b) ER model

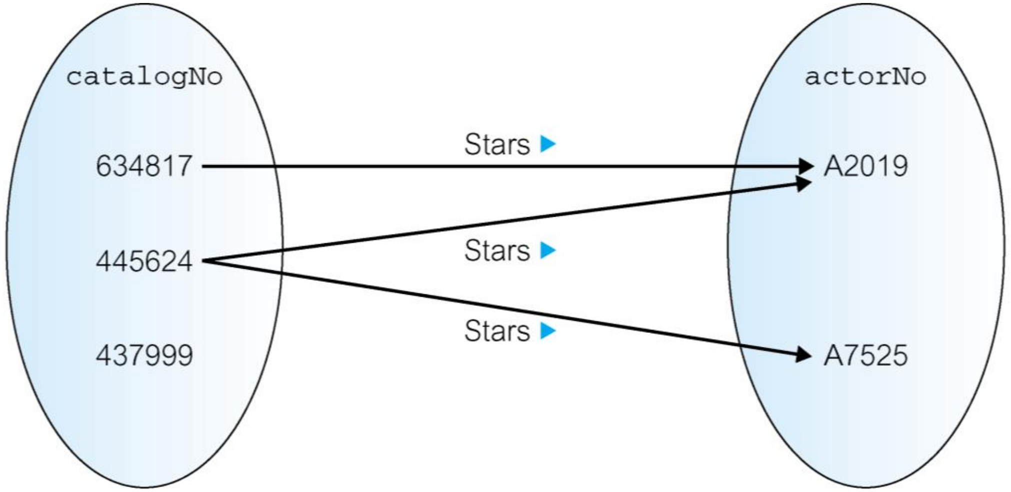


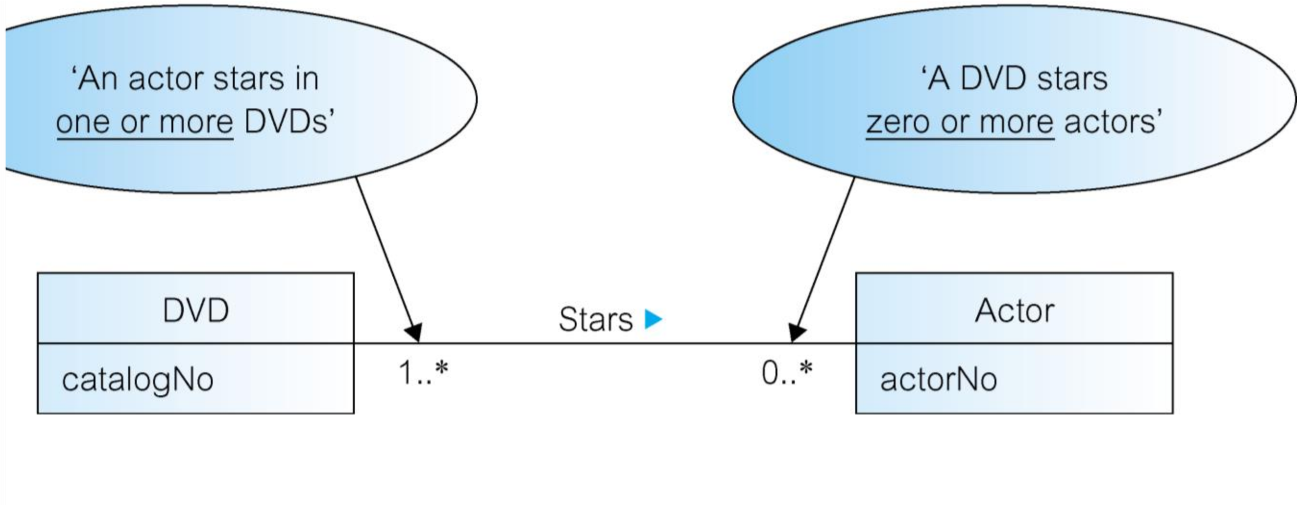
1:* relationship (a) semantic net and (b) ER model




'A member of staff works at one branch'
'A distribution center has one or more (many) staff'
DistributionCenter
dCenterNo
Has ▶
1..1
1..*
Staff
staffNo

