Introduction to Data Management
CSE 344

Unit 2: The Relational Data Model
SQL
Relational Algebra
Datalog

(9 lectures*)

*Slides may change: refresh each lecture
Introduction to Data Management
CSE 344

Lecture 2: Data Models
Class Overview

• Unit 1: Intro
• Unit 2: Relational Data Models and Query Languages
  – Data models, SQL RA, Datalog
• Unit 3: Non-relational data
• Unit 4: RDMBS internals and query optimization
• Unit 5: Parallel query processing
• Unit 6: DBMS usability, conceptual design
• Unit 7: Transactions
• Unit 8: Advanced topics (time permitting)
Review

• What is a database?
  – A collection of files storing related data

• What is a DBMS?
  – An application program that allows us to manage efficiently the collection of data files
Data Models

• Recall our example: want to design a database of books:
  – author, title, publisher, pub date, price, etc
  – How should we describe this data?

• **Data model** = mathematical formalism (or conceptual way) for describing the data
Data Models

- Relational
  - Data represented as relations

- Semi-structured (JSON)
  - Data represented as trees

- Key-value pairs
  - Used by NoSQL systems

- Graph

- Object-oriented
Example: storing FB friends

As a graph

<table>
<thead>
<tr>
<th>Person1</th>
<th>Person2</th>
<th>is_friend</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peter</td>
<td>John</td>
<td>1</td>
</tr>
<tr>
<td>John</td>
<td>Mary</td>
<td>0</td>
</tr>
<tr>
<td>Mary</td>
<td>Phil</td>
<td>1</td>
</tr>
<tr>
<td>Phil</td>
<td>Peter</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

As a relation

We will learn the tradeoffs of different data models later this quarter
3 Elements of Data Models

• Instance
  – The actual data
• Schema
  – Describe what data is being stored
• Query language
  – How to retrieve and manipulate data
Turing Awards in Data Management

Charles Bachman, 1973
*IDS and CODASYL*

Ted Codd, 1981
*Relational model*

Jim Gray, 1998
*Transaction processing*

Michael Stonebraker, 2014
*INGRES and Postgres*
Relational Model

- Data is a collection of relations / tables:

<table>
<thead>
<tr>
<th>cname</th>
<th>country</th>
<th>no_employees</th>
<th>for_profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>GizmoWorks</td>
<td>USA</td>
<td>20000</td>
<td>True</td>
</tr>
<tr>
<td>Canon</td>
<td>Japan</td>
<td>50000</td>
<td>True</td>
</tr>
<tr>
<td>Hitachi</td>
<td>Japan</td>
<td>30000</td>
<td>True</td>
</tr>
<tr>
<td>HappyCam</td>
<td>Canada</td>
<td>500</td>
<td>False</td>
</tr>
</tbody>
</table>

- mathematically, relation is a set of tuples
  - each tuple appears 0 or 1 times in the table
  - order of the rows is unspecified
The Relational Data Model

• Each attribute has a type. E.g.
  – Strings: CHAR(20), VARCHAR(50), TEXT
  – Numbers: INT, SMALLINT, FLOAT
  – MONEY, DATETIME, …
  – Few more that are vendor specific

• Types statically and strictly enforced

• #Attributes= “degree” (arity) of a relation
Keys

• Key = one (or multiple) attributes that uniquely identify a record
Keys

- Key = one (or multiple) attributes that uniquely identify a record

<table>
<thead>
<tr>
<th>cname</th>
<th>country</th>
<th>no_employees</th>
<th>for_profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>GizmoWorks</td>
<td>USA</td>
<td>20000</td>
<td>True</td>
</tr>
<tr>
<td>Canon</td>
<td>Japan</td>
<td>50000</td>
<td>True</td>
</tr>
<tr>
<td>Hitachi</td>
<td>Japan</td>
<td>30000</td>
<td>True</td>
</tr>
<tr>
<td>HappyCam</td>
<td>Canada</td>
<td>500</td>
<td>False</td>
</tr>
</tbody>
</table>
Keys

• Key = one (or multiple) attributes that uniquely identify a record

<table>
<thead>
<tr>
<th>cname</th>
<th>country</th>
<th>no_employees</th>
<th>for_profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>GizmoWorks</td>
<td>USA</td>
<td>20000</td>
<td>True</td>
</tr>
<tr>
<td>Canon</td>
<td>Japan</td>
<td>50000</td>
<td>True</td>
</tr>
<tr>
<td>Hitachi</td>
<td>Japan</td>
<td>30000</td>
<td>True</td>
</tr>
<tr>
<td>HappyCam</td>
<td>Canada</td>
<td>500</td>
<td>False</td>
</tr>
</tbody>
</table>
Keys

- Key = one (or multiple) attributes that uniquely identify a record

<table>
<thead>
<tr>
<th>cname</th>
<th>country</th>
<th>no_employees</th>
<th>for_profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>GizmoWorks</td>
<td>USA</td>
<td>20000</td>
<td>True</td>
</tr>
<tr>
<td>Canon</td>
<td>Japan</td>
<td>50000</td>
<td>True</td>
</tr>
<tr>
<td>Hitachi</td>
<td>Japan</td>
<td>30000</td>
<td>True</td>
</tr>
<tr>
<td>HappyCam</td>
<td>Canada</td>
<td>500</td>
<td>False</td>
</tr>
</tbody>
</table>
Keys

• Key = one (or multiple) attributes that uniquely identify a record

<table>
<thead>
<tr>
<th>cname</th>
<th>country</th>
<th>no_employees</th>
<th>for_profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>GizmoWorks</td>
<td>USA</td>
<td>20000</td>
<td>True</td>
</tr>
<tr>
<td>Canon</td>
<td>Japan</td>
<td>50000</td>
<td>True</td>
</tr>
<tr>
<td>Hitachi</td>
<td>Japan</td>
<td>30000</td>
<td>True</td>
</tr>
<tr>
<td>HappyCam</td>
<td>Canada</td>
<td>500</td>
<td>False</td>
</tr>
</tbody>
</table>
Multi-attribute Key

Key = fName,lName
(what does this mean?)

<table>
<thead>
<tr>
<th>fName</th>
<th>lName</th>
<th>Income</th>
<th>Department</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>Smith</td>
<td>20000</td>
<td>Testing</td>
</tr>
<tr>
<td>Alice</td>
<td>Thompson</td>
<td>50000</td>
<td>Testing</td>
</tr>
<tr>
<td>Bob</td>
<td>Thompson</td>
<td>30000</td>
<td>SW</td>
</tr>
<tr>
<td>Carol</td>
<td>Smith</td>
<td>50000</td>
<td>Testing</td>
</tr>
</tbody>
</table>
## Multiple Keys

We can choose one key and designate it as *primary key*

E.g.: primary key = SSN
Foreign Key

Company(cname, country, no_employees, for_profit)
Country(name, population)

<table>
<thead>
<tr>
<th>Company</th>
<th>Country</th>
<th>no_employees</th>
<th>for_profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canon</td>
<td>Japan</td>
<td>50000</td>
<td>Y</td>
</tr>
<tr>
<td>Hitachi</td>
<td>Japan</td>
<td>30000</td>
<td>Y</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Country</th>
<th>population</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>320M</td>
</tr>
<tr>
<td>Japan</td>
<td>127M</td>
</tr>
</tbody>
</table>
Keys: Summary

• Key = columns that uniquely identify tuple
  – Usually we underline
  – A relation can have many keys, but only one can be chosen as primary key

• Foreign key:
  – Attribute(s) whose value is a key of a record in some other relation
  – Foreign keys are sometimes called semantic pointer
Query Language

- **SQL**
  - **Structured Query Language**
  - Developed by IBM in the 70s
  - Most widely used language to query relational data

- Other relational query languages
  - Datalog, relational algebra
Our First DBMS

- SQL Lite
- Will switch to SQL Server later in the quarter
Demo 1
Discussion

• Tables are NOT ordered
  – they are sets or multisets (bags)
• Tables are FLAT
  – No nested attributes
• Tables DO NOT prescribe how they are implemented / stored on disk
  – This is called physical data independence
Table Implementation

• How would you implement this?

<table>
<thead>
<tr>
<th>cname</th>
<th>country</th>
<th>no_employees</th>
<th>for_profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>GizmoWorks</td>
<td>USA</td>
<td>20000</td>
<td>True</td>
</tr>
<tr>
<td>Canon</td>
<td>Japan</td>
<td>50000</td>
<td>True</td>
</tr>
<tr>
<td>Hitachi</td>
<td>Japan</td>
<td>30000</td>
<td>True</td>
</tr>
<tr>
<td>HappyCam</td>
<td>Canada</td>
<td>500</td>
<td>False</td>
</tr>
</tbody>
</table>
### Table Implementation

- **How would you implement this?**

<table>
<thead>
<tr>
<th>cname</th>
<th>country</th>
<th>no_employees</th>
<th>for_profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>GizmoWorks</td>
<td>USA</td>
<td>20000</td>
<td>True</td>
</tr>
<tr>
<td>Canon</td>
<td>Japan</td>
<td>50000</td>
<td>True</td>
</tr>
<tr>
<td>Hitachi</td>
<td>Japan</td>
<td>30000</td>
<td>True</td>
</tr>
<tr>
<td>HappyCam</td>
<td>Canada</td>
<td>500</td>
<td>False</td>
</tr>
</tbody>
</table>

**Row major: as an array of objects**

| GizmoWorks | USA | 20000 | True |
| Canon      | Japan | 50000 | True |
| Hitachi    | Japan | 30000 | True |
| HappyCam   | Canada | 500   | False |
**Table Implementation**

- How would you implement this?

