Relational Database Management for Epidemiologists: Normalization

Wayne Enanoria, PhD, MPH
Center for Infectious Disease Preparedness
University of California at Berkeley

Phases of Database Design

- **Define Mission Statement and Objectives**
- **Analyze the current database**
- **Create data structures**
- **Establish table relationships**
- Define business rules
- Determine and establish views
- Review data integrity

From Last Time…

- Reviewed the ERD with key personnel
- Developed a list of fields and tables
- Developed preliminary table relationships

Outline

- Normalization
  - 1NF
  - 2NF
  - 3NF
  - 4NF
  - 5NF

What is normalization?

- Normalization is a process in which a given set of relations is replaced by successive collections of relations that have a simpler and more regular structure.
- Each set, referred to as a *normal form*, defines a set of criteria that needs to be met by the different tables in the database.
- Rules of normalization eliminate redundancy and inconsistent dependency in table designs.

Objectives of Normalization

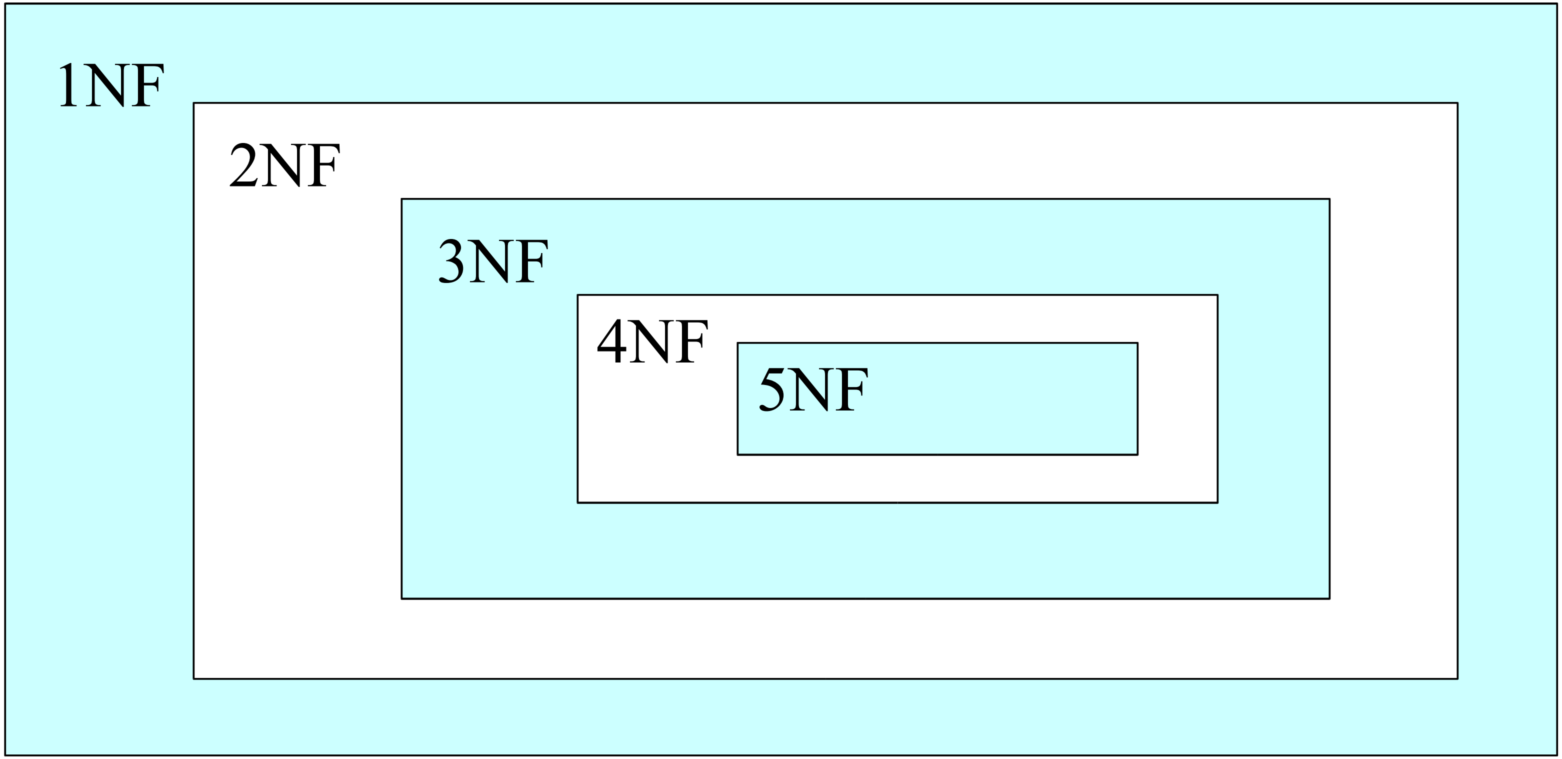
- The objectives of normalization process are*:
  - To make it feasible to represent any relation in the database.
  - To free relations from undesirable insertion, update, and deletion anomalies.
  - To reduce the need for restructuring the relations as new data types are introduced.