: relationship (a) semantic net and (b) ER model

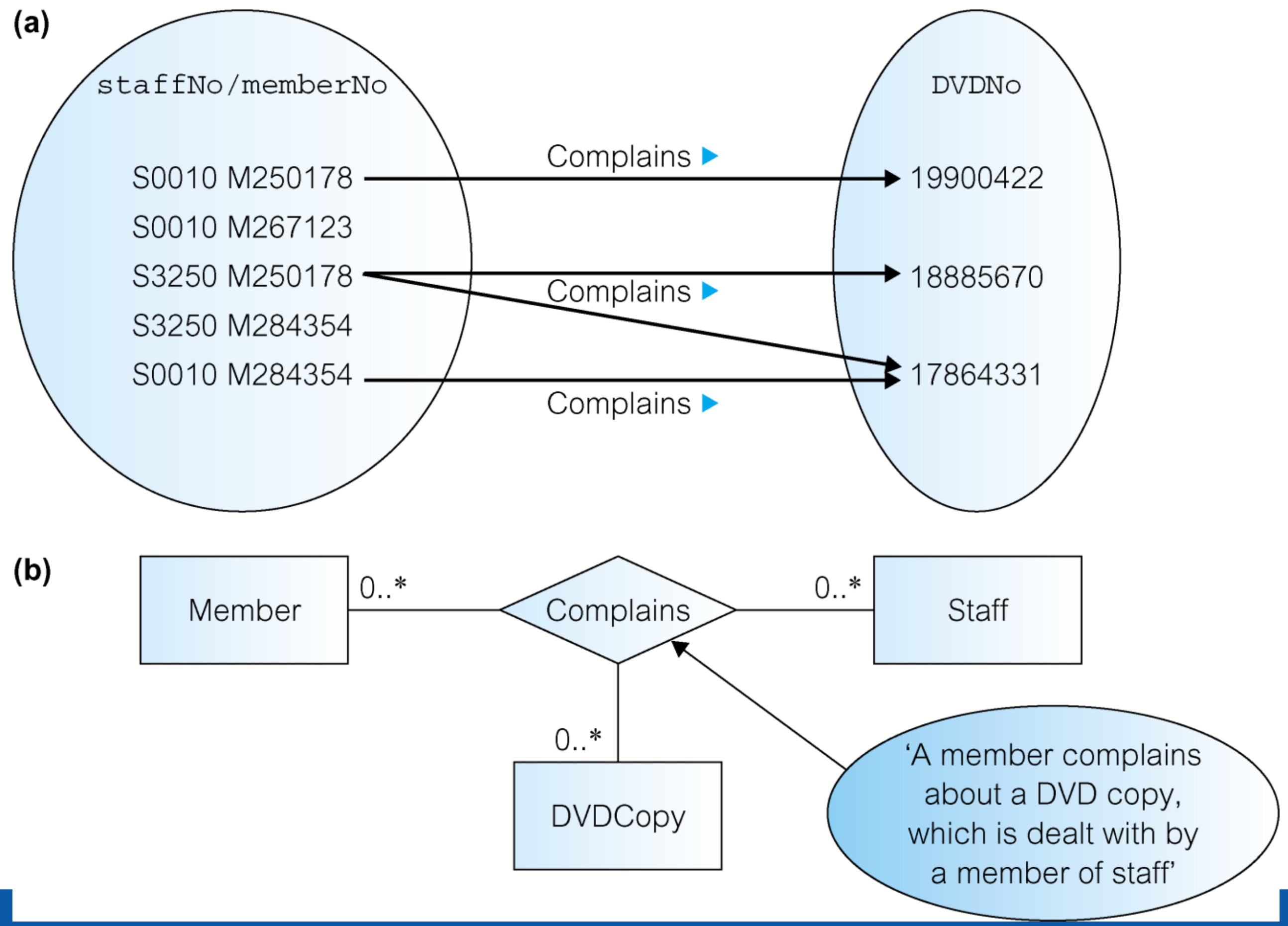




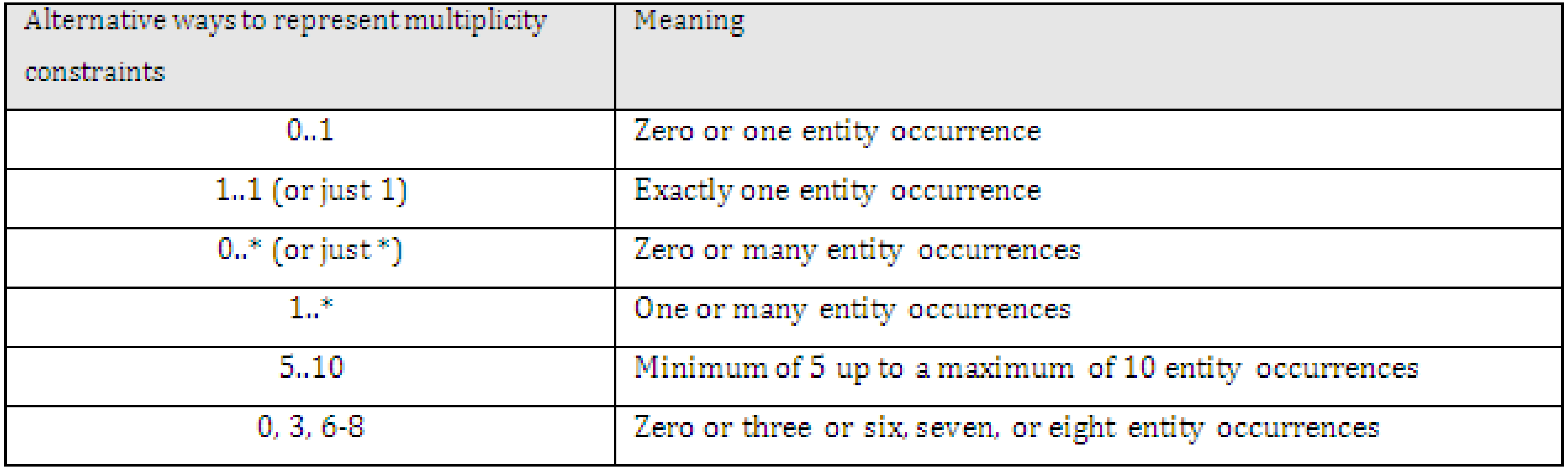
Complex relationships

Multiplicity is the number (or range) of possible occurrences of an entity type in an n-ary relationship when other (n-1) values are fixed.

Complex relationship (a) semantic net and (b) ER model



Summary of multiplicity constraints



Alternative ways to represent multiplicity constraints	Meaning
0..1	Zero or one entity occurrence
1..1 (or just 1)	Exactly one entity occurrence
0..* (or just *)	Zero or many entity occurrences
1..*	One or many entity occurrences
5..10	Minimum of 5 up to a maximum of 10 entity occurrences
0, 3, 6-8	Zero or three or six, seven, or eight entity occurrences

Multiplicity

Made up of two types of restrictions on relationships:

- cardinality,
- and participation.

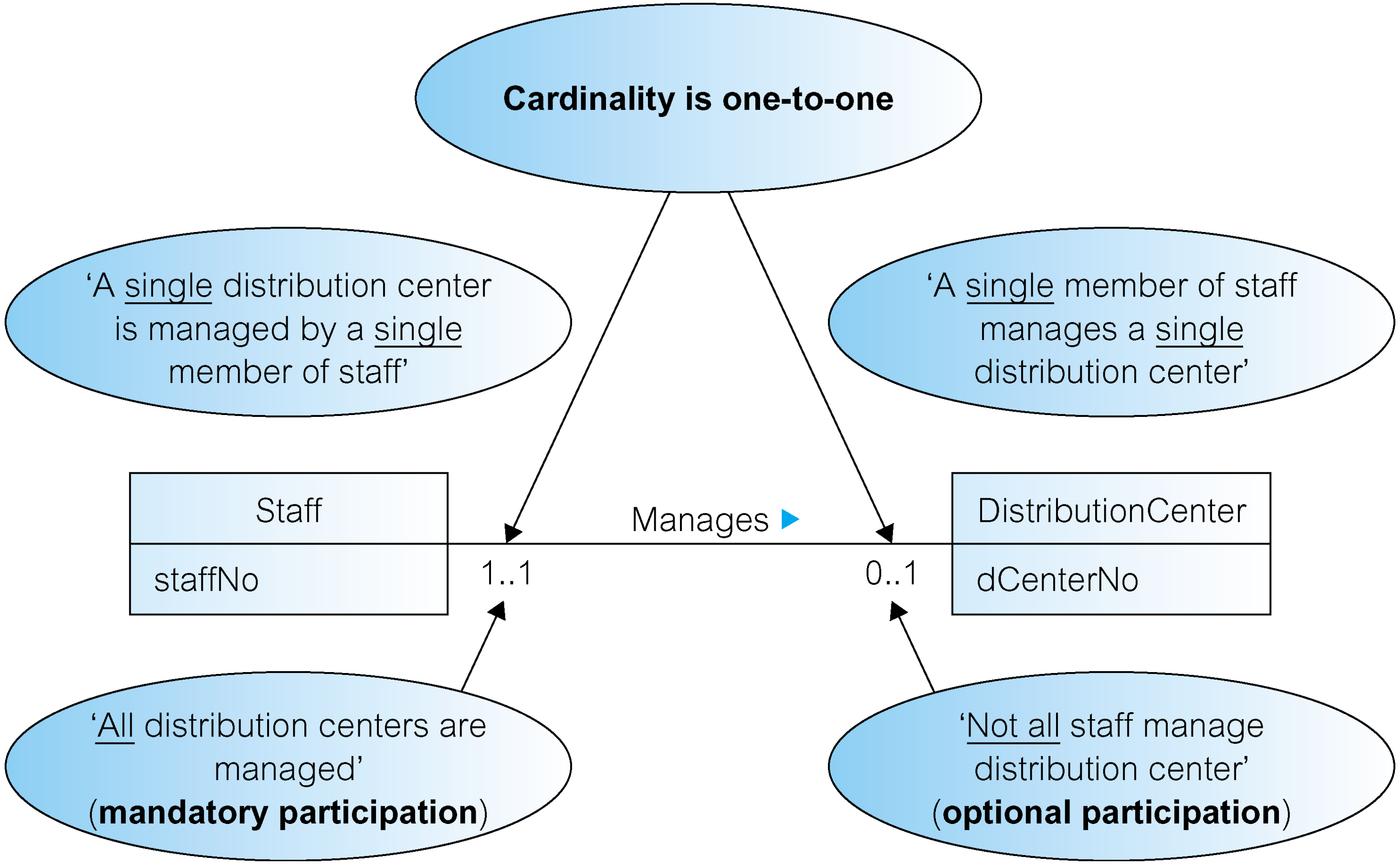
Cardinality

- Describes the number of possible relationships for each participating entity.

Participation

- Determines whether all or only some entity occurrences participate in a relationship.

