数据库原理

The Theory of Database System

第四章 关系规范化理论

中国矿业大学计算机学院

第四章 关系规范化理论

4.1 问题的提出

4.2 函数依赖和范式

4.3 数据依赖的公理系统

4.4 关系模式的分解方法

姓名	地址	电话	商铺名称	商铺电话	菜品名称	份数	备注	状态
张三	莲花小区A座	13313211111	阳光家宴	18919011111	土豆丝	1		1
张三	莲花小区A座	13313211111	四季鱼馆	18919011112	藤椒鱼	1	微辣	1
张三	莲花小区A座	13313211111	一点甜	18919011113	多肉葡萄	2		1
李四	幸福里2栋	13852033333	五味坊	18919011114	肉夹馍	2		0
李四	幸福里2栋	13852033333	五味坊	18919011114	陕西凉皮	1	不放辣	0
李四	幸福里2栋	13852033333	五味坊	18919011114	酸梅汤	2	冰镇	0
……	……	……	……		……	……	……	……

规范化的概念

一个低一级范式的关系模式，通过模式分解可以转换为若干个高一级范式的关系模式的集合，这种过程就叫做规范化。

范式的概念

- 范式表示关系模式满足的某种级别。
- 1971年E.F.Codd 提出范式概念

  1NF 2NF 3NF BCNF 4NF 5NF

$$5NF \subset 4NF \subset BCNF \subset 3NF \subset 2NF \subset 1NF$$

关系模式

R (U，D，dom，I，F)

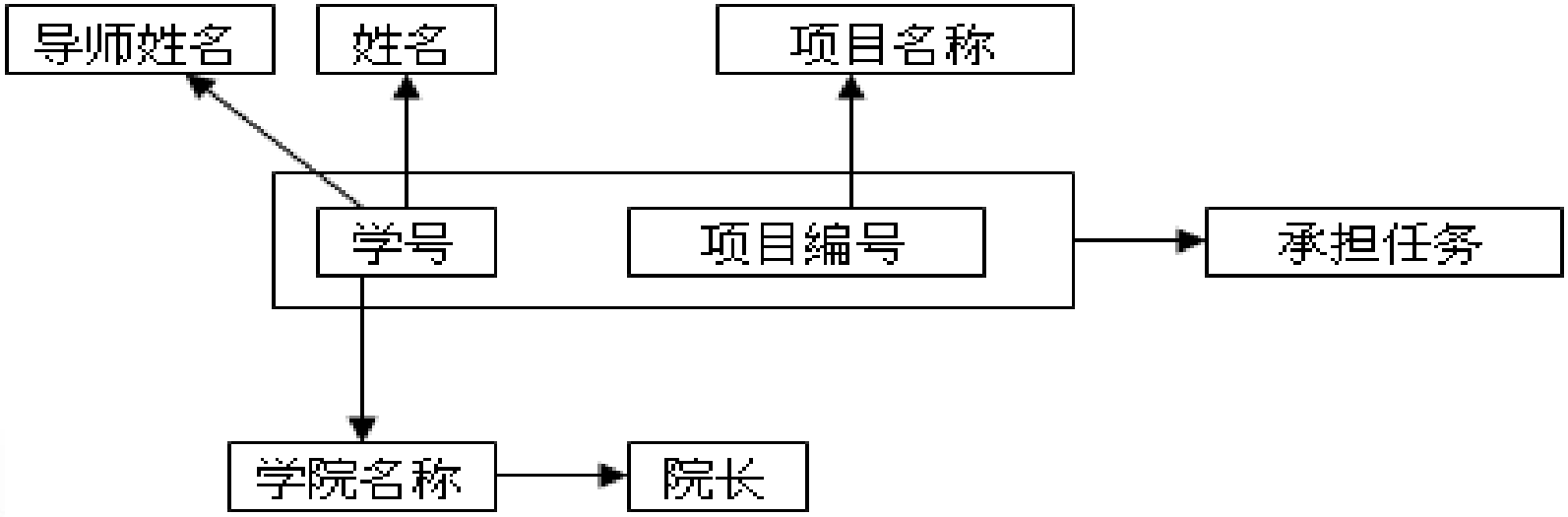
数据依赖：关系中属性间互相依存、互相制约的关系。

[*函数依赖*、多值依赖、连接依赖、分层依赖和相互依赖]

例如：

U={学号、姓名、学院名称、院长、项目编号、项目名称、承担任务、导师姓名}

F={学号→姓名，学院名称→院长，学号→学院名称，（学号，项目编号）→承担任务，项目编号→项目名称，学号→导师姓名}



学号	姓名	学院名称	院长	项目编号	项目名称	承担任务	导师姓名
20082401	周黎明	计算机学院	李洲彤	**0042**	提升机稳定性研究	实验分析	贺信维
20082402	李毅先	计算机学院	李洲彤	**0042**	提升机稳定性研究	系统设计	张琦
20082402	李毅先	计算机学院	李洲彤	**0052**	多维数据分析研究	软件编码	萨林
20083401	王鑫鑫	数学学院	吴兆民	**0091**	定理证明自动化研究	软件编码	刘玉琴
20083402	何飞雨	数学学院	吴兆民	**0083**	最大熵原理研究	软件编码	刘玉琴
20083403	杨宇奇	数学学院	吴兆民	**0063**	软件测试路径分析	实验分析	刘坤鹏