<table>
<thead>
<tr>
<th>cname</th>
<th>country</th>
<th>no_employees</th>
<th>for_profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>GizmoWorks</td>
<td>USA</td>
<td>20000</td>
<td>True</td>
</tr>
<tr>
<td>Canon</td>
<td>Japan</td>
<td>50000</td>
<td>True</td>
</tr>
<tr>
<td>Hitachi</td>
<td>Japan</td>
<td>30000</td>
<td>True</td>
</tr>
<tr>
<td>HappyCam</td>
<td>Canada</td>
<td>500</td>
<td>False</td>
</tr>
</tbody>
</table>

**Column major: as one array per attribute**

<table>
<thead>
<tr>
<th>GizmoWorks</th>
<th>Canon</th>
<th>Hitachi</th>
<th>HappyCam</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>Japan</td>
<td>Japan</td>
<td>Canada</td>
</tr>
<tr>
<td>20000</td>
<td>50000</td>
<td>30000</td>
<td>500</td>
</tr>
<tr>
<td>True</td>
<td>True</td>
<td>True</td>
<td>False</td>
</tr>
</tbody>
</table>
Table Implementation

- How would you implement this?

<table>
<thead>
<tr>
<th>cname</th>
<th>country</th>
<th>no_employees</th>
<th>for_profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>GizmoWorks</td>
<td>USA</td>
<td>20000</td>
<td>True</td>
</tr>
<tr>
<td>Canon</td>
<td>Japan</td>
<td>50000</td>
<td>True</td>
</tr>
<tr>
<td>Hitachi</td>
<td>Japan</td>
<td>30000</td>
<td>True</td>
</tr>
<tr>
<td>HappyCam</td>
<td>Canada</td>
<td>500</td>
<td>False</td>
</tr>
</tbody>
</table>

Physical data independence

The logical definition of the data remains unchanged, even when we make changes to the actual implementation.
First Normal Form

- All relations must be flat: we say that the relation is in first normal form

<table>
<thead>
<tr>
<th>cname</th>
<th>country</th>
<th>no_employees</th>
<th>for_profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canon</td>
<td>Japan</td>
<td>50000</td>
<td>Y</td>
</tr>
<tr>
<td>Hitachi</td>
<td>Japan</td>
<td>30000</td>
<td>Y</td>
</tr>
</tbody>
</table>
All relations must be flat: we say that the relation is in *first normal form*

E.g. we want to add products manufactured by each company:
First Normal Form

• All relations must be flat: we say that the relation is in *first normal form*

• E.g. we want to add products manufactured by each company:

<table>
<thead>
<tr>
<th>cname</th>
<th>country</th>
<th>no_employees</th>
<th>for_profit</th>
<th>products</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canon</td>
<td>Japan</td>
<td>50000</td>
<td>Y</td>
<td><img src="image" alt="Product Table" /></td>
</tr>
<tr>
<td>Hitachi</td>
<td>Japan</td>
<td>30000</td>
<td>Y</td>
<td><img src="image" alt="Product Table" /></td>
</tr>
</tbody>
</table>
First Normal Form

- All relations must be flat: we say that the relation is in *first normal form*
- E.g. we want to add products manufactured by each company:

<table>
<thead>
<tr>
<th>cname</th>
<th>country</th>
<th>no_employees</th>
<th>for_profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canon</td>
<td>Japan</td>
<td>50000</td>
<td>Y</td>
</tr>
<tr>
<td>Hitachi</td>
<td>Japan</td>
<td>30000</td>
<td>Y</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>products</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>pname</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SingleTouch</td>
<td>149.99</td>
<td>Photography</td>
<td></td>
</tr>
<tr>
<td>Gadget</td>
<td>200</td>
<td>Toy</td>
<td></td>
</tr>
<tr>
<td>pname</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AC</td>
<td>300</td>
<td>Appliance</td>
<td></td>
</tr>
</tbody>
</table>

Non-1NF!
First Normal Form

**Company**

<table>
<thead>
<tr>
<th>cname</th>
<th>country</th>
<th>no_employees</th>
<th>for_profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canon</td>
<td>Japan</td>
<td>50000</td>
<td>Y</td>
</tr>
<tr>
<td>Hitachi</td>
<td>Japan</td>
<td>30000</td>
<td>Y</td>
</tr>
</tbody>
</table>

**Products**

<table>
<thead>
<tr>
<th>pname</th>
<th>price</th>
<th>category</th>
<th>manufacturer</th>
</tr>
</thead>
<tbody>
<tr>
<td>SingleTouch</td>
<td>149.99</td>
<td>Photography</td>
<td>Canon</td>
</tr>
<tr>
<td>AC</td>
<td>300</td>
<td>Appliance</td>
<td>Hitachi</td>
</tr>
<tr>
<td>Gadget</td>
<td>200</td>
<td>Toy</td>
<td>Canon</td>
</tr>
</tbody>
</table>
Data Models: Summary

- Schema + Instance + Query language
- Relational model:
  - Database = collection of tables
  - Each table is flat: “first normal form”
  - Key: may consists of multiple attributes
  - Foreign key: “semantic pointer”
  - Physical data independence
Lecture 3: SQL Basics
Review

• Relational data model
  – Schema+instance+query language

• Query language: SQL
  – Create tables
  – Retrieve records from tables
  – Declare keys and foreign keys
Review

• Tables are NOT ordered
  – they are sets or multisets (bags)
  – arity: # of attributes in a relation
  – cardinality: # of records in a relation

• Tables are FLAT
  – No nested attributes

• Tables DO NOT prescribe how they are implemented / stored on disk
  – This is called **physical data independence**
SQL

- **Structured Query Language**
- Most widely used language to query relational data
- One of the many languages for querying relational data

- A *declarative* programming language
Selections in SQL

```
SELECT * 
FROM Product 
WHERE price > 100.0
```
Demo 2
Retrieve all Japanese products that cost < $150

<table>
<thead>
<tr>
<th>pname</th>
<th>price</th>
<th>category</th>
<th>manufacturer</th>
</tr>
</thead>
<tbody>
<tr>
<td>MultiTouch</td>
<td>199.99</td>
<td>gadget</td>
<td>Canon</td>
</tr>
<tr>
<td>SingleTouch</td>
<td>49.99</td>
<td>photography</td>
<td>Canon</td>
</tr>
<tr>
<td>Gizom</td>
<td>50</td>
<td>gadget</td>
<td>GizmoWorks</td>
</tr>
<tr>
<td>SuperGizmo</td>
<td>250.00</td>
<td>gadget</td>
<td>GizmoWorks</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>cname</th>
<th>country</th>
</tr>
</thead>
<tbody>
<tr>
<td>GizmoWorks</td>
<td>USA</td>
</tr>
<tr>
<td>Canon</td>
<td>Japan</td>
</tr>
<tr>
<td>Hitachi</td>
<td>Japan</td>
</tr>
</tbody>
</table>
Retrieve all Japanese products that cost < $150

```
SELECT  pname, price
FROM     Product, Company
WHERE    ...
```
Retrieve all Japanese products that cost < $150

```sql
SELECT pname, price
FROM Product, Company
WHERE manufacturer = cname AND country = 'Japan' AND price < 150
```
Product\((\text{pname}, \text{price}, \text{category}, \text{manufacturer})\)
Company\((\text{cname}, \text{country})\)

**Joins in SQL**

<table>
<thead>
<tr>
<th>pname</th>
<th>price</th>
<th>category</th>
<th>manufacturer</th>
</tr>
</thead>
<tbody>
<tr>
<td>MultiTouch</td>
<td>199.99</td>
<td>gadget</td>
<td>Canon</td>
</tr>
<tr>
<td>SingleTouch</td>
<td>49.99</td>
<td>photography</td>
<td>Canon</td>
</tr>
<tr>
<td>Gizom</td>
<td>50</td>
<td>gadget</td>
<td>GizmoWorks</td>
</tr>
<tr>
<td>SuperGizmo</td>
<td>250.00</td>
<td>gadget</td>
<td>GizmoWorks</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>cname</th>
<th>country</th>
</tr>
</thead>
<tbody>
<tr>
<td>GizmoWorks</td>
<td>USA</td>
</tr>
<tr>
<td>Canon</td>
<td>Japan</td>
</tr>
<tr>
<td>Hitachi</td>
<td>Japan</td>
</tr>
</tbody>
</table>