Multiplicity as cardinality and participation constraints



Cardinality is one-to-one

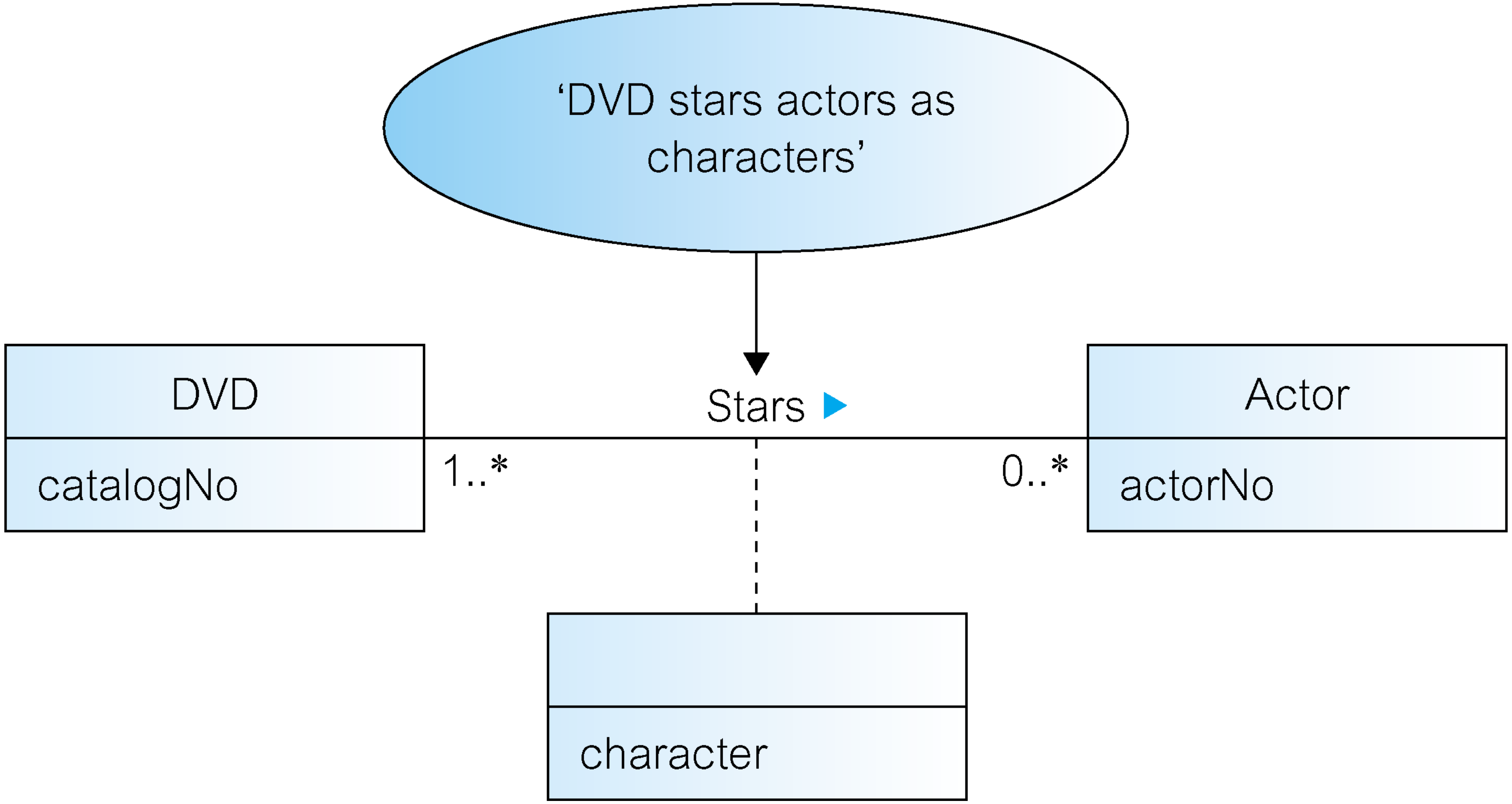
'A single distribution center is managed by a single member of staff'

'A single member of staff manages a single distribution center'

'All distribution centers are managed' (**mandatory participation**)

'Not all staff manage distribution center' (**optional participation**)

Relationship with attributes



Problems with ER models

Problems may arise when designing an ER model called connection traps.

Often due to a misinterpretation of the meaning of certain relationships.

Two main types of connection traps are called fan traps and chasm traps.

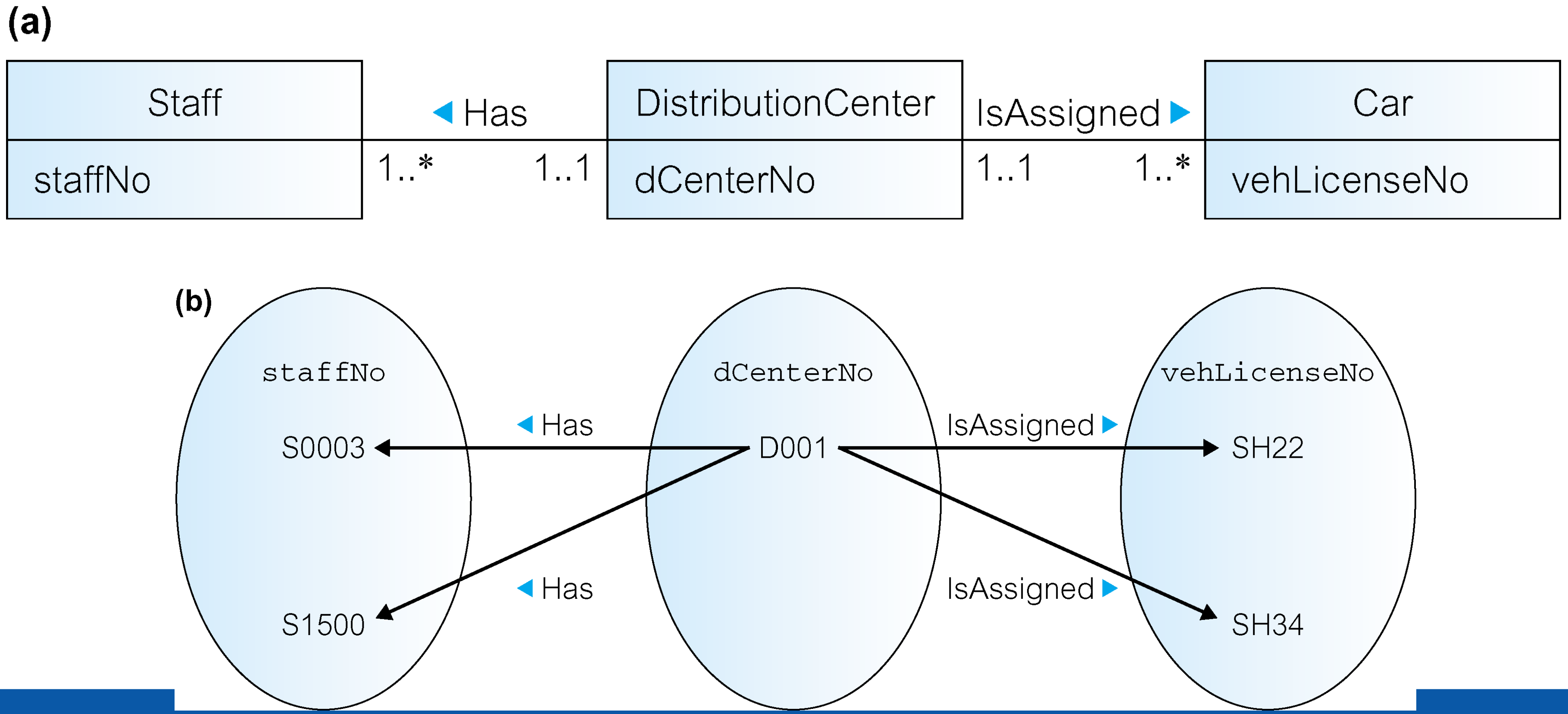
Problems with ER models

Fan trap

- Occurs between related entities that are not directly connected and the indirect pathway that connects them includes two 1:* relationships that fan out from a central entity.
- This means that certain entity occurrences that are related can only be connected using a pathway that can be ambiguous.