缺点：1、冗余太大

2、操作异常

1)插入异常　2)删除异常　3)修改异常

存在问题的原因

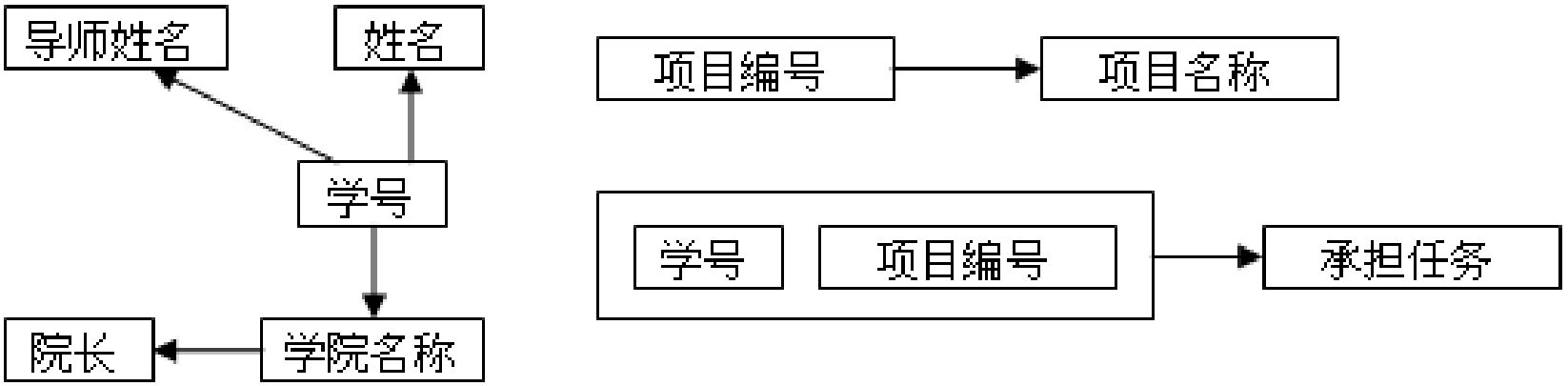
- 数据冗余和操作异常产生的重要原因就是对数据依赖的不恰当处理，最终导致不合理的关系模式的设计。
- 一个关系中各属性之间可能是相互关联的，而这种关联有“强”有“弱”，有直接关联，也有间接关联。
- 不从语义上研究和考虑属性子集间的这种关联简单地将属性随意地编排在一起，形成泛关系模式，就可能产生很大程度的数据冗余，导致“排他”现象，从而引发各种冲突和异常。

解决方法

- 解决问题的方法就是将关系模式进一步分解
- 将关系模式中的属性按照一定的约束条件重新分组，争取“**一个关系模式只描述一个独立的实体**”，使得逻辑上独立的信息放在独立的关系模式中，即进行关系模式的规范化处理。

（1）第一种分解方法

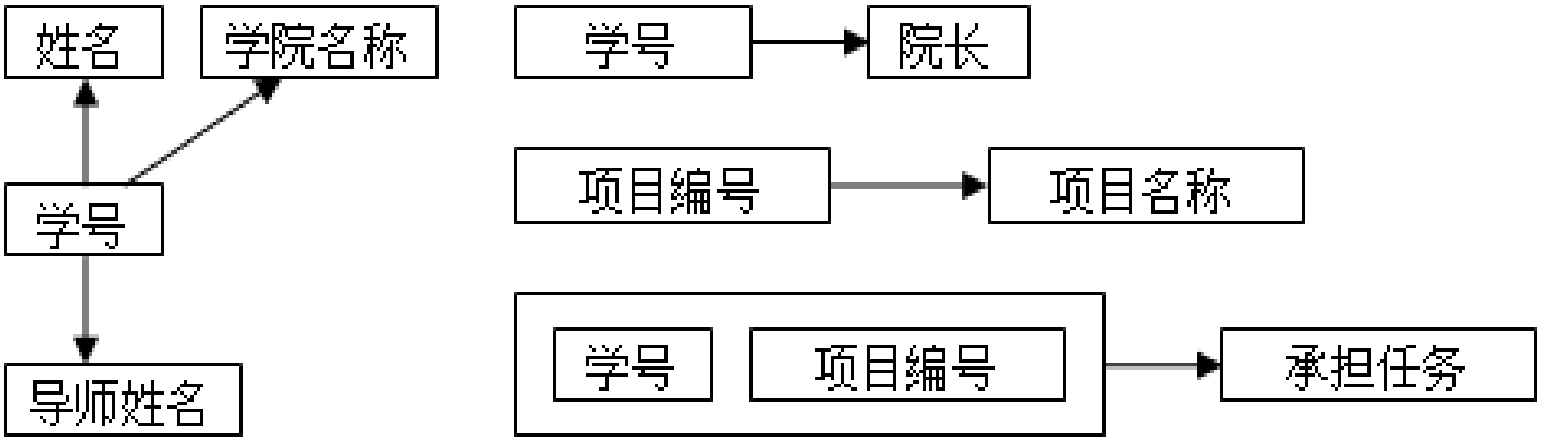
S_D（学号, 学生姓名, 学院名称, 院长, 导师姓名）
P（项目编号, 项目名称）
S_P（学号, 项目编号, 承担任务）