Retrieve all USA companies that manufacture “gadget” products.
Product(pname, price, category, manufacturer)
Company(cname, country)

## Joins in SQL

<table>
<thead>
<tr>
<th>pname</th>
<th>price</th>
<th>category</th>
<th>manufacturer</th>
</tr>
</thead>
<tbody>
<tr>
<td>MultiTouch</td>
<td>199.99</td>
<td>gadget</td>
<td>Canon</td>
</tr>
<tr>
<td>SingleTouch</td>
<td>49.99</td>
<td>photography</td>
<td>Canon</td>
</tr>
<tr>
<td>Gizom</td>
<td>50</td>
<td>gadget</td>
<td>GizmoWorks</td>
</tr>
<tr>
<td>SuperGizmo</td>
<td>250.00</td>
<td>gadget</td>
<td>GizmoWorks</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>cname</th>
<th>country</th>
</tr>
</thead>
<tbody>
<tr>
<td>GizmoWorks</td>
<td>USA</td>
</tr>
<tr>
<td>Canon</td>
<td>Japan</td>
</tr>
<tr>
<td>Hitachi</td>
<td>Japan</td>
</tr>
</tbody>
</table>

Retrieve all USA companies that manufacture “gadget” products

```sql
SELECT DISTINCT cname
FROM Product, Company
WHERE country='USA' AND category = 'gadget'
    AND manufacturer = cname
```

Why DISTINCT?
Joins in SQL

• The standard join in SQL is sometimes called an **inner join**
  – Each row in the result **must come from both tables in the join**

• Sometimes we want to include rows from only one of the two table: **outer join**
Employee(id, name)
Sales(employeeID, productID)

**Inner Join**

<table>
<thead>
<tr>
<th>id</th>
<th>name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Joe</td>
</tr>
<tr>
<td>2</td>
<td>Jack</td>
</tr>
<tr>
<td>3</td>
<td>Jill</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>employeeID</th>
<th>productID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>344</td>
</tr>
<tr>
<td>1</td>
<td>355</td>
</tr>
<tr>
<td>2</td>
<td>544</td>
</tr>
</tbody>
</table>

Retrieve employees and their sales
Inner Join

Employee

<table>
<thead>
<tr>
<th>id</th>
<th>name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Joe</td>
</tr>
<tr>
<td>2</td>
<td>Jack</td>
</tr>
<tr>
<td>3</td>
<td>Jill</td>
</tr>
</tbody>
</table>

Sales

<table>
<thead>
<tr>
<th>employeeID</th>
<th>productID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>344</td>
</tr>
<tr>
<td>2</td>
<td>544</td>
</tr>
</tbody>
</table>

Retrieve employees and their sales

```
SELECT * 
FROM Employee E, Sales S 
WHERE E.id = S.employeeID
```
Employee(id, name)
Sales(employeeID, productID)

Inner Join

<table>
<thead>
<tr>
<th>id</th>
<th>name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Joe</td>
</tr>
<tr>
<td>2</td>
<td>Jack</td>
</tr>
<tr>
<td>3</td>
<td>Jill</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>employeeID</th>
<th>productID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>344</td>
</tr>
<tr>
<td>1</td>
<td>355</td>
</tr>
<tr>
<td>2</td>
<td>544</td>
</tr>
</tbody>
</table>

Retrieve employees and their sales

```sql
SELECT * 
FROM Employee E, Sales S 
WHERE E.id = S.employeeID
```
Employee(id, name)
Sales(employeeID, productID)

Inner Join

<table>
<thead>
<tr>
<th>id</th>
<th>name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Joe</td>
</tr>
<tr>
<td>2</td>
<td>Jack</td>
</tr>
<tr>
<td>3</td>
<td>Jill</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>employeeID</th>
<th>productID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>344</td>
</tr>
<tr>
<td>1</td>
<td>355</td>
</tr>
<tr>
<td>2</td>
<td>544</td>
</tr>
</tbody>
</table>

Retrieve employees and their sales

SELECT *
FROM Employee E, Sales S
WHERE E.id = S.employeeID

Jill is missing
Inner Join

Employee(id, name)
Sales(employeeID, productID)

<table>
<thead>
<tr>
<th>id</th>
<th>name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Joe</td>
</tr>
<tr>
<td>2</td>
<td>Jack</td>
</tr>
<tr>
<td>3</td>
<td>Jill</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>employeeID</th>
<th>productID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>344</td>
</tr>
<tr>
<td>1</td>
<td>355</td>
</tr>
<tr>
<td>2</td>
<td>544</td>
</tr>
</tbody>
</table>

Retrieve employees and their sales

SELECT * FROM Employee E INNER JOIN Sales S ON E.id = S.employeeID

Jill is missing

Alternative syntax

<table>
<thead>
<tr>
<th>id</th>
<th>name</th>
<th>employeeID</th>
<th>productID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Joe</td>
<td>1</td>
<td>344</td>
</tr>
<tr>
<td>1</td>
<td>Joe</td>
<td>1</td>
<td>355</td>
</tr>
<tr>
<td>2</td>
<td>Jack</td>
<td>2</td>
<td>544</td>
</tr>
</tbody>
</table>
Employee(id, name)  
Sales(employeeID, productID)

### Outer Join

<table>
<thead>
<tr>
<th></th>
<th>id</th>
<th>name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Joe</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Jack</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Jill</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>employeeID</th>
<th>productID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>344</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>355</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>544</td>
<td></td>
</tr>
</tbody>
</table>

Retrieve employees and their sales

```sql
SELECT *  
FROM Employee E  
LEFT OUTER JOIN Sales S  
on E.id = S.employeeID
```
Introduction to Data Management
CSE 344

Lecture 4: Joins and Aggregates
Review: Our SQL Toolchest

• Selection
• Projection
• Ordering and distinct

• Inner Join
• Outer Join
(Inner) joins

Product(pname, price, category, manufacturer)
Company(cname, country)

manufacturer = foreign key to Company.cname

Return all companies in the ‘USA’ that manufacture some product in the ‘gadget’ category.
(Inner) joins

```sql
SELECT DISTINCT cname
FROM Product, Company
WHERE country='USA' AND category = 'gadget'
AND manufacturer = cname
```
INNER joins

```
SELECT DISTINCT cname
FROM Product, Company
WHERE country='USA' AND category = 'gadget'
AND manufacturer = cname
```

<table>
<thead>
<tr>
<th>Product</th>
<th>Company</th>
</tr>
</thead>
<tbody>
<tr>
<td>pname</td>
<td>cname</td>
</tr>
<tr>
<td>category</td>
<td>country</td>
</tr>
<tr>
<td>Gizmo</td>
<td>GizmoWorks</td>
</tr>
<tr>
<td>Camera</td>
<td>Hitachi</td>
</tr>
<tr>
<td>OneClick</td>
<td>Hitachi</td>
</tr>
</tbody>
</table>
(Inner) joins

```
SELECT DISTINCT cname
FROM Product, Company
WHERE country = 'USA' AND category = 'gadget'
AND manufacturer = cname
```
INNER joins

SELECT DISTINCT cname
FROM Product, Company
WHERE country='USA' AND category = 'gadget'
AND manufacturer = cname

Product

<table>
<thead>
<tr>
<th>pname</th>
<th>category</th>
<th>manufacturer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gizmo</td>
<td>gadget</td>
<td>GizmoWorks</td>
</tr>
<tr>
<td>Camera</td>
<td>Photo</td>
<td>Hitachi</td>
</tr>
<tr>
<td>OneClick</td>
<td>Photo</td>
<td>Hitachi</td>
</tr>
</tbody>
</table>

Company

<table>
<thead>
<tr>
<th>cname</th>
<th>country</th>
</tr>
</thead>
<tbody>
<tr>
<td>GizmoWorks</td>
<td>USA</td>
</tr>
<tr>
<td>Canon</td>
<td>Japan</td>
</tr>
<tr>
<td>Hitachi</td>
<td>Japan</td>
</tr>
</tbody>
</table>
(Inner) joins

```
SELECT DISTINCT c.name
FROM Product, Company
WHERE country='USA' AND category = 'gadget'
AND manufacturer = c.name
```

**Product**

<table>
<thead>
<tr>
<th>pname</th>
<th>category</th>
<th>manufacturer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gizmo</td>
<td>gadget</td>
<td>GizmoWorks</td>
</tr>
<tr>
<td>Camera</td>
<td>Photo</td>
<td>Hitachi</td>
</tr>
<tr>
<td>OneClick</td>
<td>Photo</td>
<td>Hitachi</td>
</tr>
</tbody>
</table>