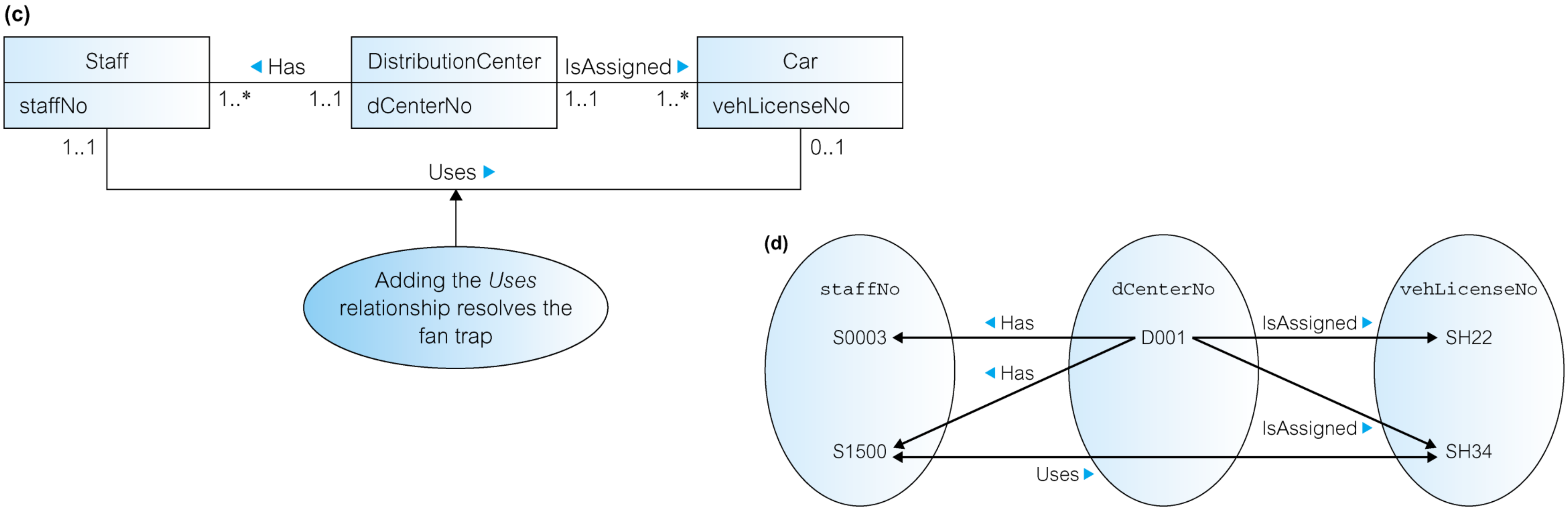
Example of a fan trap
(a) ER diagram (b) semantic net

Cannot tell which member of staff uses car SH34.



Fan trap resolved
(c) ER diagram (d) semantic net

Can now tell which car staff use.



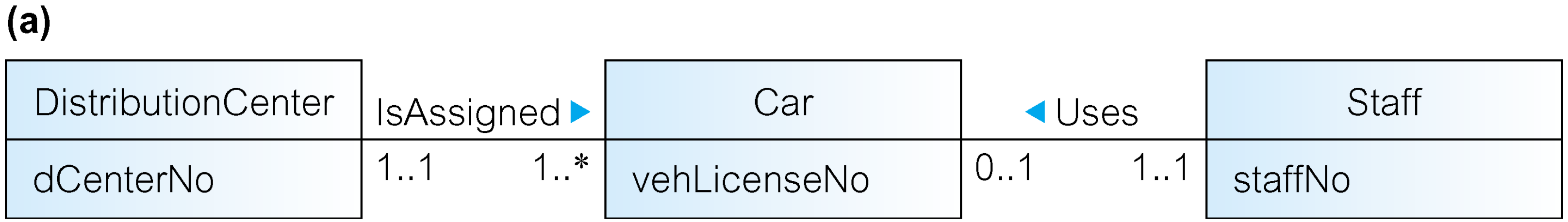
Problems with ER models

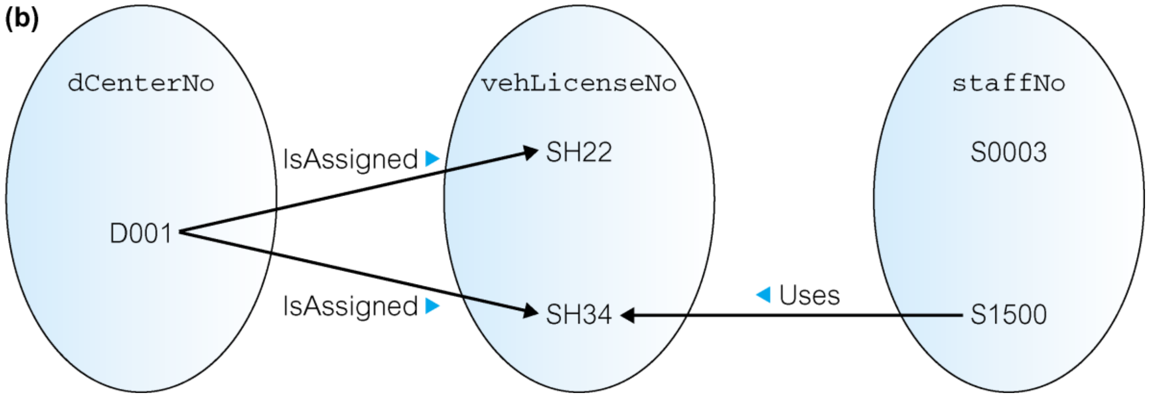
Chasm trap

- Occurs between related entities that are not directly connected and the indirect pathway that connects them includes partial participation.
- This means that certain entity occurrences that are related have no means of connection.

Example of a chasm trap (a) ER diagram (b) semantic net

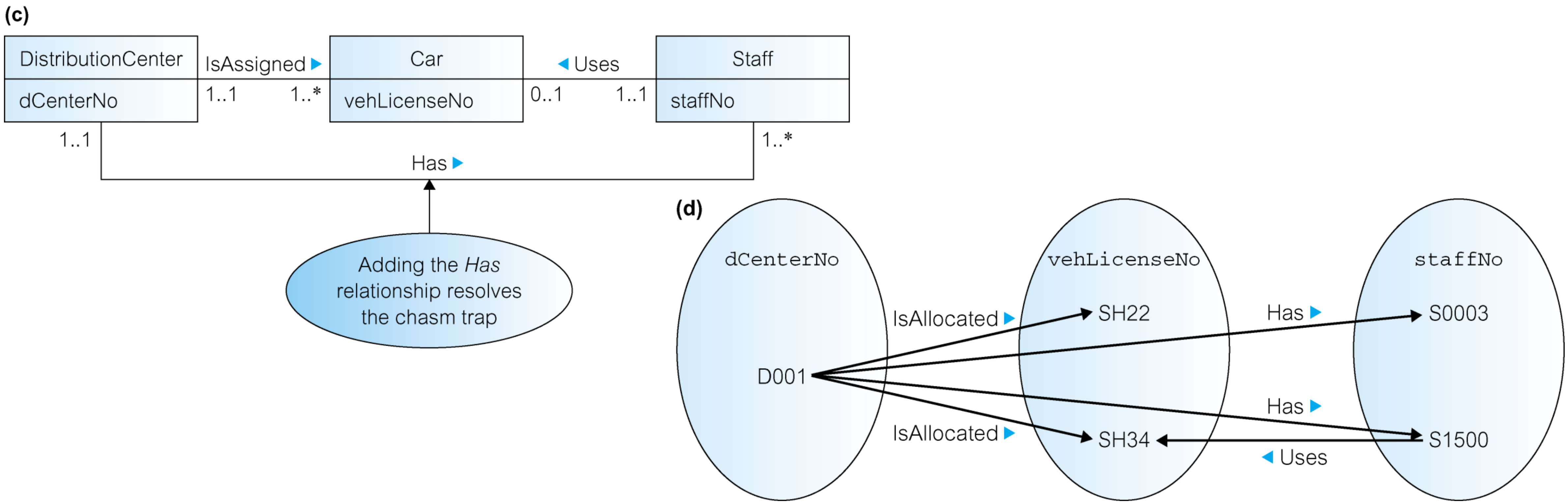
Cannot tell which distribution center has staff S0003 works at.