*Adapted from *Database Management Systems* by D. Tsichritzis and F. Lochovsky, Academic Press, 1977, and *Schaum's Outlines. Fundamentals of Relational Databases* by R.A. Mata-Toledo and P.K. Cushman, McGraw-Hill, 2000.

The Process of Normalization

- The process is based on the analysis of relations, their schemes, their primary keys and their functional dependencies.
- Whenever a relation does not meet a normal form test, the relation must be decomposed or broken down into some other relations that individually meet the criteria of the normal form test.

The Normal Form “Onion”



1NF

2NF

3NF

4NF

5NF

First Normal Form

- A table is said to be in *First Normal Form (1NF)* if and only if every entry of the table (the intersection of row and column) has at most a single value.
- Objective: to remove a table's repeating groups and ensure that all entries of the resulting table have at most a single value.

Eliminate duplicate data!

CASE Table

CaseID	CaseFname	CaseLname	ControlID	ControlFname	ControlLname	ControlAge	Relationship
101	John	Smith	1001	Fred	Smith	5	Father-Son
			1002	Larry	Smith	10	Father-Son
			1003	John	Smith, Jr.	2	Father-Son
			1004	Margaret	Smith	32	Husband-Wife
102	Maria	Sanchez	1005	Javier	Sanchez	1	Mother-Son
			1006	Izel	Sanchez	1	Mother-Daughter
			1007	Juan	Sanchez	44	Wife-Husband
103	Hilary	Connor	1008	Fred	Connor	25	Wife-Husband
			1009	Jackie	Connor	2	Mother-Daughter

“Flattening the Table”

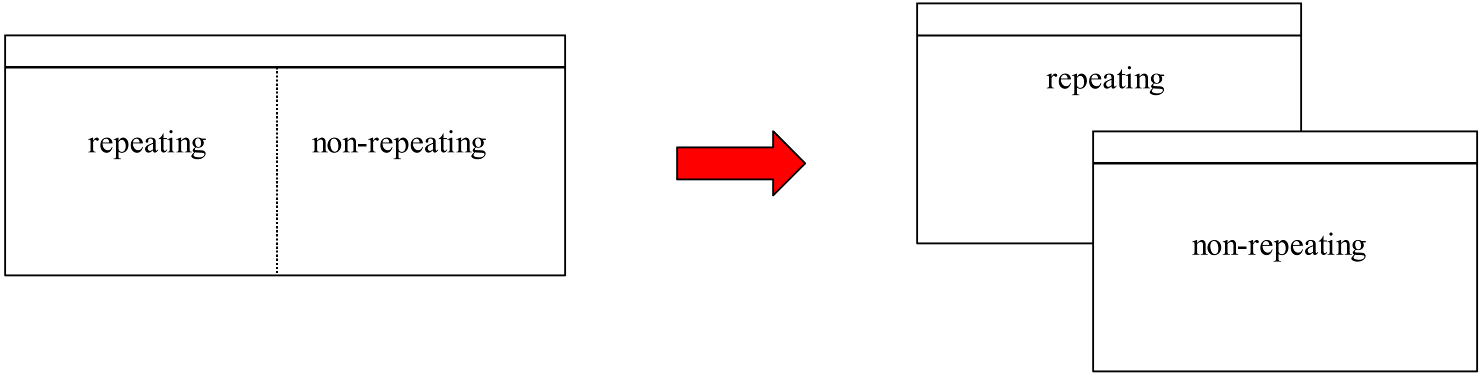
CaseID	CaseFname	CaseLname	ControlID	ControlFname	ControlLname	ControlAge	Relationship
101	John	Smith	1001	Fred	Smith	5	Father-Son
101	John	Smith	1002	Larry	Smith	10	Father-Son
101	John	Smith	1003	John	Smith, Jr.	2	Father-Son
101	John	Smith	1004	Margaret	Smith	32	Husband-Wife
102	Maria	Sanchez	1005	Javier	Sanchez	1	Mother-Son
102	Maria	Sanchez	1006	Izel	Sanchez	1	Mother-Daughter
102	Maria	Sanchez	1007	Juan	Sanchez	44	Wife-Husband
103	Hilary	Connor	1008	Fred	Connor	25	Wife-Husband
103	Hilary	Connor	1009	Jackie	Connor	2	Mother-Daughter

Sample Table

- The normalized CASE table is *not* a relation because it does not have a primary key.
- To transform this table into a relation, a primary key needs to be identified.
  - Composite key (CaseID,ControlID) is a suitable primary key for this table.