**Company**

<table>
<thead>
<tr>
<th>cname</th>
<th>country</th>
</tr>
</thead>
<tbody>
<tr>
<td>GizmoWorks</td>
<td>USA</td>
</tr>
<tr>
<td>Canon</td>
<td>Japan</td>
</tr>
<tr>
<td>Hitachi</td>
<td>Japan</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>pname</th>
<th>category</th>
<th>manufacturer</th>
<th>cname</th>
<th>country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gizmo</td>
<td>gadget</td>
<td>GizmoWorks</td>
<td>GizmoWorks</td>
<td>USA</td>
</tr>
</tbody>
</table>
INNER joins

```
SELECT DISTINCT cname
FROM Product, Company
WHERE country = 'USA' AND category = 'gadget'
AND manufacturer = cname
```
(Inner) joins

```
SELECT DISTINCT cname
FROM Product, Company
WHERE country='USA' AND category = 'gadget'
AND manufacturer = cname
```
(Inner) joins

```
SELECT DISTINCT cname
FROM Product, Company
WHERE country = 'USA' AND category = 'gadget'
AND manufacturer = cname
```
**Inner joins**

```sql
SELECT DISTINCT cname
FROM Product, Company
WHERE country='USA' AND category = 'gadget'
AND manufacturer = cname
```

### Product

<table>
<thead>
<tr>
<th>pname</th>
<th>category</th>
<th>manufacturer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gizmo</td>
<td>gadget</td>
<td>GizmoWorks</td>
</tr>
<tr>
<td>Camera</td>
<td>Photo</td>
<td>Hitachi</td>
</tr>
<tr>
<td>OneClick</td>
<td>Photo</td>
<td>Hitachi</td>
</tr>
</tbody>
</table>

### Company

<table>
<thead>
<tr>
<th>cname</th>
<th>country</th>
</tr>
</thead>
<tbody>
<tr>
<td>GizmoWorks</td>
<td>USA</td>
</tr>
<tr>
<td>Canon</td>
<td>Japan</td>
</tr>
<tr>
<td>Hitachi</td>
<td>Japan</td>
</tr>
</tbody>
</table>
(Inner) joins

```sql
SELECT DISTINCT cname
FROM Product, Company
WHERE country='USA' AND category = 'gadget'
    AND manufacturer = cname
```

<table>
<thead>
<tr>
<th>Product</th>
<th>Company</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>pname</strong></td>
<td><strong>category</strong></td>
</tr>
<tr>
<td>Gizmo</td>
<td>gadget</td>
</tr>
<tr>
<td>Camera</td>
<td>Photo</td>
</tr>
<tr>
<td>OneClick</td>
<td>Photo</td>
</tr>
</tbody>
</table>
(Inner) joins

```
SELECT DISTINCT cname
FROM Product, Company
WHERE country='USA' AND category = 'gadget'
AND manufacturer = cname
```

**Product**

<table>
<thead>
<tr>
<th>pname</th>
<th>category</th>
<th>manufacturer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gizmo</td>
<td>gadget</td>
<td>GizmoWorks</td>
</tr>
<tr>
<td>Camera</td>
<td>Photo</td>
<td>Hitachi</td>
</tr>
<tr>
<td>OneClick</td>
<td>Photo</td>
<td>Hitachi</td>
</tr>
</tbody>
</table>

**Company**

<table>
<thead>
<tr>
<th>cname</th>
<th>country</th>
</tr>
</thead>
<tbody>
<tr>
<td>GizmoWorks</td>
<td>USA</td>
</tr>
<tr>
<td>Canon</td>
<td>Japan</td>
</tr>
<tr>
<td>Hitachi</td>
<td>Japan</td>
</tr>
</tbody>
</table>

And son on...
(Inner) joins

```
SELECT DISTINCT cname
FROM   Product, Company
WHERE  country='USA' AND category = 'gadget'
       AND manufacturer = cname
```

```
SELECT DISTINCT cname
FROM   Product JOIN Company ON
       country = 'USA' AND category = 'gadget'
       AND manufacturer = cname
```
(Inner) joins

```
SELECT DISTINCT cname
FROM   Product, Company
WHERE  country='USA' AND category = 'gadget'
       AND manufacturer = cname
```

```
SELECT DISTINCT cname
FROM   Product
JOIN   Company
       ON country = 'USA'
       AND category = 'gadget'
       AND manufacturer = cname
```

Same thing, different syntax
(Inner) Joins

for x1 in R1:
    for x2 in R2:
        ...
        for xm in Rm:
            if Cond(x1, x2...):
                output(x1.a1, x2.a2, ... xm.am)

This is called nested loop semantics since we are interpreting what a join means using a nested loop
Another example

Retrieve all USA companies that manufacture products in both ‘gadget’ and ‘photography’ categories
Another example

Retrieve all USA companies that manufacture products in both ‘gadget’ and ‘photography’ categories

```
SELECT DISTINCT z.cname
FROM Product x, Company z
WHERE z.country = 'USA'
  AND x.manufacturer = z.cname
  AND x.category = 'gadget'
  AND x.category = 'photography';
```
Another example

Retrieve all USA companies that manufacture products in both ‘gadget’ and ‘photography’ categories

```
SELECT DISTINCT z.cname
FROM Product x, Company z
WHERE z.country = 'USA'
    AND x.manufacturer = z.cname
    AND x.category = 'gadget'
    AND x.category = 'photography';
```

Does this work?

Returns the empty set
Another example

Retrieve all USA companies that manufacture products in both ‘gadget’ and ‘photography’ categories

Product(pname, price, category, manufacturer)
Company(cname, country)

What about this?

SELECT DISTINCT z.cname
FROM Product x, Company z
WHERE z.country = 'USA'
  AND x.manufacturer = z.cname
  AND (x.category = 'gadget'
       OR x.category = 'photography');
Another example

Retrieve all USA companies that manufacture products in both ‘gadget’ and ‘photography’ categories

What about this?

```
SELECT DISTINCT z.cname
FROM Product x, Company z
WHERE z.country = 'USA'
  AND x.manufacturer = z.cname
  AND (x.category = 'gadget'
       OR x.category = 'photography');
```

Returns too much
Another example

Retrieve all USA companies that manufacture products in both ‘gadget’ and ‘photography’ categories

```
SELECT DISTINCT z.cname
FROM Product x, Product y, Company z
WHERE z.country = 'USA'
  AND x.manufacturer = z.cname
  AND y.manufacturer = z.cname
  AND x.category = 'gadget'
  AND y.category = 'photography';
```

Need to include Product twice!
Self-Joins and Tuple Variables

- Find USA companies that manufacture both products in the ‘gadgets’ and ‘photo’ category

- Joining Product with Company is insufficient: need to join Product, with Product, and with Company

- When a relation occurs twice in the FROM clause we call it a self-join; in that case we must use tuple variables (why?)
Self-joins

```
SELECT DISTINCT z.cname
FROM Product x, Product y, Company z
WHERE z.country = 'USA'
    AND x.category = 'gadget'
    AND y.category = 'photo'
    AND x.manufacturer = z.cname
    AND y.manufacturer = z.cname;
```
Self-joins

```
SELECT DISTINCT z.cname
FROM Product x, Product y, Company z
WHERE z.country = 'USA'
  AND x.category = 'gadget'
  AND y.category = 'photo'
  AND x.manufacturer = z.cname
  AND y.manufacturer = z.cname;
```
Self-joins

```
SELECT DISTINCT z.cname
FROM Product x, Product y, Company z
WHERE z.country = 'USA'
    AND x.category = 'gadget'
    AND y.category = 'photo'
    AND x.manufacturer = z.cname
    AND y.manufacturer = z.cname;
```

<table>
<thead>
<tr>
<th>Product</th>
<th>Company</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td></td>
</tr>
<tr>
<td>y</td>
<td></td>
</tr>
<tr>
<td><strong>pname</strong></td>
<td><strong>category</strong></td>
</tr>
<tr>
<td>Gizmo</td>
<td>gadget</td>
</tr>
<tr>
<td>SingleTouch</td>
<td>photo</td>
</tr>
<tr>
<td>MultiTouch</td>
<td>Photo</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>cname</th>
<th>country</th>
</tr>
</thead>
<tbody>
<tr>
<td>GizmoWorks</td>
<td>USA</td>
</tr>
<tr>
<td>Hitachi</td>
<td>Japan</td>
</tr>
</tbody>
</table>
SELECT DISTINCT z.cname
FROM Product x, Product y, Company z
WHERE z.country = 'USA'
    AND x.category = 'gadget'
    AND y.category = 'photo'
    AND x.manufacturer = z.cname
    AND y.manufacturer = z.cname;
Self-joins

```
SELECT DISTINCT z.cname
FROM Product x, Product y, Company z
WHERE z.country = 'USA'
AND x.category = 'gadget'
AND y.category = 'photo'
AND x.manufacturer = z.cname
AND y.manufacturer = z.cname;
```
Self-joins

SELECT DISTINCT z.cname
FROM Product x, Product y, Company z
WHERE z.country = 'USA' 
    AND x.category = 'gadget'
    AND y.category = 'photo'
    AND x.manufacturer = z.cname 
    AND y.manufacturer = z.cname;

