Normalization and Other Data Modeling Methods

There are many paths to the top of the mountain but the view is always the same

Chinese proverb

Normalization

- An alternative database design tool to data modeling
- A theoretical foundation for the relational model
- Application of a series of rules that gradually improve the design

Functional dependency

- A relationship between attributes in an entity
  - One or more attributes determine the value of another attribute
- An identifier functionally determines all the attributes of an entity
  - stock code ->firm name, stock price, stock quantity, stock dividend
  - If we know stock code we know the value of firm name, etc.
- Multivalued dependency
  - Formulae
    - (stock dividend, stock price) –> yield

Full functional dependency

- Yield is fully functionally dependent on stock dividend and stock price because both of these attributes are required to determine the value of yield
  - (stock dividend, stock price) –> yield
- Determinant
  - An attribute that fully functionally determines another attribute
    - e.g., stock code determines stock PE

Multidetermination

- A given value can determine multiple values
  - A multidetermines B
  - A –> –> B
  - e.g., Department multidetermines course
- Multivalued dependency means functional dependencies are multivalued

Attribute relationships

- One-to-one
  - A value of an attribute determines the value of another attribute and vice versa
  - A –> B and B –> A
  - e.g.,
    - CH –> Switzerland
    - Switzerland –> CH

Attribute relationships

- One-to-many
  - A value of one attribute determines the value of another attribute but **not** vice versa
    - country name –> currency unit
    - currency unit not –> country name

Attribute relationships

- Many-to-many
  - Neither attribute determines the other
  - A not –> B
  - B not –> A
    - country name not –> language
    - language not –> country name
      - French and Flemish is spoken in Belgium
      - French is spoken in many countries

Normal forms

- A classification of relations
- Stacked like a set of Russian dolls
  - Innermost is first normal form

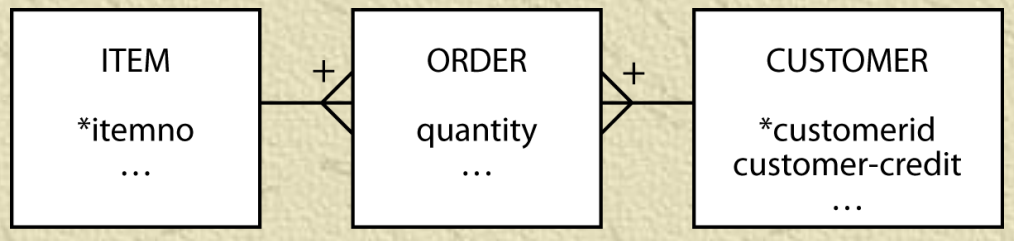
First normal form (1NF)

- All rows must have the same number of columns
- Single valued attributes only

Second normal form (2NF)

- Violated when a nonkey column is a fact about part of the primary key
- A column is not fully functionally dependent on the primary key
  - `customer-credit` in this case

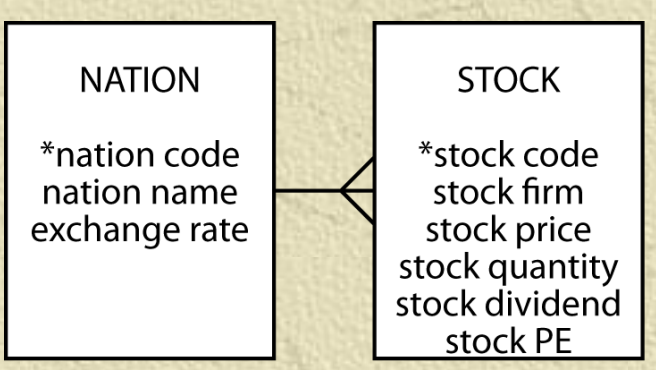
order			
itemno	customerid	quantity	customer-credit
12	57	25	OK
34	679	3	POOR



Third normal form (3NF)

- Violated when a nonkey column is a fact about another nonkey column
- A column is not fully functionally dependent on the primary key
  - `exchange rate` in this case

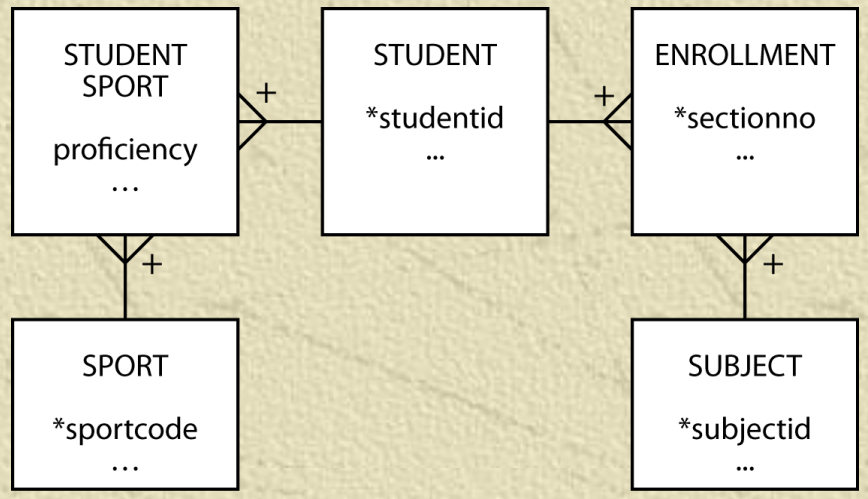
stock		
stock code	nation	exchange rate
MG	USA	0.67
IR	AUS	0.46



