CSE544
Lecture 1: Introduction

Tuesday, January 2, 2001

1

Staff

- Instructor: Dan Suciu
  - Sieg, Room 318, suciu@cs.washington.edu
  - Office hours: Tuesday, 12-1.
- TAs: Gerome Miklau
  - Office hours: Friday, 12:30-1:30.
- Mailing list: cse544@cs.washington.edu
  - Send mail to majordomo@cs.washington.edu: "subscribe cse544"
- Web page: (a lot of stuff already there) http://www.cs.washington.edu/544

2

Course Times

- In general, Tue-Thu, 10:30-11:50pm
- Special dates:
  - Thursday, Jan 25

3

Goals of the Course

- Purpose:
  - Foundations of database management systems.
  - Issues in building database systems.
  - Introduction to current research issues in databases.

4

Grading

- Homeworks: 35%
  - Very little regurgitation.
  - Meant to be challenging (I.e., fun).
- Project: 20%
  - More later.
- Final: 40%
- Intangibles: 5%

5

Textbook

- **Database Management Systems**, Ramakrishnan and Gehrke.

Also:

- **Foundations of Databases**, Abiteboul, Hull & Vianu

6

Other Useful Texts

- Pair of books by Ullman, Widom and Garcia-Molina
- *Parallel and Distributed DBMS* (Ozsu and Valduriez)
- *Transaction Processing* (Gray and Reuter)
- *Data and Knowledge based Systems* (volumes I, II) (Ullman)
- *Data on the Web* (Abiteboul, Buneman, Suciu)
- *Readings in Database Systems* (Stonebraker and Hellerstein)
- Proceedings of SIGMOD, VLDB, PODS conferences.

7

Prerequisites

- Officially: none
- Real prerequisites:
  - Programming languages
  - Logic
  - Complexity theory
  - Algorithms and data structures

8

Traditional Database Application

Suppose we are building a system to store the information about:

- students
- courses
- professors
- who takes what, who teaches what

Why use a DBMS ?

9

What we need from a database:

- store the data for a long period of time
  - large amounts (100s of GB)
  - protect against crashes
  - protect against unauthorized use
- allow users to query/update:
  - who teaches "CSE142"
  - enroll "Mary" in "CSE444"

10

- allow several (100s, 1000s) users to access the data simultaneously
- allow administrators to change the schema
  - add information about TAs

11

Trying Without a DBMS

Why Direct Implementation Won't Work:

- Storing data: file system is limited
  - size less than 4GB (on 32 bits machines)
  - when system crashes we may loose data
  - password-based authorization insufficient
- Query/update:
  - need to write a new C++/Java program for every new query
  - need to worry about performance

12

- Concurrency: limited protection
  - need to worry about interfering with other users
  - need to offer different views to different users (e.g. registrar, students, professors)
- Schema change:
  - need to rewrite virtually all applications

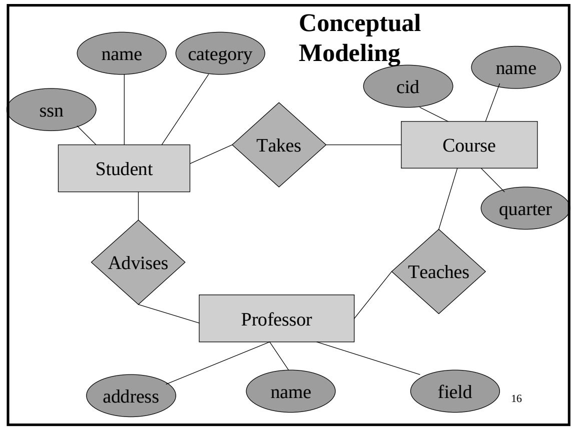
Functionality of a DBMS

- Storage management
- Data Definition Language - DDL
- Data Manipulation Language - DML
  - query language
- Transaction Management
  - concurrency control
  - recovery

Building an Application with a DBMS

- Requirements modeling (conceptual, pictures)
  - Decide what entities should be part of the application and how they should be linked.
- Schema design and implementation
  - Decide on a set of tables, attributes.
  - Define the tables in the database system.
  - Populate database (insert tuples).
- Write application programs using the DBMS
  - way easier now that the data management is taken care of.