Chasm trap resolved
(c) ER diagram (d) semantic net

Can now tell where staff work



(c)

DistributionCenter

dCenterNo

IsAssigned ▶

1..1 1..*

Car

vehLicenseNo

◀ Uses

0..1 1..1

Staff

staffNo

1..1

1..*

Has ▶

Adding the *Has* relationship resolves the chasm trap

(d)

dCenterNo

D001

IsAllocated ▶

IsAllocated ▶

vehLicenseNo

SH22

SH34

Has ▶

Has ▶

◀ Uses

staffNo

S0003

S1500

Trinity Business School

Normalization

Normalization

A technique for producing a set of tables with minimal redundancy that support the data requirements of an organization.

Data redundancy and update anomalies

- Major aim of relational database design is to group columns into tables to minimize data redundancy and reduce file storage space required by implemented base tables.
- Problems associated with data redundancy are illustrated by comparing the Staff and Branch tables with the StaffBranch table.

Staff and DistributionCenter tables with StaffDistributionCenter table

Staff

staffNo	name	position	salary	dCenterNo
S1500	Tom Daniels	Manager	48000	D001
S0003	Sally Adams	Assistant	30000	D001
S0010	Mary Martinez	Manager	51000	D002
S3250	Robert Chin	Assistant	33000	D002
S2250	Sally Stern	Manager	48000	D004
S0415	Art Peters	Manager	42000	D003

DistributionCenter

dCenterNo	dAddress	dTelNo
D001	8 Jefferson Way, Portland, OR 97201	503-555-3618
D002	City Center Plaza, Seattle, WA 98122	206-555-6756
D003	14 – 8th Avenue, New York, NY 10012	212-371-3000
D004	2 W. El Camino, San Francisco, CA 94087	822-555-3131

StaffDistributionCenter

staffNo	name	position	salary	dCenterNo	dAddress	dTelNo
S1500	Tom Daniels	Manager	48000	D001	8 Jefferson Way, Portland, OR 97201	503-555-3618
S0003	Sally Adams	Assistant	30000	D001	8 Jefferson Way, Portland, OR 97201	503-555-3618
S0010	Mary Martinez	Manager	51000	D002	City Center Plaza, Seattle, WA 98122	206-555-6756
S3250	Robert Chin	Assistant	33000	D002	City Center Plaza, Seattle, WA 98122	206-555-6756
S2250	Sally Stern	Manager	48000	D004	2 W. El Camino, San Francisco, CA 94087	822-555-3131
S0415	Art Peters	Manager	42000	D003	14 – 8th Avenue, New York, NY 10012	212-371-3000

Data redundancy and update anomalies

- StaffDistributionCenter table has redundant data; the details of a distribution center are repeated for every member of staff.
- In contrast, the details of each distribution center appears only once for each centre in the DistributionCenter table and only the distribution center number (dCenterNo) is repeated in the Staff table, to represent where each member of staff is located.

Data redundancy and update anomalies

Tables that contain redundant information may potentially suffer from update anomalies.

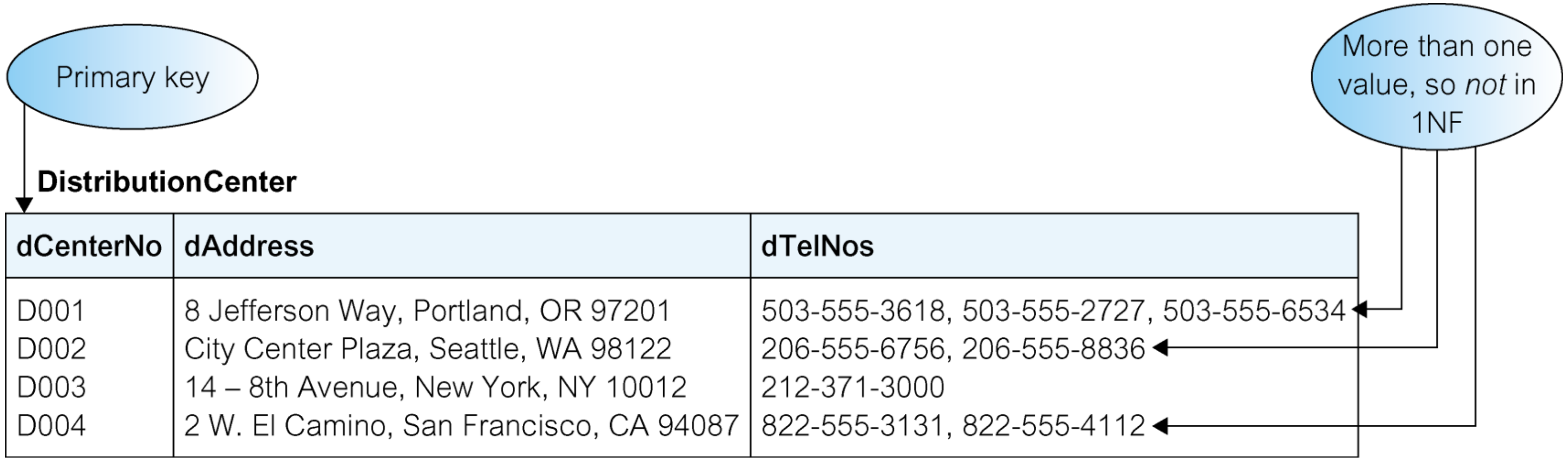
Types of update anomalies include:

- insertion,
- deletion,
- modification.

First normal form (1NF)

- Only 1NF is critical in creating appropriate tables for relational databases. All subsequent normal forms are optional.
- A table in which the intersection of every column and record contains only one value.