Fourth normal form (4NF)

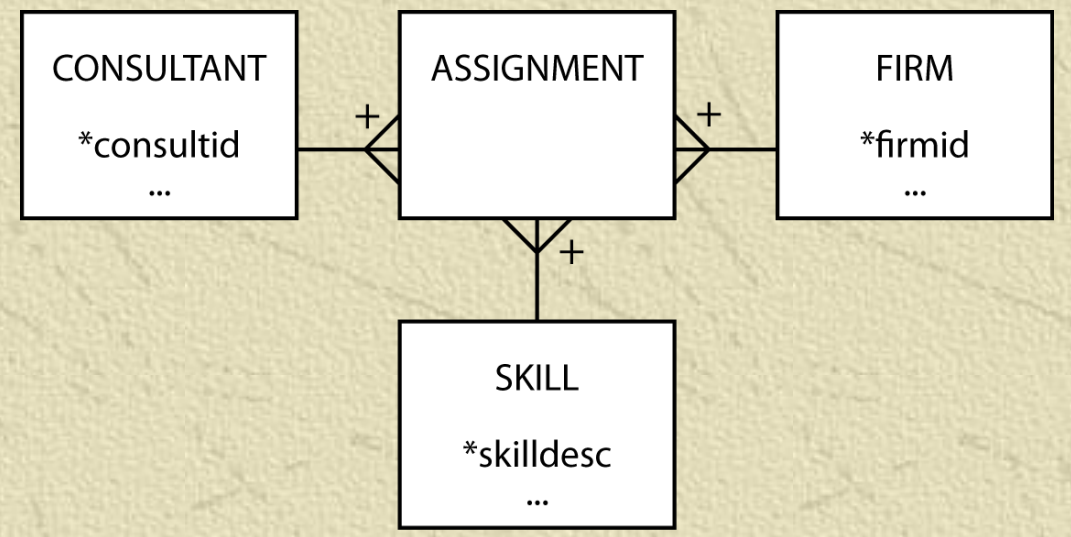
- A row should not contain two or more multivalued independent facts

student			
studentid	sport	subject	...
50	Football	English	...
50	Football	Music	...
50	Tennis	Botany	...
50	Karate	Botany	...



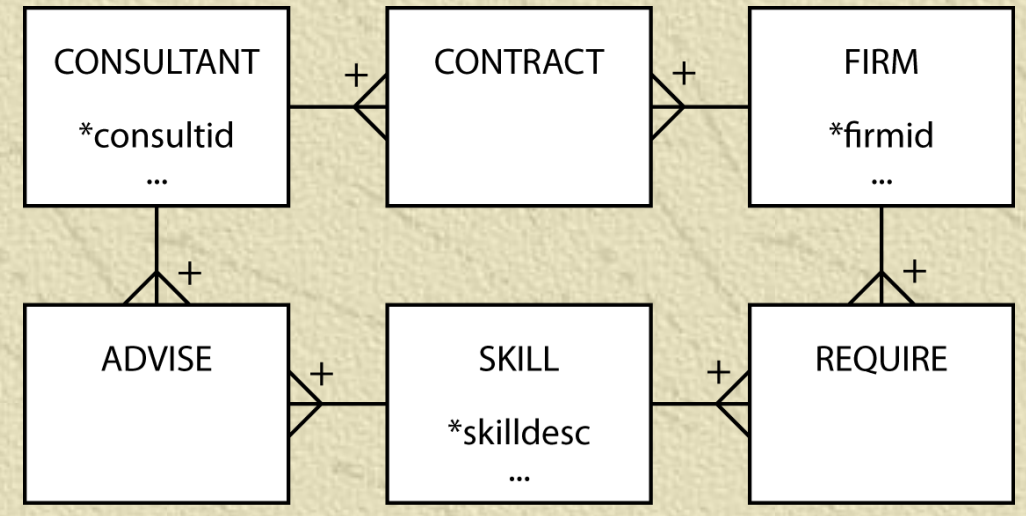
Fifth normal form (5NF)

- A table can be reconstructed from other tables
- There exists some rule that enables a relation to be inferred
- Base case
  - Consultants provide skills to one or more firms and firms can use many consultants; a consultant has many skills and a skill can be used by many firms; and a firm can have a need for many skills and the same skill can be required by many firms



Fifth normal form (5NF)

- The rule
  - If a consultant has a certain skill (e.g., database) and has a contract with the firm that requires that skill (e.g., IBM), then the consultant advises the firm on that skill (i.e., he advises IBM on database)



Data modeling and normalization

- Data modeling is often an easier path to good database design
- A high-fidelity data model will be of high normal form
- 5NF is likely to create the most problems
  - Check for special rules

Goal

- Learn to think like a data modeler
- Different dialects and greater precision (e.g., cardinality) come easily once the basics are mastered

Key points

- Normalization is one approach to data modeling
- The are multiple representations for data model
- Learning to model is difficult
- Learning to represent a model is easy