Conceptual Modeling



Schema Design and Implementation

- Tables:

Students:

SSN	Name	Category
123-45-6789	Charles	undergrad
234-56-7890	Dan	grad
	…	…

Takes:

SSN	CID
123-45-6789	CSE444
123-45-6789	CSE444
234-56-7890	CSE142
	…

Courses:

CID	Name	Quarter
CSE444	Databases	fall
CSE541	Operating systems	winter

- Separates the logical view from the physical view of the data.

Querying a Database

- Find all courses that "Mary" takes
- S(tructured) Q(uery) L(anguage)

```
select  C.name
from Students S, Takes T, Courses C
where S.name="Mary" and
   S.ssn = T.ssn and T.cid = C.cid
```

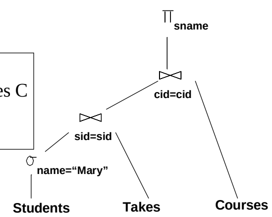
- Query processor figures out how to answer the query efficiently.

Query Optimization

Goal:

Declarative SQL query → ***Imperative query execution plan:***

select C.name
from Students S, Takes T, Courses C
where S.name="Mary" and
S.ssn = T.ssn and T.cid = C.cid



<u>Plan</u>: Tree of Relational Algebra operators, with a choice of algorithm implementation for each operator

Ideally: Want to find best plan. Practically: Avoid worst plans!

19

Database Industry

- Relational databases are a great success of theoretical ideas.
- Oracle has a market cap of over $200B
- Other players: IBM, MS, Sybase, Informix
- Trends:
  - warehousing and decision support
  - data integration
  - XML, XML, XML.

20

What is the Field of Databases ?

- To a theoretical researcher (PODS/ICDT/LICS)
  - Focus on the query languages
  - Query language = logic = complexity classes
- To an applied researcher (SIGMOD/VLDB/ICDE)
  - Query optimization
  - Query processing (yet-another join algorithm)
  - Transaction processing, recovery
  - Novel applications: data mining, high-dimensional search
- To a systems programmer at Oracle:
  - Millions lines of code
- To an application builder:
  - E/R, SQL, ODBC/JDBC

21

Current and Future Data Management

- Current Data Management:
  - relational data for enterprise applications
  - storage
  - query processing/optimization
  - transaction processing
- Future Data Management:
  - XML data for exchange on the Web
  - transport
  - query/data translation
  - information retrieval

22

XML: Semi-structured Data

eXtensible Markup Language:

– *Emerging format for data exchange on the web and between applications.*

```
<db>
 <book>
    <title>Complete Guide to DB2</title>
    <author>Chamberlin</author>
 </book>
 <book>
    <title>Transaction Processing</title>
    <author>Bernstein</author>
    <author>Newcomer</author>
 </book>
 <publisher>
   <name>Morgan Kaufman</name>
   <state>CA</state>
 </publisher>
</db>
```

23

Course (Rough) Outline

- The basics: (quickly)
  - E/R, ODL, the relational model
  - Relational algebra, SQL
  - Views, integrity constraints
- Semistructured data and XML
- Some theory
  - Theory of conjunctive queries
  - Recursive queries (datalog)
  - Query languages, logic, and complexity classes

24

Course Outline (cont)

- Query processing
- Query optimization
- Transaction processing

25

Projects

- Goal: apply some database principles to a new problem
- Suggested topics are from XML (see website), but anything goes.
- Groups of 2-3
- Groups assembled end of week 2;
- Proposals, beginning of week 4
- Touch base with me: *every* two weeks.
- Start Early.

26

Today: Database Design

- E/R - Entity relationship diagrams (Chapter 2)

27

Database Design

- Why do we need it?
  - Agree on structure of the database before deciding on a particular implementation.
- Consider issues such as:
  - What entities to model
  - How entities are related
  - What constraints exist in the domain
  - How to achieve *good* designs

28

Entity-Relationship (E/R) Model

- Basic design paradigm in E/R:
  - Model ***entities*** and their ***properties.***
- For abstraction purposes:
  - Group objects into ***entity sets.***
- What qualifies as a ***good*** entity set ?
  - Entities in an entity set should have common properties.