消除部分冗余数据

（2）第二种分解方法

S（学号，学生姓名，学院名称，导师姓名）
P（项目编号，项目名称）
S_MN（学号，院长）
S_P（学号，项目编号，承担任务）



消除冗余数据，但丢失数据依赖关系

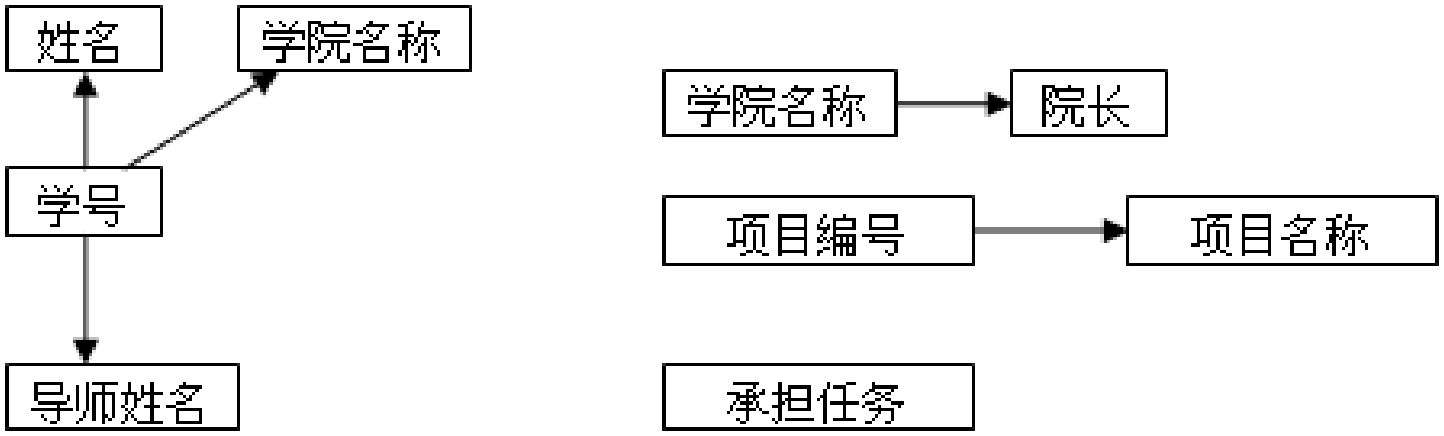
（3）第三种分解方法

S（学号，学生姓名，学院名称，导师姓名）

P（项目编号，项目名称）

D（学院名称，院长）

T（承担任务）



消除冗余数据，但丢失了信息

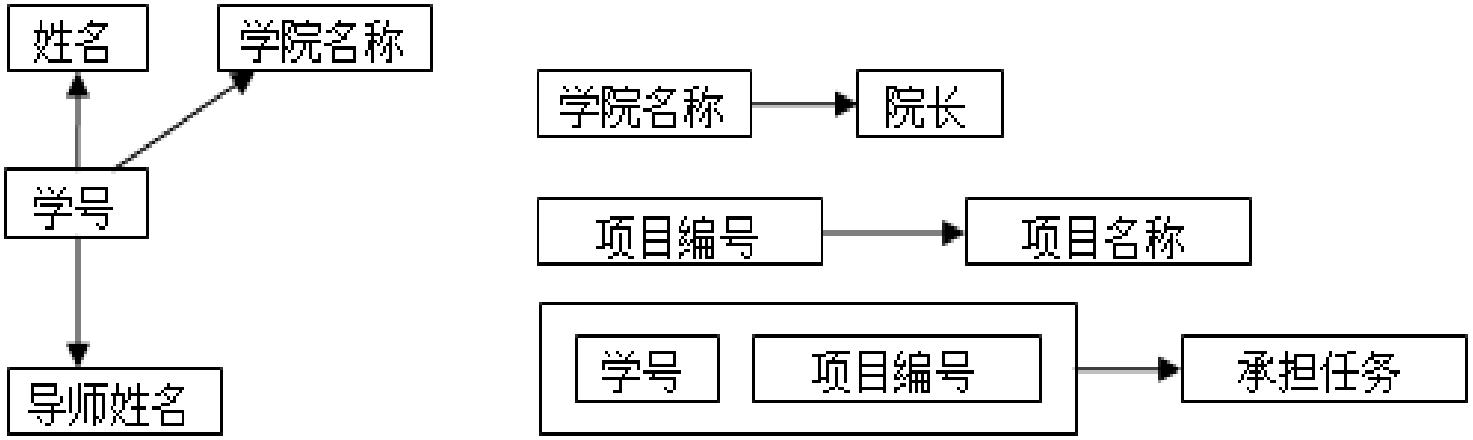
（4）第四种分解方法

S（学号，学生姓名，学院名称、导师姓名）

P（项目编号，项目名称）

D（学院名称，院长）

S_P（学号，项目编号，承担任务）



消除冗余，保持数据依赖，保证信息不丢失

规范化理论的提出

“关系规范化”理论包含两个核心的问题：

- 一、如何判断关系模式中存在的问题。
  - 通过分析关系模式中的数据依赖关系，判断关系模式的“范式”级别，从而得到这种模式中可能存在的数据冗余和操作异常问题；
- 二、如何解决关系模式中存在的问题，即对关系模式进行分解。
  - 如何分解？“关系规范化”理论为解决这些问题提供了理论依据和相应的算法。

4.2 函数依赖和范式

一、函数依赖：

属性或属性组之间可能存在的依赖性。

1、定义

定义4.1： 设R(U)是属性集U上的关系模式。X，Y是U的子集。若对于R(U)的任意一个可能的关系r，r 中不可能存在两个元组在X上的属性值相等，而在Y上的属性值不等，则称X函数确定Y或Y函数依赖于X，记作X→Y。

或者说：关系模式R(U)的任一具体关系，属性集X在任意元组上的值能唯一决定属性集Y在该元组上的值，则称X函数确定Y或Y函数依赖于X，记作X→Y。

或者说：设R(U)是一个关系模式，X，Y是U的子集，对于R中X的每一个值都有Y的唯一值与之对应，则称X函数确定Y或Y函数依赖于X，记作X→Y。

例：U={学号，学院，院长，课程号，课程名，成绩}

F={学号→学院，学院→院长，课程号→课程名，
（学号，课程号）→成绩}

注意：函数依赖不是指关系模式R的某个或某些关系满足的条件，而是指R的一切关系均要满足的约束条件。

由定义可以导出下列概念：

1. **决定因素**：若$X \rightarrow Y$，则X叫做决定因素

2. **平凡的函数依赖**：$X \rightarrow Y$，$Y \subseteq X$，则称$X \rightarrow Y$是平凡的函数依赖。

3. **非平凡的函数依赖**：$X \rightarrow Y$，但$Y \not\subseteq X$，则称$X \rightarrow Y$是非平凡的函数依赖。

4. **互相依赖**：若$X \rightarrow Y$，$Y \rightarrow X$，则记作$X \longleftrightarrow Y$。

5. 若Y不函数依赖于X，则记作$X \not\rightarrow Y$。

定义4.2： 完全函数依赖

在R(U)中，如果$X \rightarrow Y$，并且对于X的任何一个真子集X'，都有$X' \not\rightarrow Y$，则称Y对X完全函数依赖。记作：