Product

<table>
<thead>
<tr>
<th></th>
<th>pname</th>
<th>category</th>
<th>manufacturer</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>Gizmo</td>
<td>gadget</td>
<td>GizmoWorks</td>
</tr>
<tr>
<td></td>
<td>SingleTouch</td>
<td>photo</td>
<td>Hitachi</td>
</tr>
<tr>
<td>y</td>
<td>MultiTouch</td>
<td>Photo</td>
<td>GizmoWorks</td>
</tr>
</tbody>
</table>

Company

<table>
<thead>
<tr>
<th></th>
<th>cname</th>
<th>country</th>
</tr>
</thead>
<tbody>
<tr>
<td>z</td>
<td>GizmoWorks</td>
<td>USA</td>
</tr>
<tr>
<td></td>
<td>Hitachi</td>
<td>Japan</td>
</tr>
</tbody>
</table>
Self-joins

```
SELECT DISTINCT z.cname
FROM Product x, Product y, Company z
WHERE z.country = 'USA'
  AND x.category = 'gadget'
  AND y.category = 'photo'
  AND x.manufacturer = z.cname
  AND y.manufacturer = z.cname;
```
SELECT DISTINCT z.cname
FROM Product x, Product y, Company z
WHERE z.country = 'USA'
AND x.category = 'gadget'
AND y.category = 'photo'
AND x.manufacturer = z.cname
AND y.manufacturer = z.cname;
### Self-joins

```sql
SELECT DISTINCT z.cname
FROM Product x, Product y, Company z
WHERE z.country = 'USA'
  AND x.category = 'gadget'
  AND y.category = 'photo'
  AND x.manufacturer = z.cname
  AND y.manufacturer = z.cname;
```

<table>
<thead>
<tr>
<th>Product</th>
<th>Company</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>X</strong></td>
<td><strong>Y</strong></td>
</tr>
<tr>
<td><strong>Gizmo</strong></td>
<td><strong>GizmoWorks</strong></td>
</tr>
<tr>
<td><strong>SingleTouch</strong></td>
<td><strong>Hitachi</strong></td>
</tr>
<tr>
<td><strong>MultiTouch</strong></td>
<td><strong>GizmoWorks</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>X</strong>.pname</th>
<th><strong>X</strong>.category</th>
<th><strong>X</strong>.manufacturer</th>
<th><strong>Y</strong>.pname</th>
<th><strong>Y</strong>.category</th>
<th><strong>Y</strong>.manufacturer</th>
<th><strong>z</strong>.cname</th>
<th><strong>z</strong>.country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gizmo</td>
<td>gadget</td>
<td>GizmoWorks</td>
<td>MultiTouch</td>
<td>Photo</td>
<td>GizmoWorks</td>
<td>GizmoWorks</td>
<td>USA</td>
</tr>
<tr>
<td>SingleTouch</td>
<td>photo</td>
<td>Hitachi</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Japan</td>
</tr>
<tr>
<td>MultiTouch</td>
<td>Photo</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Self-joins

```sql
SELECT DISTINCT z.cname
FROM Product x, Product y, Company z
WHERE z.country = 'USA'
    AND x.category = 'gadget'
    AND y.category = 'photo'
    AND x.manufacturer = z.cname
    AND y.manufacturer = z.cname;
```

<table>
<thead>
<tr>
<th>Product X</th>
<th>Product Y</th>
<th>Company Z</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>pname</strong></td>
<td><strong>category</strong></td>
<td><strong>cname</strong></td>
</tr>
<tr>
<td>Gizmo</td>
<td>gadget</td>
<td>GizmoWorks</td>
</tr>
<tr>
<td>SingleTouch</td>
<td>photo</td>
<td>Hitachi</td>
</tr>
<tr>
<td><strong>MultiTouch</strong></td>
<td><strong>Photo</strong></td>
<td><strong>country</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>USA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Japan</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>x.pname</th>
<th>x.category</th>
<th>x.manufacturer</th>
<th>y.pname</th>
<th>y.category</th>
<th>y.manufacturer</th>
<th>z.cname</th>
<th>z.country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gizmo</td>
<td>gadget</td>
<td>GizmoWorks</td>
<td>MultiTouch</td>
<td>Photo</td>
<td>GizmoWorks</td>
<td>GizmoWorks</td>
<td>USA</td>
</tr>
</tbody>
</table>
Outer joins

Product(name, category)
Purchase(prodName, store)

-- prodName is foreign key

Retrieve all product names and the stores where they were purchased.
Include products that never sold.
Outer joins

Product(name, category)
Purchase(prodName, store)

-- prodName is foreign key

Retrieve all product names and the stores where they were purchased.
Include products that never sold.

```
SELECT Product.name, Purchase.store
FROM Product, Purchase
WHERE Product.name = Purchase.prodName
```
Outer joins

Product(name, category)
Purchase(prodName, store)

-- prodName is foreign key

Retrieve all product names and the stores where they were purchased.
Include products that never sold.

SELECT Product.name, Purchase.store
FROM Product, Purchase
WHERE Product.name = Purchase.prodName

Does not include products that never sold! (why?)
Outer joins

Product(name, category)
Purchase(prodName, store)

-- prodName is foreign key

```
SELECT Product.name, Purchase.store
FROM   Product, Purchase
WHERE  Product.name = Purchase.prodName
```
Outer joins

Product(name, category)
Purchase(prodName, store)

-- prodName is foreign key

```
SELECT Product.name, Purchase.store
FROM Product LEFT OUTER JOIN Purchase ON
Product.name = Purchase.prodName
```
```
SELECT Product.name, Purchase.store
FROM Product JOIN Purchase ON
    Product.name = Purchase.prodName
```
### SELECT Product.name, Purchase.store
FROM Product JOIN Purchase ON Product.name = Purchase.prodName

<table>
<thead>
<tr>
<th>Product</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gizmo</td>
<td>gadget</td>
</tr>
<tr>
<td>Camera</td>
<td>Photo</td>
</tr>
<tr>
<td>OneClick</td>
<td>Photo</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Purchase</th>
<th>Store</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gizmo</td>
<td>Wiz</td>
</tr>
<tr>
<td>Camera</td>
<td>Ritz</td>
</tr>
<tr>
<td>Camera</td>
<td>Wiz</td>
</tr>
</tbody>
</table>
SELECT Product.name, Purchase.store
FROM Product JOIN Purchase ON Product.name = Purchase.prodName

<table>
<thead>
<tr>
<th>Name</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gizmo</td>
<td>gadget</td>
</tr>
<tr>
<td>Camera</td>
<td>Photo</td>
</tr>
<tr>
<td>OneClick</td>
<td>Photo</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ProdName</th>
<th>Store</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gizmo</td>
<td>Wiz</td>
</tr>
<tr>
<td>Camera</td>
<td>Ritz</td>
</tr>
<tr>
<td>Camera</td>
<td>Wiz</td>
</tr>
</tbody>
</table>

Output

<table>
<thead>
<tr>
<th>Name</th>
<th>Store</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gizmo</td>
<td>Wiz</td>
</tr>
</tbody>
</table>
### SELECT Query

```
SELECT Product.name, Purchase.store
FROM Product JOIN Purchase ON
  Product.name = Purchase.prodName
```

### Product Table

<table>
<thead>
<tr>
<th>Name</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gizmo</td>
<td>gadget</td>
</tr>
<tr>
<td>Camera</td>
<td>Photo</td>
</tr>
<tr>
<td>OneClick</td>
<td>Photo</td>
</tr>
</tbody>
</table>

### Purchase Table

<table>
<thead>
<tr>
<th>ProdName</th>
<th>Store</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gizmo</td>
<td>Wiz</td>
</tr>
<tr>
<td>Camera</td>
<td>Ritz</td>
</tr>
<tr>
<td>Camera</td>
<td>Wiz</td>
</tr>
</tbody>
</table>

### Output Table

<table>
<thead>
<tr>
<th>Name</th>
<th>Store</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gizmo</td>
<td>Wiz</td>
</tr>
</tbody>
</table>
SELECT Product.name, Purchase.store
FROM Product JOIN Purchase ON Product.name = Purchase.prodName

Product

<table>
<thead>
<tr>
<th>Name</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gizmo</td>
<td>gadget</td>
</tr>
<tr>
<td>Camera</td>
<td>Photo</td>
</tr>
<tr>
<td>OneClick</td>
<td>Photo</td>
</tr>
</tbody>
</table>

Purchase

<table>
<thead>
<tr>
<th>ProdName</th>
<th>Store</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gizmo</td>
<td>Wiz</td>
</tr>
<tr>
<td>Camera</td>
<td>Ritz</td>
</tr>
<tr>
<td>Camera</td>
<td>Wiz</td>
</tr>
</tbody>
</table>

Output

<table>
<thead>
<tr>
<th>Name</th>
<th>Store</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gizmo</td>
<td>Wiz</td>
</tr>
</tbody>
</table>
```sql
SELECT Product.name, Purchase.store
FROM Product JOIN Purchase
ON Product.name = Purchase.prodName
```