DistributionCenter table is not in 1NF



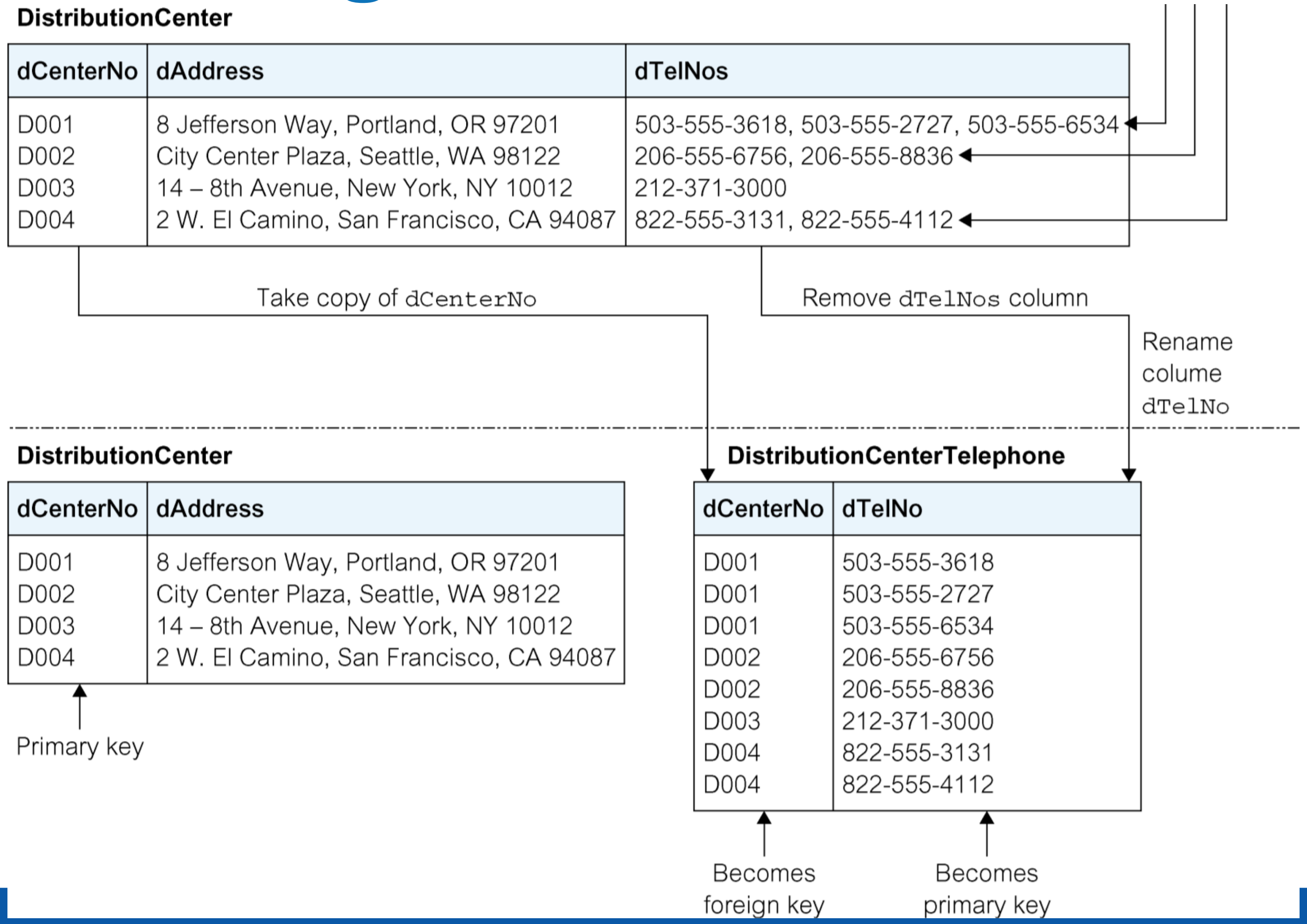
Primary key

More than one value, so *not* in 1NF

DistributionCenter

dCenterNo	dAddress	dTelNos
D001	8 Jefferson Way, Portland, OR 97201	503-555-3618, 503-555-2727, 503-555-6534
D002	City Center Plaza, Seattle, WA 98122	206-555-6756, 206-555-8836
D003	14 – 8th Avenue, New York, NY 10012	212-371-3000
D004	2 W. El Camino, San Francisco, CA 94087	822-555-3131, 822-555-4112

Converting DistributionCenter table to 1NF



DistributionCenter

dCenterNo	dAddress	dTelNos
D001	8 Jefferson Way, Portland, OR 97201	503-555-3618, 503-555-2727, 503-555-6534
D002	City Center Plaza, Seattle, WA 98122	206-555-6756, 206-555-8836
D003	14 – 8th Avenue, New York, NY 10012	212-371-3000
D004	2 W. El Camino, San Francisco, CA 94087	822-555-3131, 822-555-4112

DistributionCenter

dCenterNo	dAddress
D001	8 Jefferson Way, Portland, OR 97201
D002	City Center Plaza, Seattle, WA 98122
D003	14 – 8th Avenue, New York, NY 10012
D004	2 W. El Camino, San Francisco, CA 94087

DistributionCenterTelephone

dCenterNo	dTelNo
D001	503-555-3618
D001	503-555-2727
D001	503-555-6534
D002	206-555-6756
D002	206-555-8836
D003	212-371-3000
D004	822-555-3131
D004	822-555-4112

Second normal form (2NF)

- A table that is in 1NF and in which the values of each non-primary-key column are determined by the values in all the columns that make up the primary key.
- To assess whether a table breaks 2NF form requires identification of the primary key and functional dependencies associated with that table.
- 2NF only applies to tables with composite primary keys.

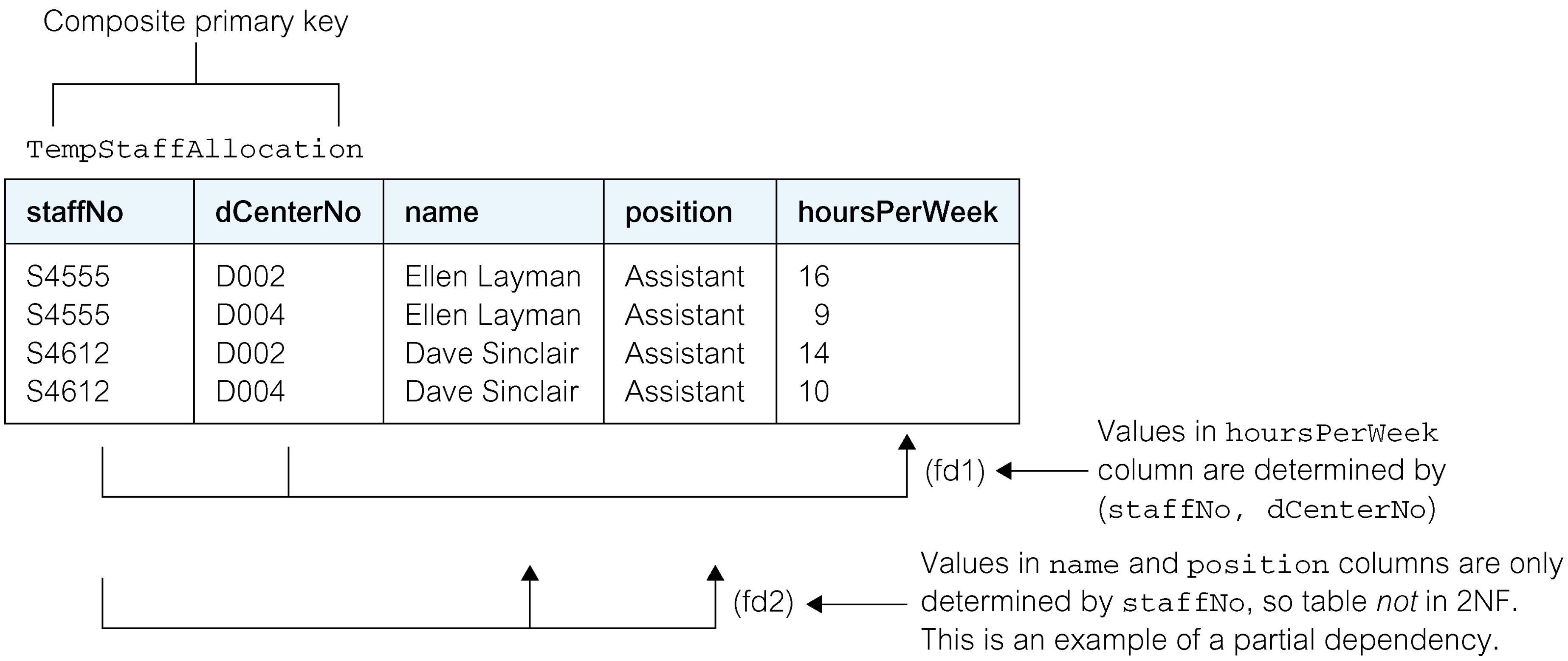
Functional dependency

- Describes the relationship between columns in a table and indicates how columns relate to one another.
- For example, consider a table with columns a and b, where b is functionally dependent on a (denoted a → b). If we know the value of a, we find only one value of b in all the records that has this value of a, at any moment in time. However, for a given value of b there may be several different values of a.

Second normal form (2NF)

- Formal definition of 2NF is a table that is in 1NF and every non-primary-key column is fully functional dependent on the primary key.
- Full functional dependency indicates that if a and b are columns of a table, b is fully determined by a, if b is not determined by any subset of a. If b is determined by a subset of a, this is referred to as a partial dependency.
- Identification of partial dependencies on the primary key is evidence that a table is breaking 2NF and may suffer from update anomalies.