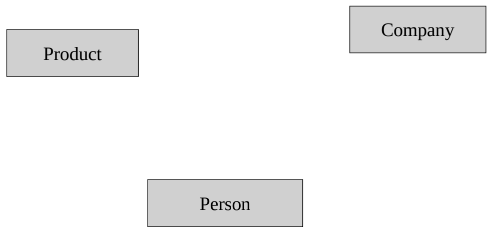
29

E/R Design

- Three steps:
  - Design the entity sets
  - Design their attributes
  - Design the relationships

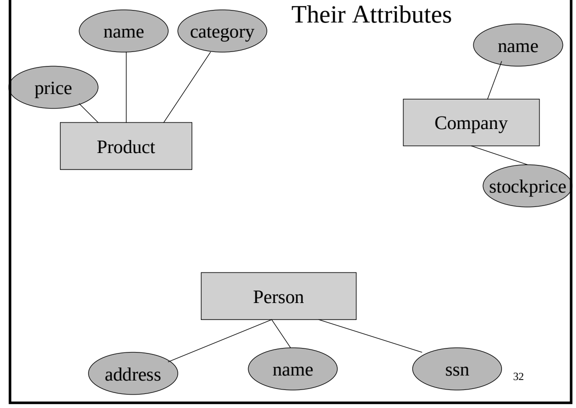
30

E/R Example: The Entity Sets



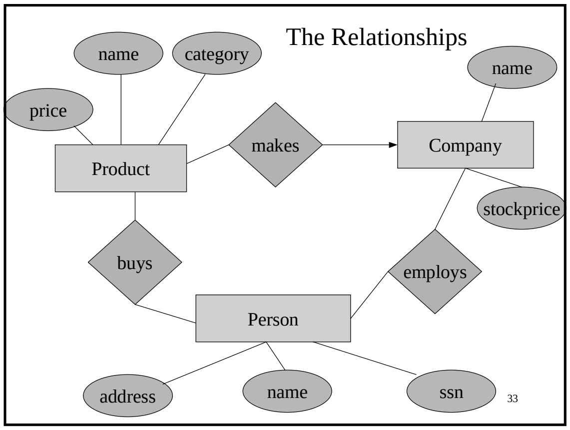
31

Their Attributes



32

The Relationships



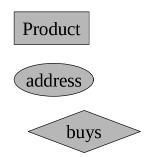
33

Entity / Relationship Diagrams in Summary

Entity sets:

Properties:

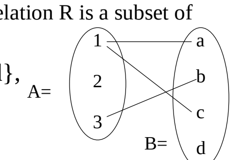
Relationships:



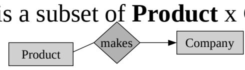
34

What is a Relation ?

- A mathematical definition:
  - if A, B are sets, then a relation R is a subset of A x B
- A={1,2,3}, B={a,b,c,d},
  R = {(1,a), (1,c), (3,b)}



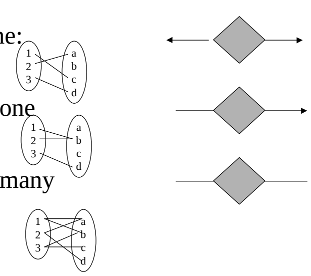
- **makes** is a subset of **Product** x **Company**:



35

Multiplicity of E/R Relations

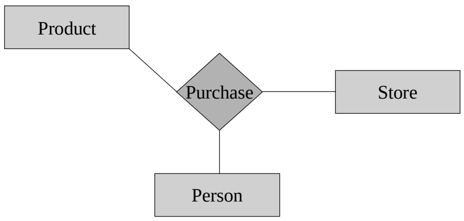
- one-one:
- many-one
- many-many



36

Multi-way Relationships

How do we model a purchase relationship between buyers, products and stores?

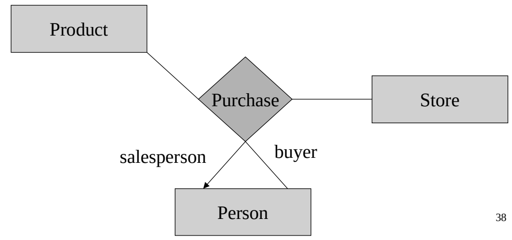


Can still model as a mathematical set (how ?)