Decomposition

- An alternative method to flattening is decomposition, where the table is decomposed into two or more tables that will replace the original table.



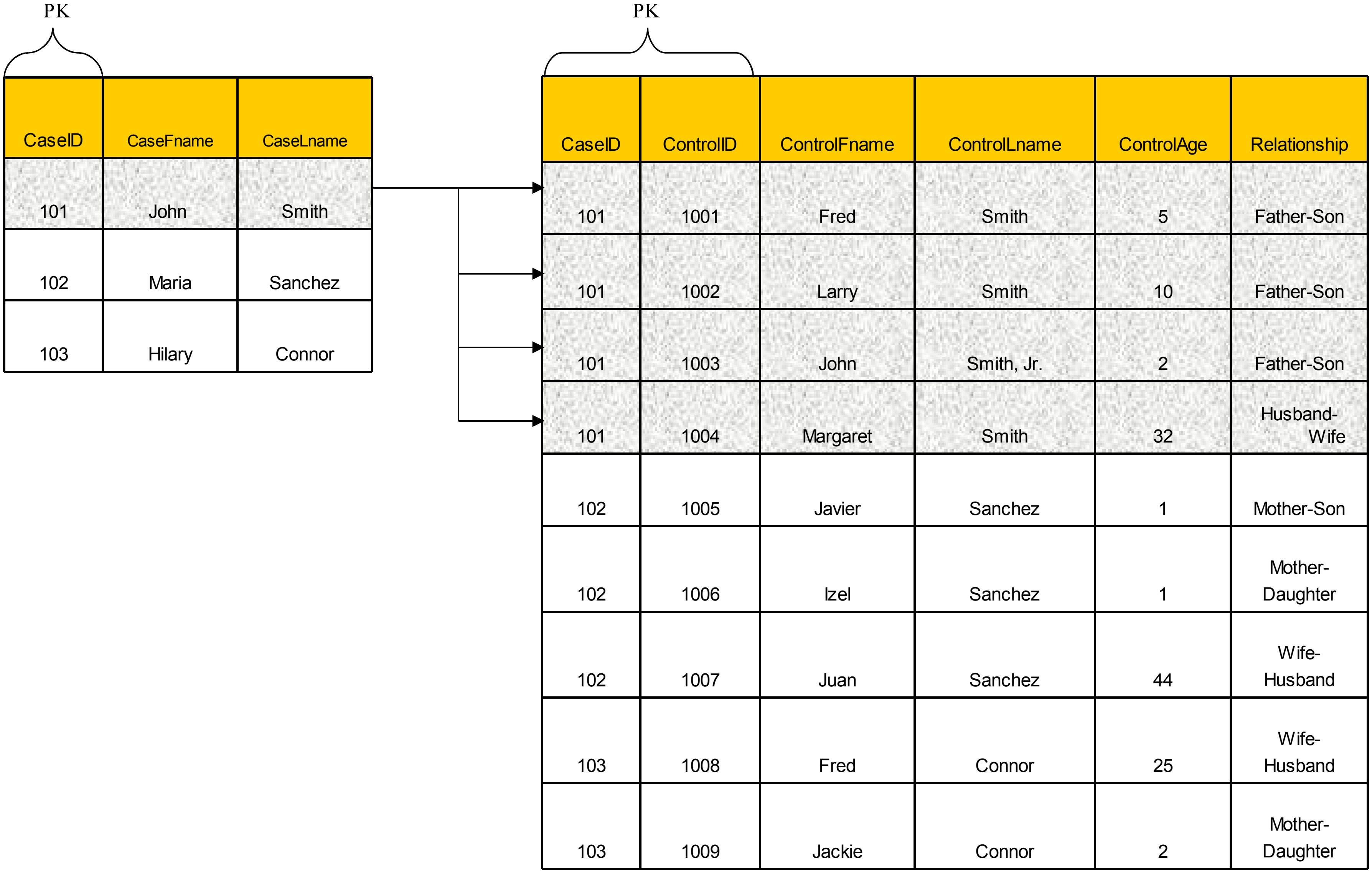
repeating

non-repeating

repeating

non-repeating

Case and Control Table



CaseID	CaseFname	CaseLname
101	John	Smith
102	Maria	Sanchez
103	Hilary	Connor

CaseID	ControlID	ControlFname	ControlLname	ControlAge	Relationship
101	1001	Fred	Smith	5	Father-Son
101	1002	Larry	Smith	10	Father-Son
101	1003	John	Smith, Jr.	2	Father-Son
101	1004	Margaret	Smith	32	Husband-Wife
102	1005	Javier	Sanchez	1	Mother-Son
102	1006	Izel	Sanchez	1	Mother-Daughter
102	1007	Juan	Sanchez	44	Wife-Husband
103	1008	Fred	Connor	25	Wife-Husband
103	1009	Jackie	Connor	2	Mother-Daughter