<table>
<thead>
<tr>
<th>Name</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gizmo</td>
<td>gadget</td>
</tr>
<tr>
<td>Camera</td>
<td>Photo</td>
</tr>
<tr>
<td>OneClick</td>
<td>Photo</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ProdName</th>
<th>Store</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gizmo</td>
<td>Wiz</td>
</tr>
<tr>
<td>Camera</td>
<td>Ritz</td>
</tr>
<tr>
<td>Camera</td>
<td>Wiz</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Store</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gizmo</td>
<td>Wiz</td>
</tr>
</tbody>
</table>
### Product

<table>
<thead>
<tr>
<th>Name</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gizmo</td>
<td>gadget</td>
</tr>
<tr>
<td>Camera</td>
<td>Photo</td>
</tr>
<tr>
<td>OneClick</td>
<td>Photo</td>
</tr>
</tbody>
</table>

### Purchase

<table>
<thead>
<tr>
<th>ProdName</th>
<th>Store</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gizmo</td>
<td>Wiz</td>
</tr>
<tr>
<td>Camera</td>
<td>Ritz</td>
</tr>
</tbody>
</table>

### Output

<table>
<thead>
<tr>
<th>Name</th>
<th>Store</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gizmo</td>
<td>Wiz</td>
</tr>
<tr>
<td>Camera</td>
<td>Ritz</td>
</tr>
</tbody>
</table>
### Product

<table>
<thead>
<tr>
<th>Name</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gizmo</td>
<td>gadget</td>
</tr>
<tr>
<td>Camera</td>
<td>Photo</td>
</tr>
<tr>
<td>OneClick</td>
<td>Photo</td>
</tr>
</tbody>
</table>

### Purchase

<table>
<thead>
<tr>
<th>ProdName</th>
<th>Store</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gizmo</td>
<td>Wiz</td>
</tr>
<tr>
<td>Camera</td>
<td>Ritz</td>
</tr>
<tr>
<td>Camera</td>
<td>Wiz</td>
</tr>
</tbody>
</table>

### Output

<table>
<thead>
<tr>
<th>Name</th>
<th>Store</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gizmo</td>
<td>Wiz</td>
</tr>
<tr>
<td>Camera</td>
<td>Ritz</td>
</tr>
<tr>
<td>Camera</td>
<td>Wiz</td>
</tr>
</tbody>
</table>

```sql
SELECT Product.name, Purchase.store
FROM Product JOIN Purchase ON Product.name = Purchase.prodName
```
**Select** 

\[
\text{SELECT Product.name, Purchase.store} \\
\text{FROM Product JOIN Purchase ON} \\
\text{Product.name = Purchase.prodName}
\]

### Product

<table>
<thead>
<tr>
<th>Name</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gizmo</td>
<td>gadget</td>
</tr>
<tr>
<td>Camera</td>
<td>Photo</td>
</tr>
<tr>
<td>OneClick</td>
<td>Photo</td>
</tr>
</tbody>
</table>

### Purchase

<table>
<thead>
<tr>
<th>ProdName</th>
<th>Store</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gizmo</td>
<td>Wiz</td>
</tr>
<tr>
<td>Camera</td>
<td>Ritz</td>
</tr>
<tr>
<td>Camera</td>
<td>Wiz</td>
</tr>
</tbody>
</table>

### Output

<table>
<thead>
<tr>
<th>Name</th>
<th>Store</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gizmo</td>
<td>Wiz</td>
</tr>
<tr>
<td>Camera</td>
<td>Ritz</td>
</tr>
<tr>
<td>Camera</td>
<td>Wiz</td>
</tr>
</tbody>
</table>
```sql
SELECT Product.name, Purchase.store
FROM Product
LEFT OUTER JOIN Purchase
ON Product.name = Purchase.prodName
```
### Product

<table>
<thead>
<tr>
<th>Name</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gizmo</td>
<td>gadget</td>
</tr>
<tr>
<td>Camera</td>
<td>Photo</td>
</tr>
<tr>
<td>OneClick</td>
<td>Photo</td>
</tr>
</tbody>
</table>

### Purchase

<table>
<thead>
<tr>
<th>ProdName</th>
<th>Store</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gizmo</td>
<td>Wiz</td>
</tr>
<tr>
<td>Camera</td>
<td>Ritz</td>
</tr>
<tr>
<td>Camera</td>
<td>Wiz</td>
</tr>
</tbody>
</table>

### Output

<table>
<thead>
<tr>
<th>Name</th>
<th>Store</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gizmo</td>
<td>Wiz</td>
</tr>
<tr>
<td>Camera</td>
<td>Ritz</td>
</tr>
<tr>
<td>Camera</td>
<td>Wiz</td>
</tr>
<tr>
<td>OneClick</td>
<td>NULL</td>
</tr>
</tbody>
</table>

**SQL Query**

```sql
SELECT Product.name, Purchase.store
FROM Product
LEFT OUTER JOIN Purchase
ON Product.name = Purchase.prodName
```
### SELECT

```sql
SELECT Product.name, Purchase.store
FROM Product FULL OUTER JOIN Purchase ON
Product.name = Purchase.prodName
```

<table>
<thead>
<tr>
<th>Product</th>
<th>Category</th>
<th>Purchase</th>
<th>Store</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gizmo</td>
<td>gadget</td>
<td>Gizmo</td>
<td>Wiz</td>
</tr>
<tr>
<td>Camera</td>
<td>Photo</td>
<td>Camera</td>
<td>Ritz</td>
</tr>
<tr>
<td>Camera</td>
<td></td>
<td>Camera</td>
<td>Wiz</td>
</tr>
<tr>
<td>OneClick</td>
<td>Photo</td>
<td>OneClick</td>
<td>NULL</td>
</tr>
<tr>
<td>NULL</td>
<td></td>
<td>NULL</td>
<td>Foo</td>
</tr>
</tbody>
</table>
Outer Joins

```sql
(tableA (LEFT/RIGHT/FULL) OUTER JOIN tableB ON p)
```

- **Left outer join:**
  - Include tuples from `tableA` even if no match
- **Right outer join:**
  - Include tuples from `tableB` even if no match
- **Full outer join:**
  - Include tuples from both even if no match

- **In all cases:**
  - Patch tuples without matches using NULL
Other DBMSs have other ways of importing data

Specify a filename where the database will be stored

Loading Data into SQLite

>sqlite3 lecture04

sqlite> create table Purchase
    (pid int primary key,
     product text,
     price float,
     quantity int,
     month varchar(15));

sqlite> -- download data.txt
sqlite> .import lec04-data.txt Purchase
Comment about SQLite

• Cannot load NULL values such that they are actually loaded as null values

• So we need to use two steps:
  – Load null values using some type of special value
  – Update the special values to actual null values

```sql
update Purchase
  set price = null
where price = 'null'
```
Simple Aggregations

Five basic aggregate operations in SQL

```sql
select count(*) from Purchase
select sum(quantity) from Purchase
select avg(price) from Purchase
select max(quantity) from Purchase
select min(quantity) from Purchase
```

Except count, all aggregations apply to a single attribute.

count of an empty table is 0
Aggregates and NULL Values

Null values are not used in aggregates

```sql
insert into Purchase
values(12, 'gadget', NULL, NULL, NULL, 'april')
```

Let’s try the following

```sql
select count(*) from Purchase
select count(quantity) from Purchase
```

```sql
select sum(quantity) from Purchase
```

```sql
select count(*)
from Purchase
where quantity is not null;
```
COUNT applies to duplicates, unless otherwise stated:

```
SELECT count(product) 
FROM Purchase 
WHERE price > 4.99
```

same as count(*) if no nulls

We probably want:

```
SELECT count(DISTINCT product) 
FROM Purchase 
WHERE price > 4.99
```
More Examples

\[
\text{SELECT} \quad \text{Sum(price} \times \text{quantity)} \\
\text{FROM} \quad \text{Purchase}
\]

\[
\text{SELECT} \quad \text{Sum(price} \times \text{quantity)} \\
\text{FROM} \quad \text{Purchase} \\
\text{WHERE} \quad \text{product} = \text{‘bagel’}
\]
Introduction to Data Management
CSE 344

Lecture 5: Grouping and Query Evaluation
Announcement

• The Webquiz is due tonight!
Grouping and Aggregation

Purchase(product, price, quantity)

Find total quantities for all sales over $1, by product.
SELECT product, Sum(quantity) AS TotalSales
FROM Purchase
WHERE price > 1
GROUP BY product
Other Examples

Compare these two queries:

```
SELECT product, count(*)
FROM Purchase
GROUP BY product
```

```
SELECT month, count(*)
FROM Purchase
GROUP BY month
```
Other Examples

Compare these two queries:

```
SELECT product, count(*)
FROM Purchase
GROUP BY product
```

```
SELECT month, count(*)
FROM Purchase
GROUP BY month
```

One answer for each product.  One answer for each month.
Other Examples

SELECT product,
    sum(quantity) AS SumQuantity,
    max(price) AS MaxPrice
FROM Purchase
GROUP BY product