$$X \xrightarrow{F} Y$$

定义4.3： 部分函数依赖

在R(U)中，如果$X \rightarrow Y$，并且对于X的一个真子集X'，有$X' \rightarrow Y$，则称Y对X部分函数依赖。记作：

$$X \xrightarrow{P} Y$$

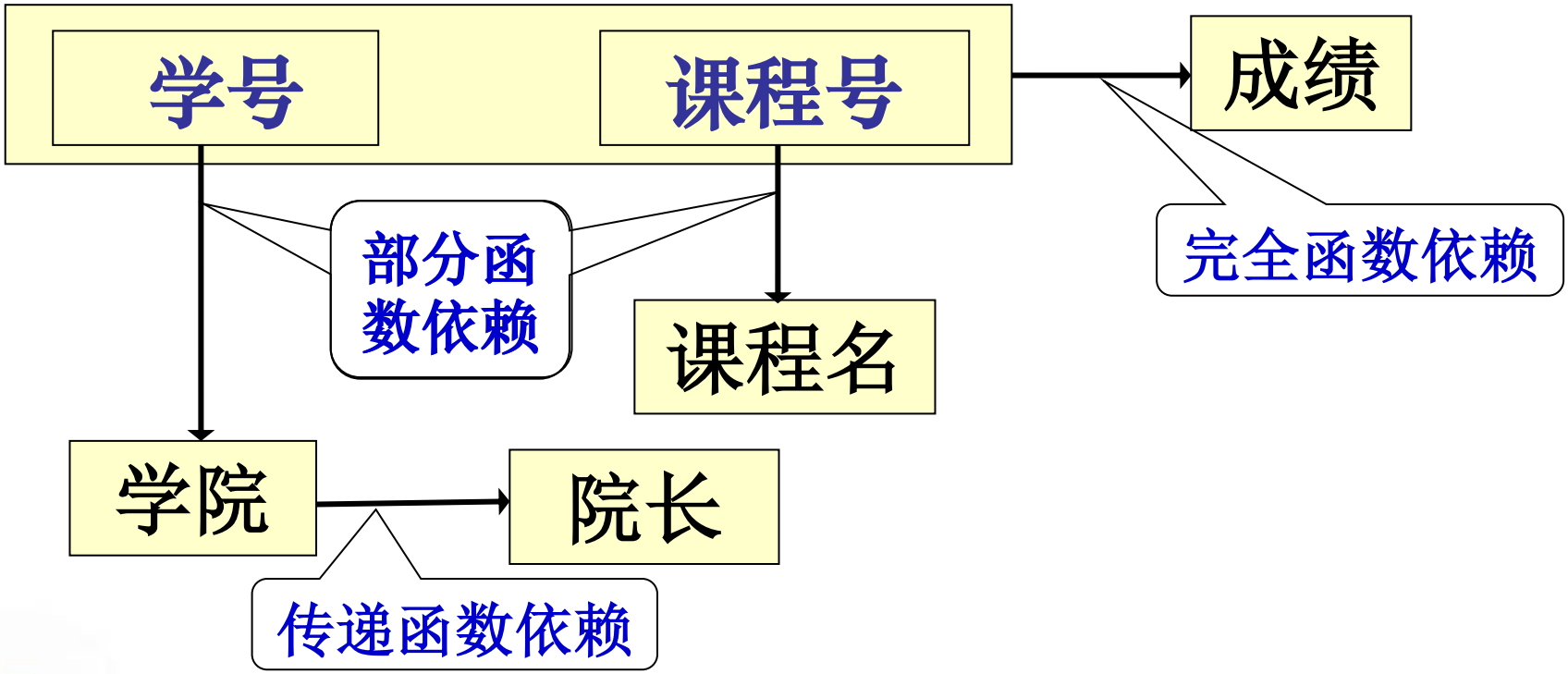
定义4.4：传递函数依赖

在R(U)中，如果$X \rightarrow Y$，$(Y \not\subseteq X)$，$Y \not\rightarrow X$，$Y \rightarrow Z$，则称Z对X传递函数依赖。