TempStaffAllocation table is <u>not</u> in 2NF



Composite primary key

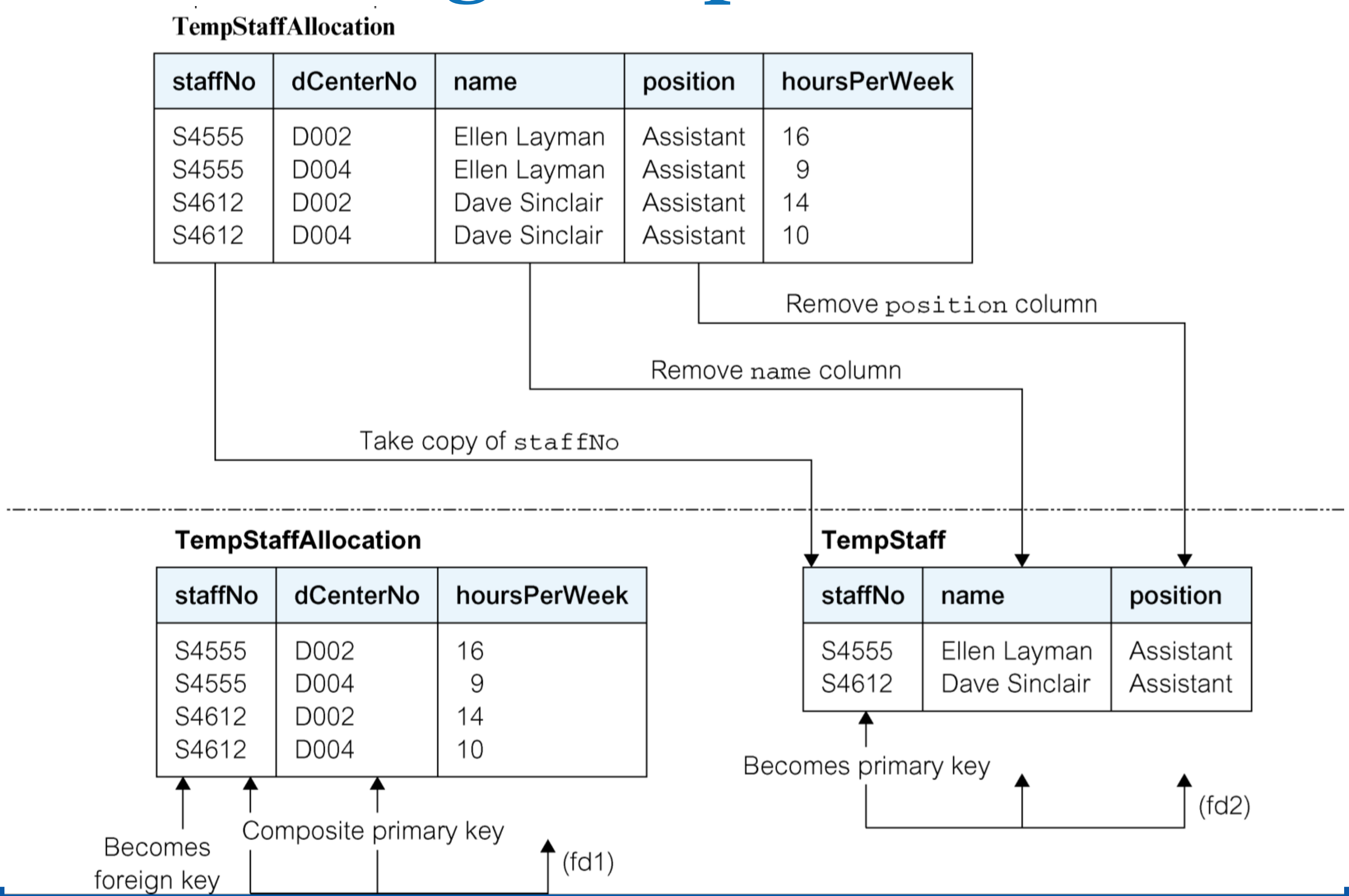
`TempStaffAllocation`

staffNo	dCenterNo	name	position	hoursPerWeek
S4555	D002	Ellen Layman	Assistant	16
S4555	D004	Ellen Layman	Assistant	9
S4612	D002	Dave Sinclair	Assistant	14
S4612	D004	Dave Sinclair	Assistant	10

(fd1) ← Values in `hoursPerWeek` column are determined by (`staffNo, dCenterNo`)

(fd2) ← Values in `name` and `position` columns are only determined by `staffNo`, so table *not* in 2NF. This is an example of a partial dependency.

Converting TempStaffAllocation table to 2NF



TempStaffAllocation

staffNo	dCenterNo	name	position	hoursPerWeek
S4555	D002	Ellen Layman	Assistant	16
S4555	D004	Ellen Layman	Assistant	9
S4612	D002	Dave Sinclair	Assistant	14
S4612	D004	Dave Sinclair	Assistant	10

Remove position column

Remove name column

Take copy of staffNo

TempStaffAllocation

staffNo	dCenterNo	hoursPerWeek
S4555	D002	16
S4555	D004	9
S4612	D002	14
S4612	D004	10

Becomes foreign key

Composite primary key

(fd1)

TempStaff

staffNo	name	position
S4555	Ellen Layman	Assistant
S4612	Dave Sinclair	Assistant

Becomes primary key

(fd2)

Third normal form (3NF)

- A table that is in 1NF and 2NF and in which the values in all non-primary-key column can be determined from only the primary key column(s) and no other columns.
- The formal definition of 3NF is a table that is in 1NF and 2NF and in which no non-primary-key column is transitively dependent on the primary key.

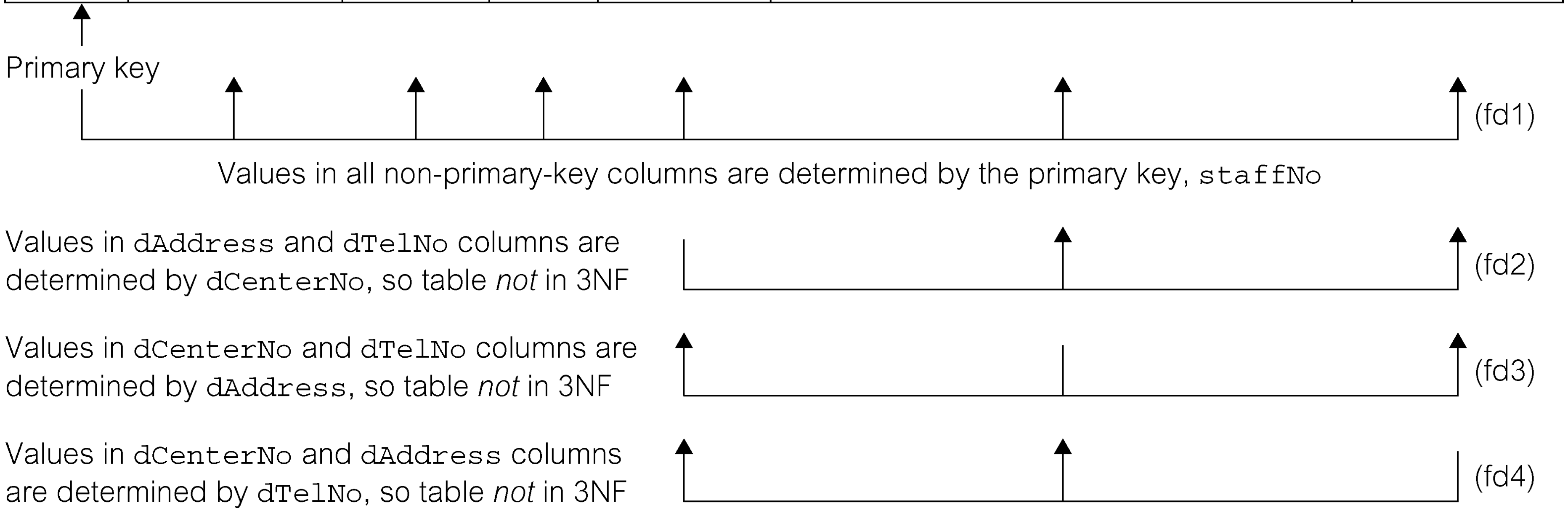
Third normal form (3NF)

- A transitive dependency describes a relationship between columns a, b, and c. If a determines b (a → b) and b determines c (b → c), then c is transitively dependent on a via b (provided that b or c does not determine a).
- Identification of transitive dependencies on the primary key is evidence that a table is breaking 3NF and may suffer from update anomalies.