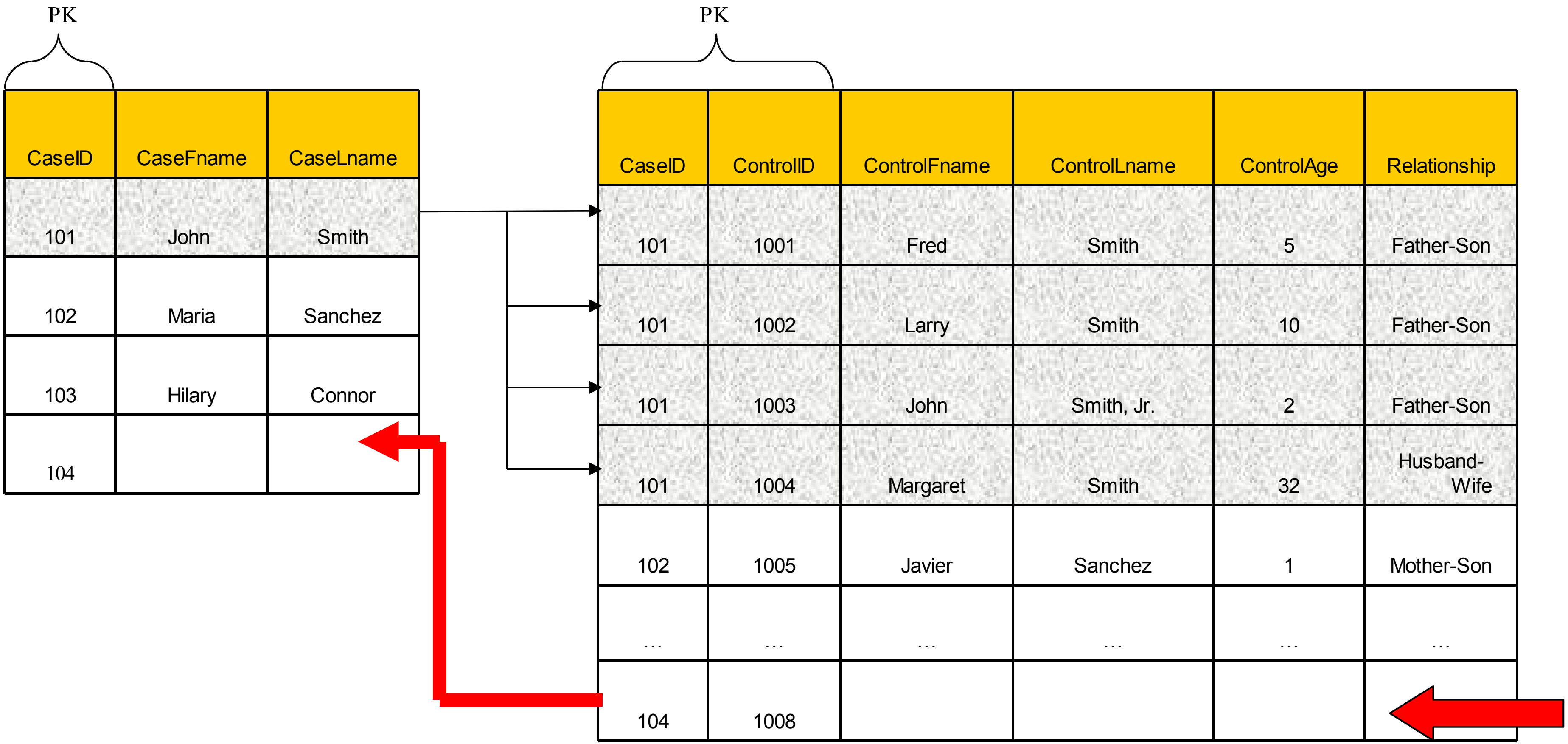
Steps of First Normal Form

- Identify any field that contains multiple pieces of information.
- Break up any fields found in (1) into separate fields.
- Create a separate table for each set of related data.
- Identify each set of related data with a primary key.