例： U={学号，学院，院长，课程号，课程名，成绩}

F={学号→学院，学院→院长，　课程号→课程名，

（学号，课程号）→成绩 }

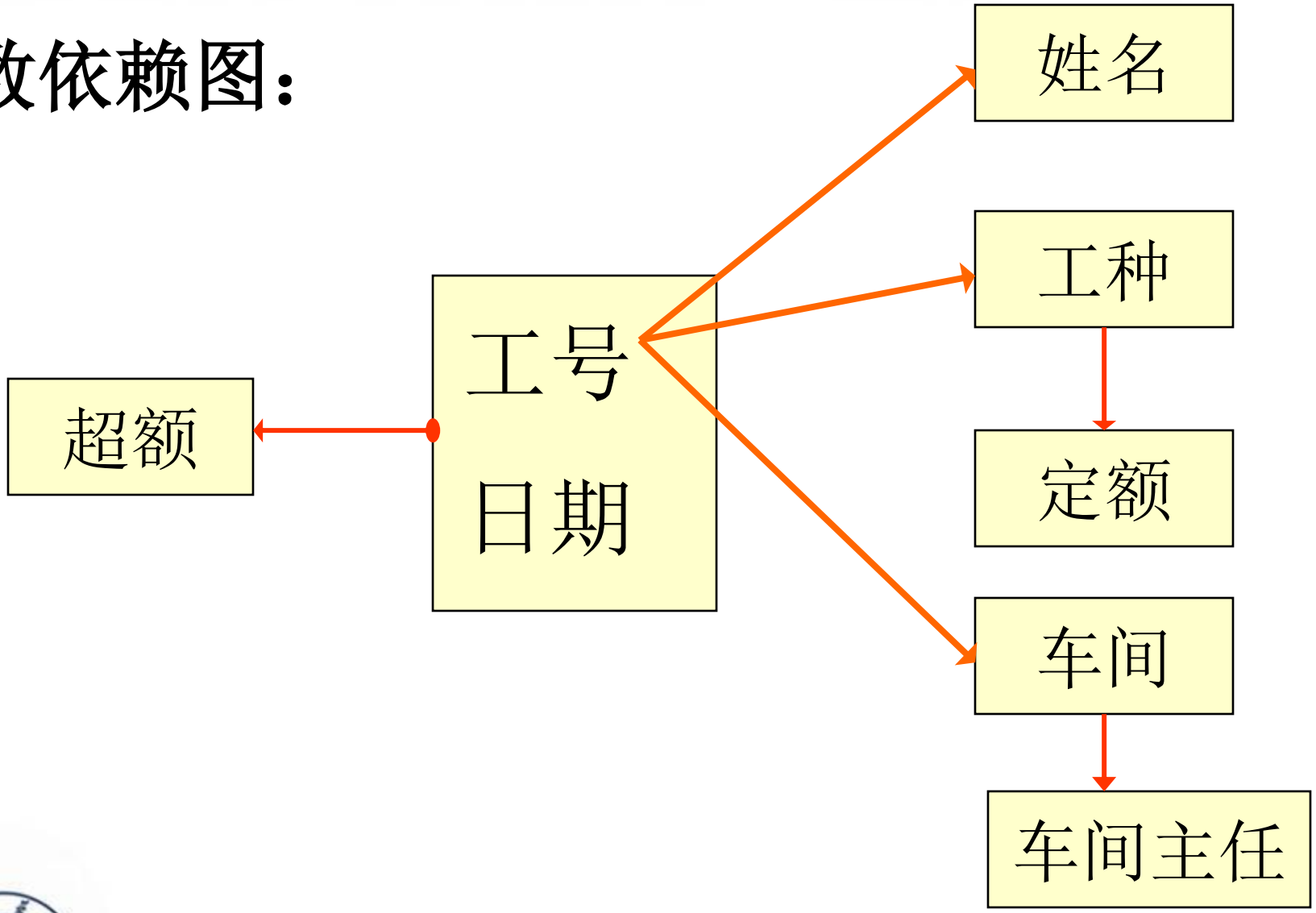


例：设车间考核职工完成工作量的关系模式如下：

U={日期，工号，姓名，工种，超额定额，车间，车间主任}

F={（日期，工号）→超额，工号→姓名，
工号→车间，工号→工种，
工种→定额，车间→车间主任　}

函数依赖图：



分析：

日期、工号 → 超额　　完全函数依赖

日期、工号 → 姓名
日期、工号 → 工种
日期、工号 → 车间　　部分函数依赖

日期、工号→ 定额
日期、工号 → 车间主任　　传递函数依赖

二、码

定义4.5：设K为R(U，F)中的属性或属性组，若 $K \xrightarrow{F} U$，则K为R的**候选码**。

- **主码：**若候选码多于一个，则选定其中的一个为主码。
- **主属性：**包含在任何一个侯选码中的属性。
- **非主属性：**不包含在任何码中的属性。
- **全码：**整个属性组是码。

定义4.6: 关系模式R中属性或属性组X并非R的码，但X是另一个关系模式的码，则称X是R的外码。

主码与外码提供了一个表示关系间联系的手段。

超码（Super Key）

- 包含候选码的属性集合称为超码。

- 例如：学号为学生表的候选码，则包含学号的任一个属性组都是超码。

三、第一范式(1NF)

定义：满足关系的每一个分量是不可分的数据项这一条件的关系模式就属于第一范式(1NF)。

sid	name		class	telephone	enrollment	
	lname	fname			cno	major
1	Jones	Allan	2	555-1234	101	No
					108	Yes
2	Smith	John	3	555-4321	105	No
3	Borwn	Harry	2	555-1122	101	Yes
					108	No
4	White	Edward	3	555-3344	102	No
					105	No

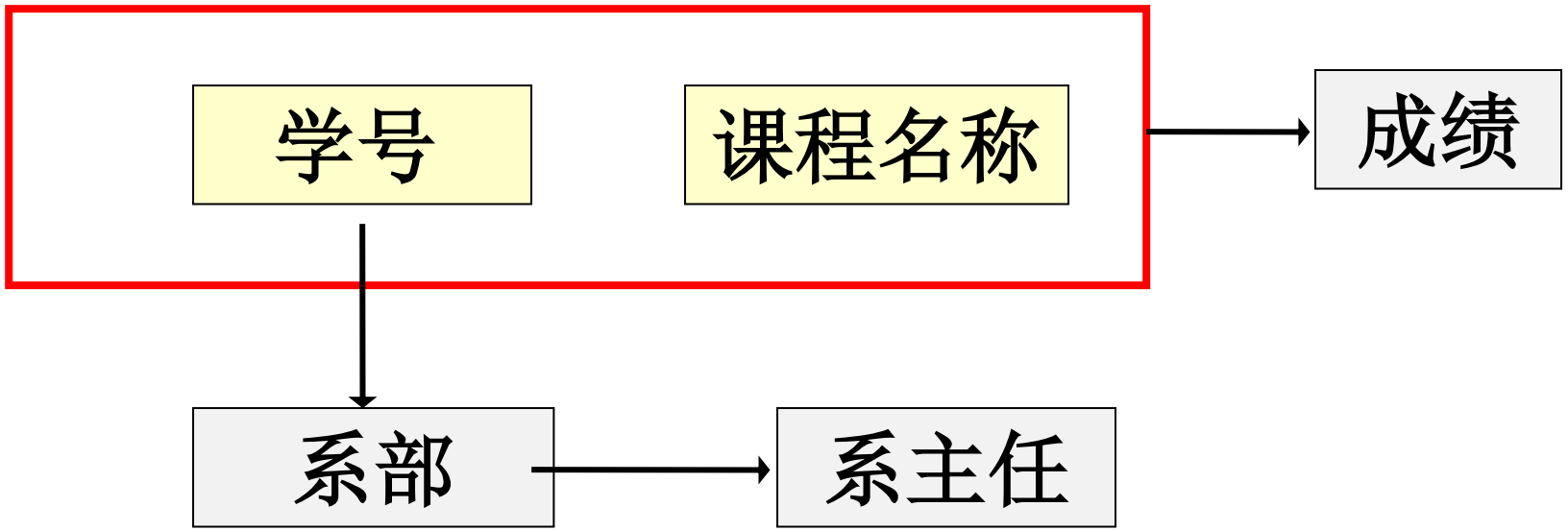
学号	系部	系主任	课程名称	成绩
S001	CS	Jone	operating system	76
S001	CS	Jone	compilation techniques	85
S001	CS	Jone	computer network	90
S002	IS	Mike	compilation techniques	67
S002	IS	Mike	operating system	86
S003	DB	May	computer network	78