The StaffDistributionCenter table is <u>not </u>in 3NF

StaffDistributionCenter

staffNo	name	position	salary	dCenterNo	dAddress	dTelNo
S1500	Tom Daniels	Manager	48000	D001	8 Jefferson Way, Portland, OR 97201	503-555-3618
S0003	Sally Adams	Assistant	30000	D001	8 Jefferson Way, Portland, OR 97201	503-555-3618
S0010	Mary Martinez	Manager	51000	D002	City Center Plaza, Seattle, WA 98122	206-555-6756
S3250	Robert Chin	Assistant	33000	D002	City Center Plaza, Seattle, WA 98122	206-555-6756
S2250	Sally Stern	Manager	48000	D004	2 W. El Camino, San Francisco, CA 94087	822-555-3131
S0415	Art Peters	Manager	42000	D003	14 – 8th Avenue, New York, NY 10012	212-371-3000



Primary key

(fd1)

Values in all non-primary-key columns are determined by the primary key, `staffNo`

Values in `dAddress` and `dTelNo` columns are determined by `dCenterNo`, so table *not* in 3NF

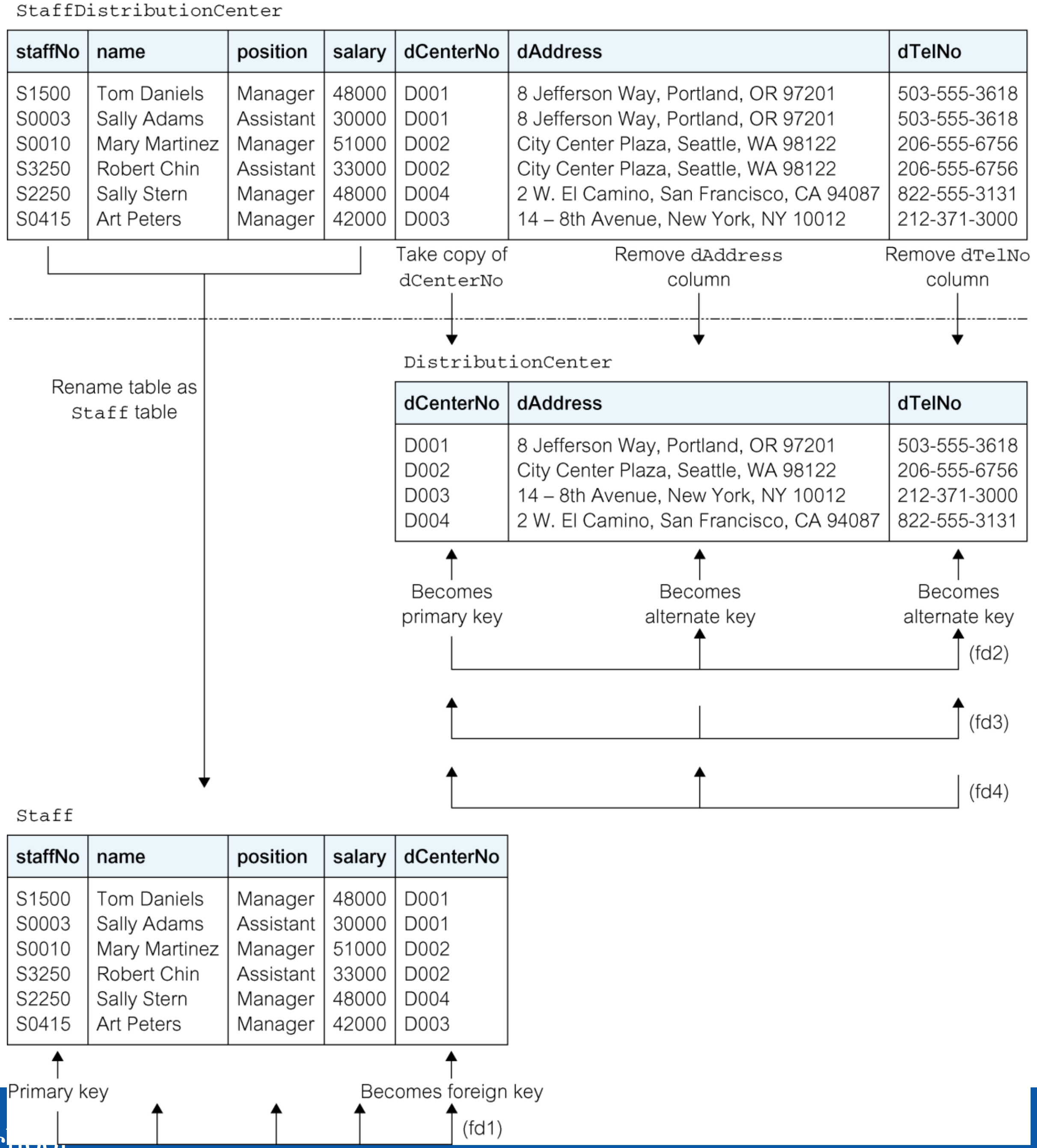
(fd2)

Values in `dCenterNo` and `dTelNo` columns are determined by `dAddress`, so table *not* in 3NF

(fd3)

Values in `dCenterNo` and `dAddress` columns are determined by `dTelNo`, so table *not* in 3NF

(fd4)



StaffDistributionCenter

staffNo	name	position	salary	dCenterNo	dAddress	dTelNo
S1500	Tom Daniels	Manager	48000	D001	8 Jefferson Way, Portland, OR 97201	503-555-3618
S0003	Sally Adams	Assistant	30000	D001	8 Jefferson Way, Portland, OR 97201	503-555-3618
S0010	Mary Martinez	Manager	51000	D002	City Center Plaza, Seattle, WA 98122	206-555-6756
S3250	Robert Chin	Assistant	33000	D002	City Center Plaza, Seattle, WA 98122	206-555-6756
S2250	Sally Stern	Manager	48000	D004	2 W. El Camino, San Francisco, CA 94087	822-555-3131
S0415	Art Peters	Manager	42000	D003	14 – 8th Avenue, New York, NY 10012	212-371-3000

Take copy of dCenterNo

Remove dAddress column

Remove dTelNo column

Rename table as Staff table

DistributionCenter

dCenterNo	dAddress	dTelNo
D001	8 Jefferson Way, Portland, OR 97201	503-555-3618
D002	City Center Plaza, Seattle, WA 98122	206-555-6756
D003	14 – 8th Avenue, New York, NY 10012	212-371-3000
D004	2 W. El Camino, San Francisco, CA 94087	822-555-3131

Becomes primary key

Becomes alternate key

Becomes alternate key

(fd2)

(fd3)

(fd4)

Staff

staffNo	name	position	salary	dCenterNo
S1500	Tom Daniels	Manager	48000	D001
S0003	Sally Adams	Assistant	30000	D001
S0010	Mary Martinez	Manager	51000	D002
S3250	Robert Chin	Assistant	33000	D002
S2250	Sally Stern	Manager	48000	D004
S0415	Art Peters	Manager	42000	D003

Primary key

Becomes foreign key

(fd1)