Data Anomalies in 1NF Relations

- Redundancies in 1NF relations lead to data anomalies, ie, side effects that the data experience due to some relational operations.
- Two main categories:
  - Insertion/deletion
  - Update

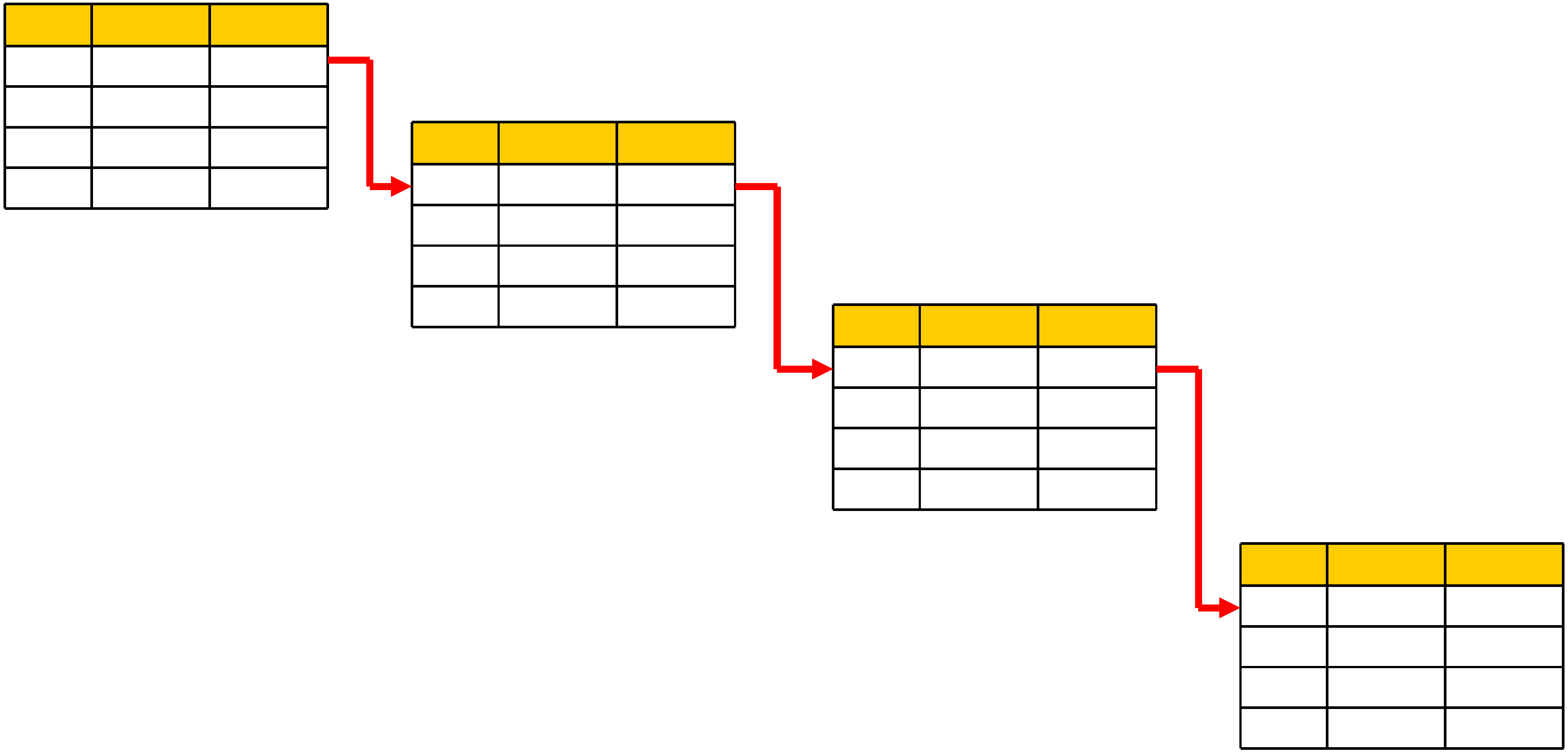
Example



CaseID	CaseFname	CaseLname
101	John	Smith
102	Maria	Sanchez
103	Hilary	Connor
104		

CaseID	ControlID	ControlFname	ControlLname	ControlAge	Relationship
101	1001	Fred	Smith	5	Father-Son
101	1002	Larry	Smith	10	Father-Son
101	1003	John	Smith, Jr.	2	Father-Son
101	1004	Margaret	Smith	32	Husband-Wife
102	1005	Javier	Sanchez	1	Mother-Son
…	…	…	…	…	…
104	1008				