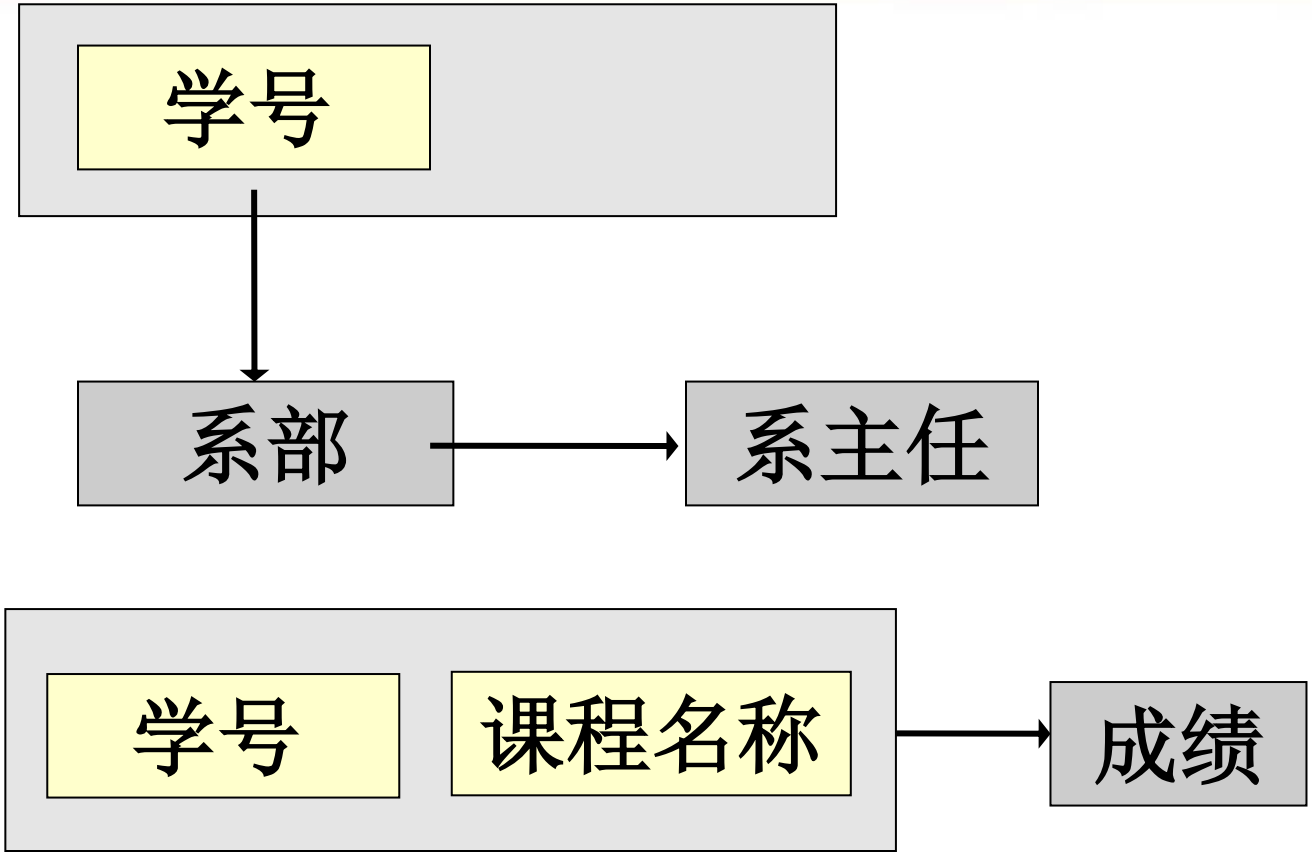
缺点：

1. 插入异常
2. 删除异常
3. 冗余太大
4. 修改复杂

四、第二范式(2NF)

定义：若R∈1NF，且每一个非主属性完全函数依赖于码，则R∈2NF。



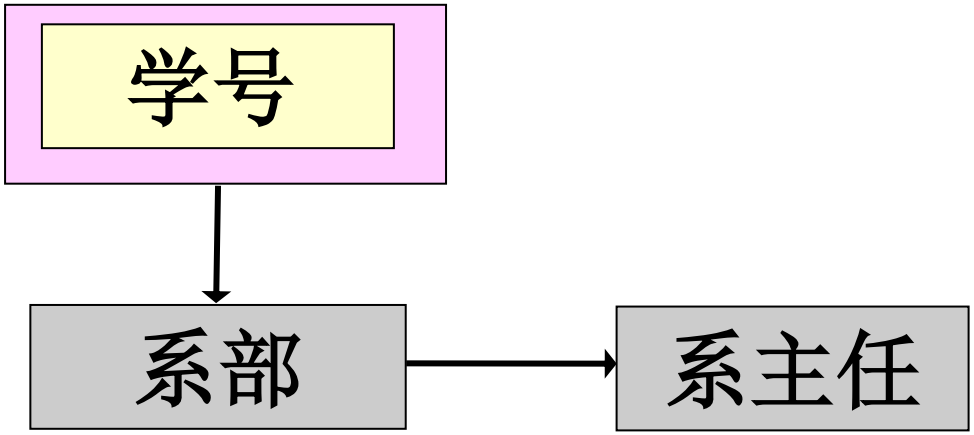


（学号，系部，系主任）

（学号，课程名称，成绩）

四、第三范式(3NF)

定义：关系模式R(U，F)中若不存在这样的码X，属性组Y及非主属性组Z(Z⊆Y)使得X→Y，(Y→X) Y →Z成立，则称R(U，F)∈3NF。



四、第三范式(3NF)

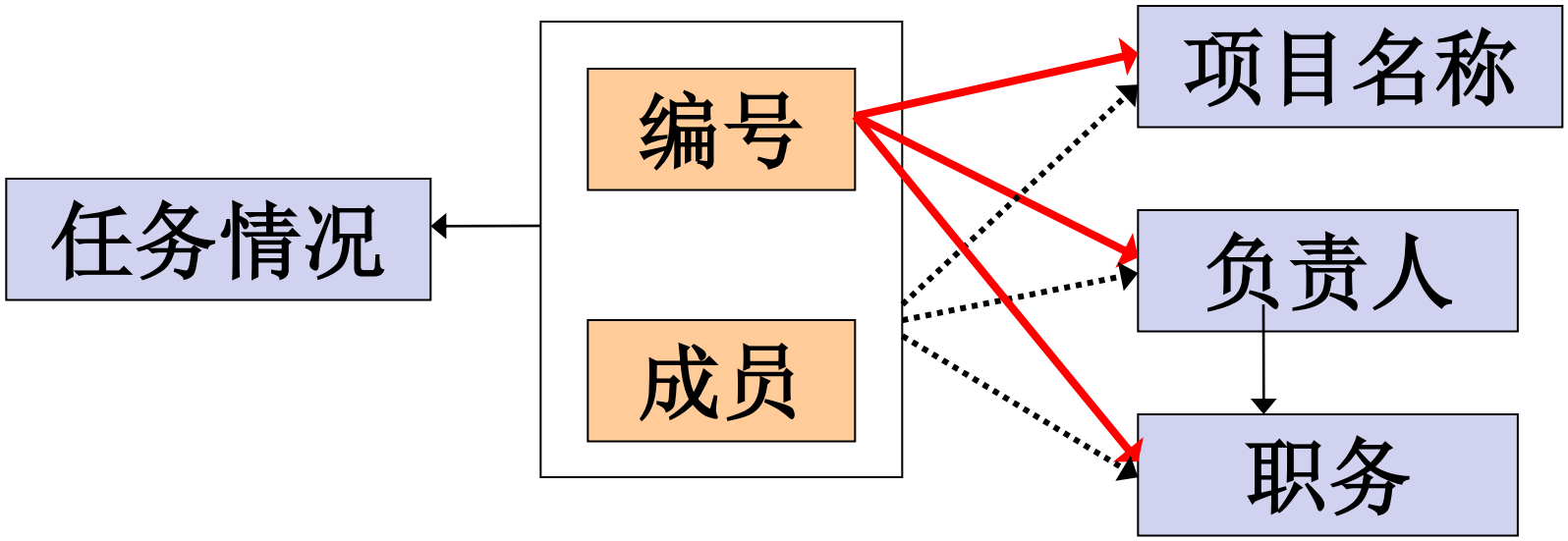
或者：

- 若R ∈2NF，且每一个非主属性不传递依赖于码，则R ∈3NF。
- 若R ∈1NF，且每一个非主属性既不部分依赖于码也不传递依赖于码。

例：

项目(编号，项目名称，负责人，职务， 成员，任务情况)