37

Roles in Relationships

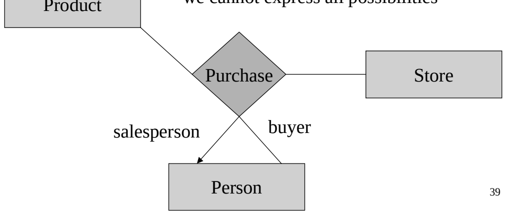
What if we need an entity set twice in one relationship?



38

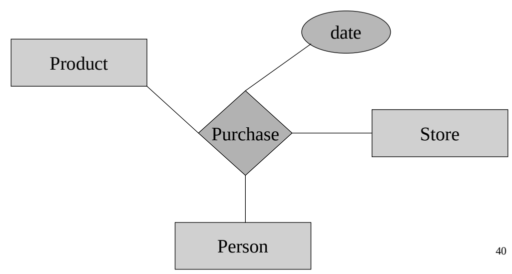
Roles in Relationships

Note the multiplicity of the relationships:
we cannot express all possibilities



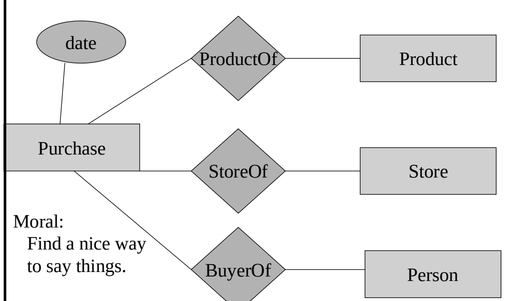
39

Attributes on Relationships



40

Converting Multi-way Relationships to Binary

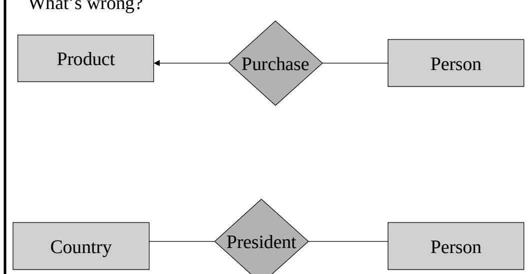


Moral:
Find a nice way
to say things.

41

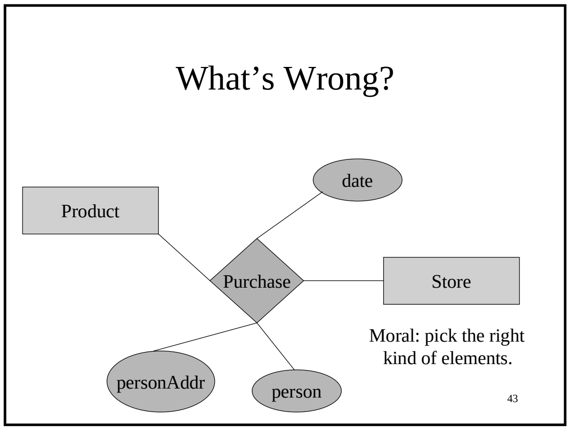
Design Principles

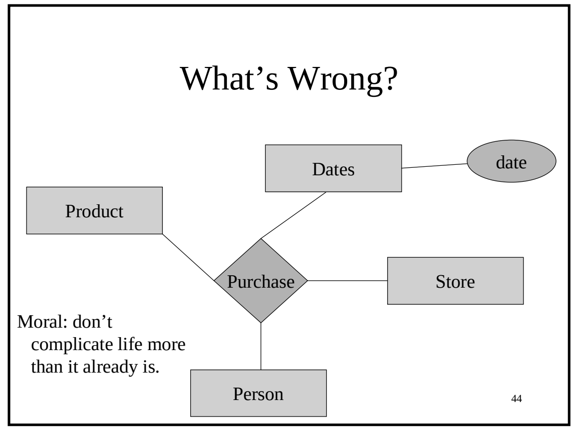
What's wrong?



Moral: be faithful!

42


What’s Wrong?
date
Product
Purchase
Store
Moral: pick the right kind of elements.
personAddr
person
43


What’s Wrong?
Dates
date
Product
Purchase
Store
Moral: don’t complicate life more than it already is.
Person
44