Multiple aggregates OK
Need to be Careful…

```
SELECT product, 
    max(quantity)
FROM   Purchase
GROUP BY product
```

<table>
<thead>
<tr>
<th>Product</th>
<th>Price</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bagel</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>Bagel</td>
<td>1.50</td>
<td>20</td>
</tr>
<tr>
<td>Banana</td>
<td>0.5</td>
<td>50</td>
</tr>
<tr>
<td>Banana</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>Banana</td>
<td>4</td>
<td>10</td>
</tr>
</tbody>
</table>
Need to be Careful...

```
SELECT product, max(quantity)
FROM Purchase
GROUP BY product
```

<table>
<thead>
<tr>
<th>Product</th>
<th>Price</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bagel</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>Bagel</td>
<td>1.50</td>
<td>20</td>
</tr>
<tr>
<td>Banana</td>
<td>0.5</td>
<td>50</td>
</tr>
<tr>
<td>Banana</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>Banana</td>
<td>4</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Product</th>
<th>Max(quantity)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bagel</td>
<td>20</td>
</tr>
<tr>
<td>Banana</td>
<td>50</td>
</tr>
</tbody>
</table>
Need to be Careful…

**SELECT** product, quantity
FROM Purchase
**GROUP BY** product

**SELECT** product, max(quantity)
FROM Purchase
**GROUP BY** product

-- what does this mean?

<table>
<thead>
<tr>
<th>Product</th>
<th>Price</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bagel</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>Bagel</td>
<td>1.50</td>
<td>20</td>
</tr>
<tr>
<td>Banana</td>
<td>0.5</td>
<td>50</td>
</tr>
<tr>
<td>Banana</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>Banana</td>
<td>4</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Product</th>
<th>Max(quantity)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bagel</td>
<td>20</td>
</tr>
<tr>
<td>Banana</td>
<td>50</td>
</tr>
</tbody>
</table>
Need to be Careful…

```
SELECT product, max(quantity)
FROM Purchase
GROUP BY product
```

```
SELECT product, quantity
FROM Purchase
GROUP BY product
-- what does this mean?
```

<table>
<thead>
<tr>
<th>Product</th>
<th>Price</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bagel</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>Bagel</td>
<td>1.50</td>
<td>20</td>
</tr>
<tr>
<td>Banana</td>
<td>0.5</td>
<td>50</td>
</tr>
<tr>
<td>Banana</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>Banana</td>
<td>4</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Product</th>
<th>Max(quantity)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bagel</td>
<td>20</td>
</tr>
<tr>
<td>Banana</td>
<td>50</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Product</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bagel</td>
<td>20</td>
</tr>
<tr>
<td>Banana</td>
<td>??</td>
</tr>
</tbody>
</table>
Need to be Careful…

```sql
SELECT product, max(quantity)
FROM Purchase
GROUP BY product
```

```sql
SELECT product, quantity
FROM Purchase
GROUP BY product
```

<table>
<thead>
<tr>
<th>Product</th>
<th>Max(quantity)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bagel</td>
<td>20</td>
</tr>
<tr>
<td>Banana</td>
<td>50</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Product</th>
<th>Price</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bagel</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>Bagel</td>
<td>1.50</td>
<td>20</td>
</tr>
<tr>
<td>Banana</td>
<td>0.5</td>
<td>50</td>
</tr>
<tr>
<td>Banana</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>Banana</td>
<td>4</td>
<td>10</td>
</tr>
</tbody>
</table>

Product
-- what does this mean?

<table>
<thead>
<tr>
<th>Product</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bagel</td>
<td>20</td>
</tr>
<tr>
<td>Banana</td>
<td>??</td>
</tr>
</tbody>
</table>

CSE 344 - 2018au
Everything in SELECT must be either a GROUP-BY attribute, or an aggregate.

Need to be Careful…

SELECT product, quantity
FROM Purchase
GROUP BY product

-- what does this mean?

<table>
<thead>
<tr>
<th>Product</th>
<th>Max(quantity)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bagel</td>
<td>20</td>
</tr>
<tr>
<td>Banana</td>
<td>50</td>
</tr>
</tbody>
</table>

SELECT product, max(quantity)
FROM Purchase
GROUP BY product

<table>
<thead>
<tr>
<th>Product</th>
<th>Price</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bagel</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>Bagel</td>
<td>1.50</td>
<td>20</td>
</tr>
<tr>
<td>Banana</td>
<td>0.5</td>
<td>50</td>
</tr>
<tr>
<td>Banana</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>Banana</td>
<td>4</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Product</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bagel</td>
<td>20</td>
</tr>
<tr>
<td>Banana</td>
<td>??</td>
</tr>
</tbody>
</table>

CSE 344 - 2018au
Number of Groups

Purchase(product, price, quantity)

Find total quantities for all sales over $1, by product.

```
SELECT product, Sum(quantity) AS TotalSales
FROM Purchase
WHERE price > 1
GROUP BY product
```
Number of Groups

Purchase(product, price, quantity)

Find total quantities for all sales over $1, by product.

\[
\text{SELECT product, Sum(quantity) AS TotalSales FROM Purchase WHERE price > 1 GROUP BY product}
\]

Clearly, queries return different answers. What about # groups?

\[
\text{SELECT product, Sum(quantity) AS TotalSales FROM Purchase GROUP BY product}
\]
Number of Groups

Purchase(product, price, quantity)

Find total quantities for all sales over $1, by product.

SELECT product, Sum(quantity) AS TotalSales
FROM Purchase
WHERE price > 1
GROUP BY product

Cleary, queries return different answers. What about # groups?

SELECT product, Sum(quantity) AS TotalSales
FROM Purchase
GROUP BY product

Empty groups are removed, hence first query may return fewer groups
Grouping and Aggregation

1. Compute the **FROM** and **WHERE** clauses.

2. Group by the attributes in the **GROUPBY**

3. Compute the **SELECT** clause: grouped attributes and aggregates.

FWGS™
SELECT product, Sum(quantity) AS TotalSales
FROM Purchase
WHERE price > 1
GROUP BY product
### 3.4. Grouping, Select

**SQL Query:**
```sql
SELECT product, Sum(quantity) AS TotalSales
FROM Purchase
WHERE price > 1
GROUP BY product
```

<table>
<thead>
<tr>
<th>Product</th>
<th>Price</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bagel</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>Bagel</td>
<td>1.50</td>
<td>20</td>
</tr>
<tr>
<td>Banana</td>
<td>0.5</td>
<td>50</td>
</tr>
<tr>
<td>Banana</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>Banana</td>
<td>4</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Product</th>
<th>TotalSales</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bagel</td>
<td>40</td>
</tr>
<tr>
<td>Banana</td>
<td>20</td>
</tr>
</tbody>
</table>
Ordering Results

```
SELECT product, sum(price*quantity) as rev
FROM Purchase
GROUP BY product
ORDER BY rev DESC
```

Note: some SQL engines want you to say

```
ORDER BY sum(price*quantity) DESC
```
HAVING Clause

Same query as before, except that we consider only products that had at least 30 sales.

```sql
SELECT product, sum(price*quantity)
FROM Purchase
WHERE price > 1
GROUP BY product
HAVING sum(quantity) > 30
```

HAVING clause contains conditions on aggregates.
General form of Grouping and Aggregation

\[
\text{SELECT } S \\
\text{FROM } R_1, \ldots, R_n \\
\text{WHERE } C_1 \\
\text{GROUP BY } a_1, \ldots, a_k \\
\text{HAVING } C_2
\]

S = may contain attributes \(a_1, \ldots, a_k\) and/or any aggregates but NO OTHER ATTRIBUTES

C1 = is any condition on the attributes in \(R_1, \ldots, R_n\)

C2 = is any condition on aggregate expressions and on attributes \(a_1, \ldots, a_k\)
Semantics of SQL With Group-By

Evaluation steps:

1. Evaluate FROM-WHERE using Nested Loop Semantics
2. Group by the attributes $a_1, \ldots, a_k$
3. Apply condition $C2$ to each group (may have aggregates)
4. Compute aggregates in $S$ and return the result
Exercise

Compute the total income per month
Show only months with less than 10 items sold
Order by quantity sold and display as “TotalSold”
Exercise

Compute the total income per month
Show only months with less than 10 items sold
Order by quantity sold and display as “TotalSold”

```
FROM Purchase
```
Exercise

Compute the total income per month
Show only months with less than 10 items sold
Order by quantity sold and display as “TotalSold”

```
FROM Purchase
GROUP BY month
```
Exercise

Compute the total income per month
Show only months with less than 10 items sold
Order by quantity sold and display as “TotalSold”

```
FROM Purchase
GROUP BY month
HAVING sum(quantity) < 10
```
Exercise

Compute the total income per month
Show only months with less than 10 items sold
Order by quantity sold and display as “TotalSold”

```
SELECT month, sum(price*quantity),
       sum(quantity) as TotalSold
FROM Purchase
GROUP BY month
HAVING sum(quantity) < 10
```
Exercise