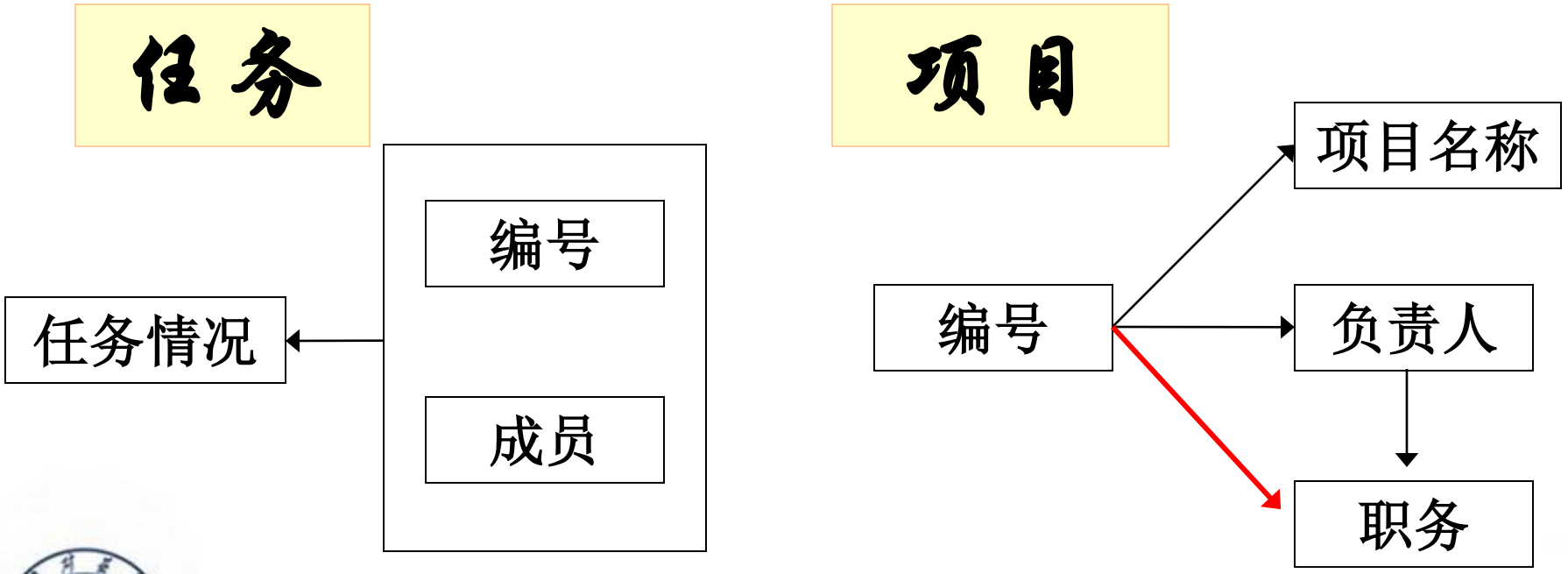
(假设:负责人无重名情况)