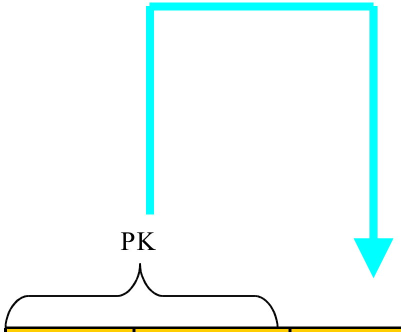
Table Relationships



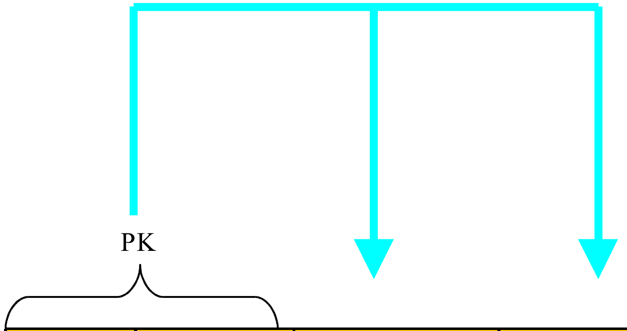
Second Normal Form

- A table is in *Second Normal Form* (2NF) if and only if the following two conditions are met:
  - The table is in 1NF.
  - No non-key attribute is partially dependent on any key (that is, every attribute is fully dependent upon every key).

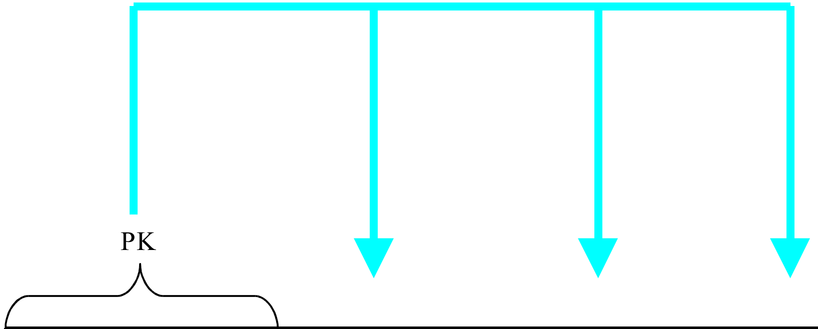
All data in the table must apply directly to the subject (entity) of the table!



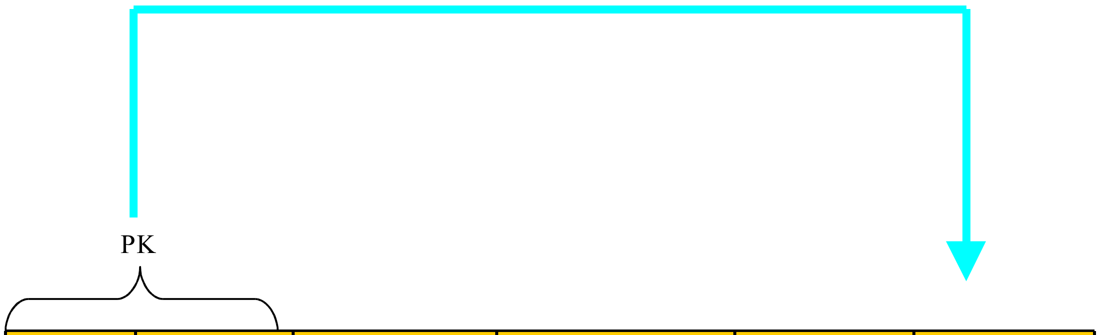
CaseID	ControlID	ControlFname	ControlLname	ControlAge	Relationship
101	1001	Fred	Smith	5	Father-Son
101	1002	Larry	Smith	10	Father-Son
101	1003	John	Smith, Jr.	2	Father-Son
101	1004	Margaret	Smith	32	Husband-Wife
102	1005	Javier	Sanchez	1	Mother-Son
…	…	…	…	…	…



CaseID	ControlID	ControlFname	ControlLname	ControlAge	Relationship
101	1001	Fred	Smith	5	Father-Son
101	1002	Larry	Smith	10	Father-Son
101	1003	John	Smith, Jr.	2	Father-Son
101	1004	Margaret	Smith	32	Husband-Wife
102	1005	Javier	Sanchez	1	Mother-Son
…	…	…	…	…	…



CaseID	ControlID	ControlFname	ControlLname	ControlAge	Relationship
101	1001	Fred	Smith	5	Father-Son
101	1002	Larry	Smith	10	Father-Son
101	1003	John	Smith, Jr.	2	Father-Son
101	1004	Margaret	Smith	32	Husband-Wife
102	1005	Javier	Sanchez	1	Mother-Son
…	…	…	…	…	…