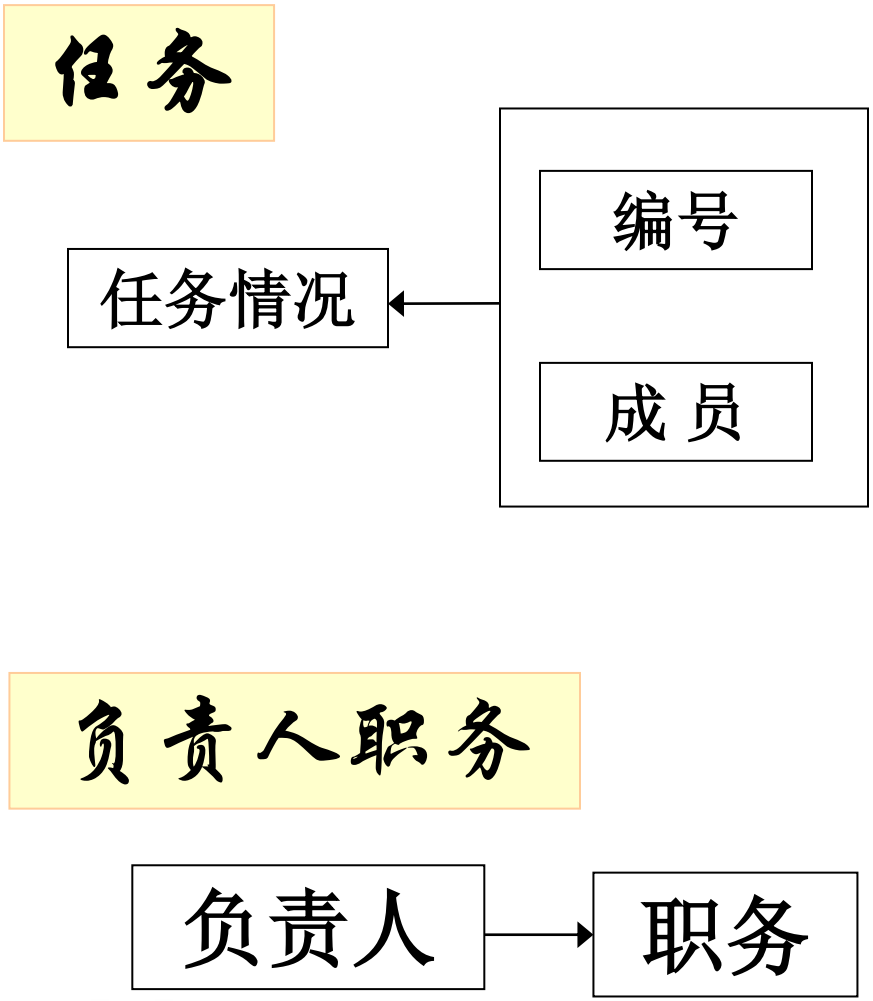
根据2NF要求

任务（编号，成员，任务情况）

项目（编号，项目名称，负责人，职务）



根据3NF要求

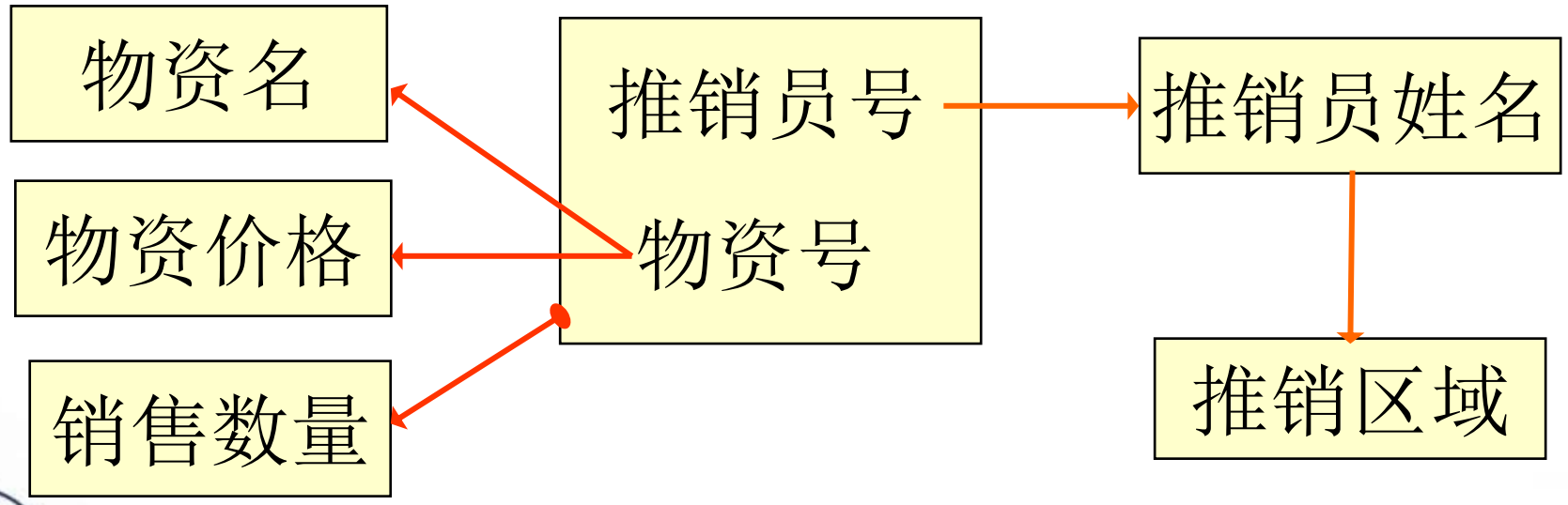


例：分析下列关系属于第几范式

推销员管理信息：

（推销员编号，推销员姓名，物资号，物资名，物资价格，销售数量，销售区域）假设推销员姓名无重名。

函数依赖关系如下图：



分析

（1）候选码：

推销员号＋物资号

（2）存在的函数依赖：

非主属性对码存在部分依赖

（3）达到的范式级别：

属于1NF

分解：

推销员号+物资号

销售数量

推销员号

推销员姓名

推销员姓名

推销区域

物资号

物资名

物资价格

- 凡是满足3NF的关系，一般都能获得满意的效果。但是某些情况下，3NF仍会出现问题。
- 原因是没有对主属性与关键字之间给出任何限制，如果出现**主属性部分或传递依赖于码**，则也会使关系性能变坏

五、BCNF（扩充的3NF）

定义：关系模式R(U, F) ∈1NF。若X→Y且Y ⊈ X时X必含有码，则R(U, F) ∈BCNF。

即：关系模式R(U，F) 中，若每一个决定因素都包含码，则R(U, F) ∈BCNF。

一个满足BCNF的关系模式有:

- 所有非主属性对每一个码都是完全函数依赖。
- 所有主属性对每一个不包含它的码也是完全函数依赖。
- 没有任何属性完全函数依赖于非码的任何一组属性。

*例：*关系模式SJP（S，J，P）

S：学生　[学生选修课程有一定的名次]

J：课程　[每门课程中每一名次只有一个学生]

P：名次　　（名次没有并列）

函数依赖：　（S，J）→ P

（J，P）→ S

分析得知：　SJP ∈ 3NF

SJP ∈ BCNF

*例：*关系模式STJ（S，T，J）

S：学生 [某一学生选定某门课，就对应一个固定教师]

T：教师 [每个教师只教一门课]

J： 课程 [每门课有若干教师]

函数依赖：（S，J）→ T

T → J

分析得知：STJ ∈ 3NF

但是：STJ ∉ BCNF 因为： T → J

STJ可以分解为：ST（S，T）TJ（T，J）

规范化小结(续)

- 规范化程度过低的关系不一定能够很好地描述现实世界，可能会存在插入异常、删除异常、修改复杂、数据冗余等问题。
- 一个低一级范式的关系模式，通过模式分解可以转换为若干个高一级范式的关系模式集合，这种过程就叫关系模式的规范化。
- 关系数据库的规范化理论是数据库逻辑设计的工具。

规范化小结(续)

关系模式规范化的基本步骤

消除决定属性集非码的非平凡函数依赖

1NF

↓ 消除非主属性对码的部分函数依赖

2NF

↓ 消除非主属性对码的传递函数依赖

3NF

↓ 消除主属性对码的部分和传递函数依赖

BCNF

↓ 消除非平凡且非函数依赖的多值依赖

4NF

规范化的基本思想

- 消除不合适的数据依赖
- 使得各关系模式达到某种程度的“分离”
- 采用“一事一地”的模式设计原则

  让一个关系描述一个概念、一个实体或者实体间的一种联系。若多于一个概念就把它“分离”出去
- 所谓规范化实质上是概念的单一化

规范化的基本思想(续)

- 不能说规范化程度越高的关系模式就越好
- 在设计数据库模式结构时，必须对现实世界的实际情况和用户应用需求作进一步分析，确定一个合适的、能够反映现实世界的模式
- 上面的规范化步骤可以在其中任何一步终止