CaseID	ControlID	ControlFname	ControlLname	ControlAge	Relationship
101	1001	Fred	Smith	5	Father-Son
101	1002	Larry	Smith	10	Father-Son
101	1003	John	Smith, Jr.	2	Father-Son
101	1004	Margaret	Smith	32	Husband-Wife
102	1005	Javier	Sanchez	1	Mother-Son
…	…	…	…	…	…

Relationship

PK

CaseID	ControlID	Relationship
101	1001	Father-Son
101	1002	Father-Son
101	1003	Father-Son
101	1004	Husband-Wife
102	1005	Mother-Son
…	…	…

Controls

PK

ControlID	ControlFname	ControlLname	ControlAge
1001	Fred	Smith	5
1002	Larry	Smith	10
1003	John	Smith, Jr.	2
1004	Margaret	Smith	32
1005	Javier	Sanchez	1
…	…	…	…

Steps of Second Normal Form

- Identify any fields that do not relate directly to the primary key.
- Create new tables accordingly.
- Assign or create new primary keys.
- Repeat steps (1) through (3) as needed.
- Create the requisite foreign keys indicating the relationships.

PK

ProjID	EmpID	EmpName	EmpDpt	EmpHrlyRate	TotalHrs
100	1234	Hyde	MIS	65	10
100	9808	Jones	TechSupport	45	6
100	2348	Smith	Engineering	45	6
100	5422	McCulloch	Cabling	30	12
100	4323	Sherwood	MIS	65	5
…	…	…	…	…	…

Employee

PK

EmpID	Empname	EmpDpt	EmpHrlyRate
1234	Hyde	MIS	65
9808	Jones	TechSupport	45
2348	Smith	Engineering	45
5422	McCulloch	Cabling	30
4323	Sherwood	MIS	65
…	…	…	…

Hours-Assigned

PK

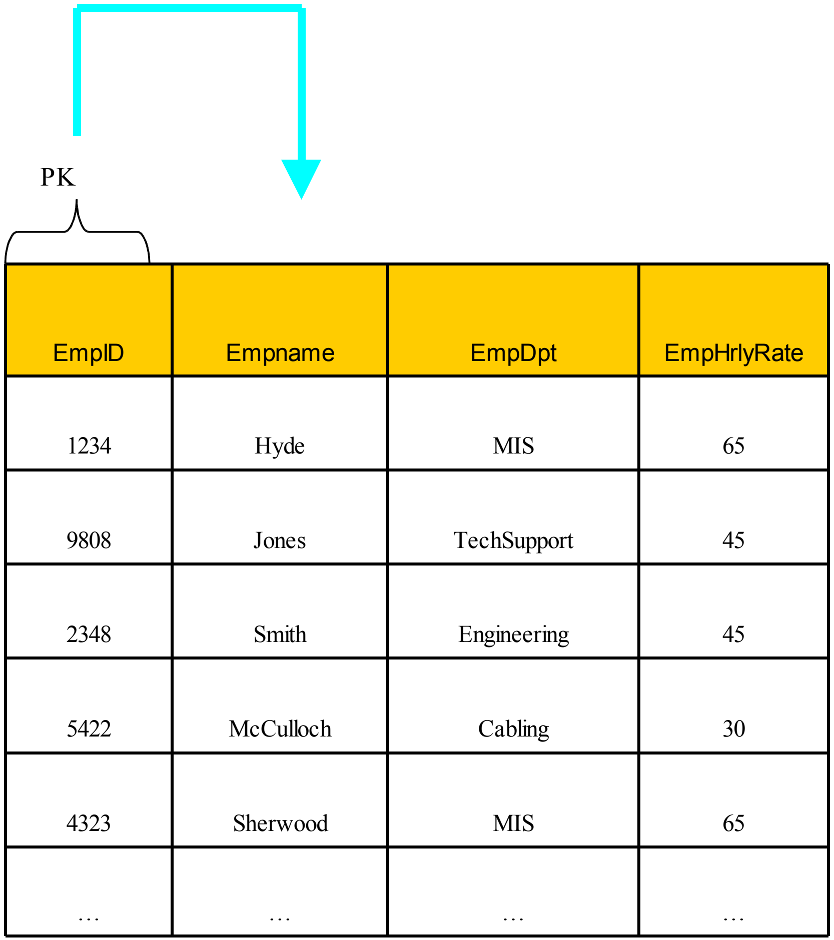
ProjID	EmpID	TotalHrs
100	1234	10
100	9808	6
100	2348	6
100	5422	12
100	4323	5
…	…	…

Third Normal Form

- A table is in *Third Normal Form* (3NF) if and only if the following two conditions are met:
  - The table is in 2NF.
  - Every nonkey column is independent of every other nonkey column. In other words, the fields of a table other than the keys should be mutually independent.

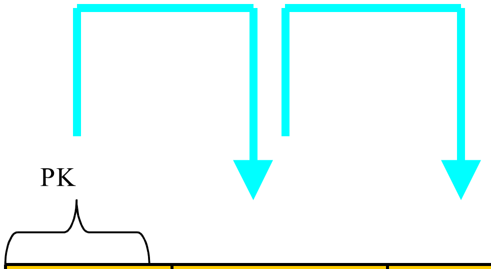
  Eliminates fields that can be derived from other fields!

Employee



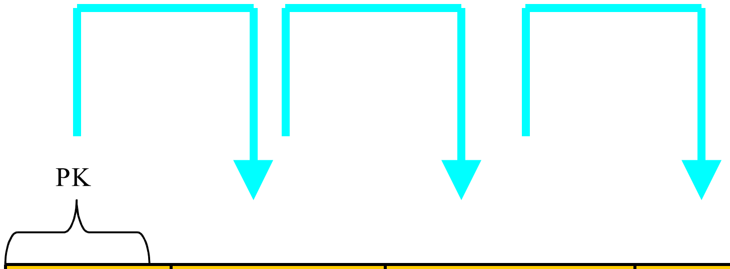
EmpID	Empname	EmpDpt	EmpHrlyRate
1234	Hyde	MIS	65
9808	Jones	TechSupport	45
2348	Smith	Engineering	45
5422	McCulloch	Cabling	30
4323	Sherwood	MIS	65
…	…	…	…

Employee



EmpID	Empname	EmpDpt	EmpHrlyRate
1234	Hyde	MIS	65
9808	Jones	TechSupport	45
2348	Smith	Engineering	45
5422	McCulloch	Cabling	30
4323	Sherwood	MIS	65
…	…	…	…

Employee



EmpID	Empname	EmpDpt	EmpHrlyRate
1234	Hyde	MIS	65
9808	Jones	TechSupport	45
2348	Smith	Engineering	45
5422	McCulloch	Cabling	30
4323	Sherwood	MIS	65
…	…	…	…

Employee

PK

EmpID	Empname	EmpDpt
1234	Hyde	MIS
9808	Jones	TechSupport
2348	Smith	Engineering
5422	McCulloch	Cabling
4323	Sherwood	MIS
…	…	…

Charges

PK

EmpDpt	EmpHrlyRate
MIS	65
TechSupport	45
Engineering	45
Cabling	30

Steps of Third Normal Form

- Identify any fields that depend on any of the nonkey fields of the table (or alternatively, separate fields that do not depend on the key).
- Create new tables accordingly.
- Assign or create new primary keys.
- Repeat steps (1) through (3) as needed.

Fourth Normal Form

- A table is in *Fourth Normal Form* (4NF) if and only if the following two conditions are met:
  - The table is in 3NF.
  - In a many-to-many relationship, independent entities cannot be stored in the same table.

  A table cannot contain fields for two or more independent subjects (entities).

X	Y	Z
X1	Y1	Z1
X2	Y2	Z2
X3	Y2	Z3
X4	Y3	Z4

Decompose

X	Y
X1	Y1
X2	Y2
X3	Y2
X4	Y3

Y	Z
Y1	Z1
Y2	Z2
Y2	Z3
Y3	Z4

Join

X	Y	Z
X1	Y1	Z1
X2	Y2	Z2
X2	Y2	Z3
X3	Y2	Z2
X3	Y2	Z3
X4	Y3	Z4

“Spurious Records”

Table 1

PK1	X	Y
100	X1	Y1
200	X2	Y2
300	X3	Y2
400	X4	Y3

Table 2

PK2	Y	Z
10	Y1	Z1
20	Y2	Z2
30	Y2	Z3
40	Y3	Z4

Linking Table

FK1	FK2
100	10
200	20
300	30
400	40

Fifth Normal Form

- A table is in *Fifth Normal Form* (5NF) if and only if the following condition is met:
  - The original table must be reconstructed from the tables into which it has been broken down.

The source data should be able to be recreated from the tables that have met 1NF, 2NF, 3NF, and 4NF!

Tradeoff of Normalization

- Normalized databases will most likely be slower for updating, retrieving data from, and modifying.
- Stability and endurance are achieved at the expense of convenience and performance.
- However, normalization favors data integrity and scalability over simplicity and speed.

At the End of the Day…

- From our tables, we:
  - Eliminated multivalued fields
  - Ensured that every column in a table that is not a key related to the primary key
  - Ensured the that the fields of a table that are not keys are mutually independent.
  - Retained original relationships and maintained data integrity.

Next Time

- Creating a database