Compute the total income per month
Show only months with less than 10 items sold
Order by quantity sold and display as “TotalSold”

```
SELECT month, sum(price*quantity),
       sum(quantity) as TotalSold
FROM Purchase
GROUP BY month
HAVING sum(quantity) < 10
ORDER BY sum(quantity)
```
WHERE vs HAVING

• WHERE condition is applied to individual rows
  – The rows may or may not contribute to the aggregate
  – No aggregates allowed here
  – Occasionally, some groups become empty and are removed

• HAVING condition is applied to the entire group
  – Entire group is returned, or removed
  – May use aggregate functions on the group
Mystery Query

What do they compute?

```
SELECT month, sum(quantity), max(price)
FROM Purchase
GROUP BY month
```

```
SELECT month, sum(quantity)
FROM Purchase
GROUP BY month
```

```
SELECT month
FROM Purchase
GROUP BY month
```
Mystery Query

What do they compute?

```
SELECT month, sum(quantity), max(price)
FROM Purchase
GROUP BY month
```

Lesson:
DISTINCT is a special case of GROUP BY
Aggregate + Join

For each manufacturer, compute how many products with price $> \$100$ they sold
Aggregate + Join

For each manufacturer, compute how many products with price > $100 they sold

Problem: manufacturer is in Product, price is in Purchase...
Aggregate + Join

For each manufacturer, compute how many products with price > $100 they sold

Problem: manufacturer is in Product, price is in Purchase...

-- step 1: think about their join
SELECT ...
FROM Product x, Purchase y
WHERE x.product_id = y.product_id
    and y.price > 100

<table>
<thead>
<tr>
<th>manufacturer</th>
<th>...</th>
<th>price</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hitachi</td>
<td></td>
<td>150</td>
</tr>
<tr>
<td>Canon</td>
<td></td>
<td>300</td>
</tr>
<tr>
<td>Hitachi</td>
<td></td>
<td>180</td>
</tr>
</tbody>
</table>
Aggregate + Join

For each manufacturer, compute how many products with price > $100 they sold

Problem: manufacturer is in Product, price is in Purchase...

-- step 1: think about their join
SELECT ...
FROM Product x, Purchase y
WHERE x.product_id = y.product_id
    and y.price > 100

-- step 2: do the group-by on the join
SELECT x.manufacturer, count(*)
FROM Product x, Purchase y
WHERE x.product_id = y.product_id
    and y.price > 100
GROUP BY x.manufacturer

<table>
<thead>
<tr>
<th>manufacturer</th>
<th>price</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hitachi</td>
<td>150</td>
</tr>
<tr>
<td>Canon</td>
<td>300</td>
</tr>
<tr>
<td>Hitachi</td>
<td>180</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>manufacturer</th>
<th>count(*)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hitachi</td>
<td>2</td>
</tr>
<tr>
<td>Canon</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Aggregate + Join

Variant:
For each manufacturer, compute how many products with price > $100 they sold in each month

```
SELECT x.manufacturer, y.month, count(*)
FROM Product x, Purchase y
WHERE x.product_id = y.product_id
    and y.price > 100
GROUP BY x.manufacturer, y.month
```
Including Empty Groups

• In the result of a group by query, there is one row per group in the result

```sql
SELECT x.manufacturer, count(*)
FROM Product x, Purchase y
WHERE x.product_id = y.product_id
GROUP BY x.manufacturer
```

Count(*) is never 0
Including Empty Groups

```sql
SELECT x.manufacturer, count(*)
FROM Product x, Purchase y
WHERE x.product_id = y.product_id
GROUP BY x.manufacturer
```

<table>
<thead>
<tr>
<th>Pname</th>
<th>Manufacturer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gizmo</td>
<td>GizmoWorks</td>
</tr>
<tr>
<td>Camera</td>
<td>Canon</td>
</tr>
<tr>
<td>OneClick</td>
<td>Hitachi</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Product</th>
<th>Purchase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pname</td>
<td>Manufacturer</td>
</tr>
<tr>
<td>---------</td>
<td>--------------</td>
</tr>
<tr>
<td>Camera</td>
<td>Canon</td>
</tr>
<tr>
<td>Camera</td>
<td>Canon</td>
</tr>
<tr>
<td>OneClick</td>
<td>Hitachi</td>
</tr>
</tbody>
</table>

Final results

<table>
<thead>
<tr>
<th>Manufacturer</th>
<th>Count(*)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canon</td>
<td>2</td>
</tr>
<tr>
<td>Hitachi</td>
<td>1</td>
</tr>
</tbody>
</table>

No GizmoWorks!
Including Empty Groups

```
SELECT x.manufacturer, count(y.pid)
FROM Product x LEFT OUTER JOIN Purchase y
ON x.product_id = y.product_id
GROUP BY x.manufacturer
```

Count(pid) is 0 when all pid’s in the group are NULL
Including Empty Groups

```sql
SELECT x.manufacturer, count(y.pid)
FROM Product x LEFT OUTER JOIN Purchase y
ON x.product_id = y.product_id
GROUP BY x.manufacturer
```

<table>
<thead>
<tr>
<th>prod_id</th>
<th>manufacturer</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gizmo</td>
<td>GizmoWorks</td>
<td></td>
</tr>
<tr>
<td>Camera</td>
<td>Canon</td>
<td></td>
</tr>
<tr>
<td>OneClick</td>
<td>Hitachi</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>prod_id</th>
<th>price</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Camera</td>
<td>150</td>
<td></td>
</tr>
<tr>
<td>Camera</td>
<td>300</td>
<td></td>
</tr>
<tr>
<td>OneClick</td>
<td>180</td>
<td></td>
</tr>
</tbody>
</table>

Final results

<table>
<thead>
<tr>
<th>manufacturer</th>
<th>Count(y.pid)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canon</td>
<td>2</td>
</tr>
<tr>
<td>Hitachi</td>
<td>1</td>
</tr>
<tr>
<td>GizmoWorks</td>
<td>0</td>
</tr>
</tbody>
</table>

Why 0 for GizmoWorks?

GizmoWorks is paired with NULLs
Including Empty Groups

```
SELECT x.manufacturer, count(*)
FROM Product x LEFT OUTER JOIN Purchase y
ON x.product_id = y.product_id
GROUP BY x.manufacturer
```

Product

<table>
<thead>
<tr>
<th>prod_id</th>
<th>manufacturer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gizmo</td>
<td>GizmoWorks</td>
</tr>
<tr>
<td>Camera</td>
<td>Canon</td>
</tr>
<tr>
<td>OneClick</td>
<td>Hitachi</td>
</tr>
</tbody>
</table>

Purchase

<table>
<thead>
<tr>
<th>product</th>
<th>price</th>
</tr>
</thead>
<tbody>
<tr>
<td>Camera</td>
<td>150</td>
</tr>
<tr>
<td>Camera</td>
<td>300</td>
</tr>
<tr>
<td>OneClick</td>
<td>180</td>
</tr>
</tbody>
</table>

Left Outer Join(Product, Purchase)

<table>
<thead>
<tr>
<th>prod_id</th>
<th>manufacturer</th>
<th>product</th>
<th>price</th>
</tr>
</thead>
<tbody>
<tr>
<td>Camera</td>
<td>Canon</td>
<td>Camera</td>
<td>150</td>
</tr>
<tr>
<td>Camera</td>
<td>Canon</td>
<td>Camera</td>
<td>300</td>
</tr>
<tr>
<td>OneClick</td>
<td>Hitachi</td>
<td>OneClick</td>
<td>180</td>
</tr>
<tr>
<td>Gizmo</td>
<td>GizmoWorks</td>
<td>NULL</td>
<td>NULL</td>
</tr>
</tbody>
</table>

Final results

<table>
<thead>
<tr>
<th>manufacturer</th>
<th>Count(*)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canon</td>
<td>2</td>
</tr>
<tr>
<td>Hitachi</td>
<td>1</td>
</tr>
<tr>
<td>GizmoWorks</td>
<td>1</td>
</tr>
</tbody>
</table>

Probably not what we want!
Introduction to Data Management
CSE 344

Lecture 6: Nested Queries in SQL
Announcements

• HW2 is due tomorrow (Tuesday)

• HW3: you received an email from invites@microsoft.com: accept it

• Webquiz 2 due on Friday
What have we learned so far

- Data models
- Relational data model
  - Instance: relations
  - Schema: table with attribute names
  - Language: SQL
What have we learned so far

SQL features
• Projections
• Selections
• Joins (inner and outer)
• Aggregates
• Group by
• Inserts, updates, and deletes

Make sure you read the textbook!
Lecture Goals

• Today we will learn how to write (even) more powerful SQL queries

• Reading: Ch. 6.3
Subqueries

• A subquery is a SQL query nested inside a larger query
• Such inner-outer queries are called nested queries
• A subquery may occur in:
  – A SELECT clause
  – A FROM clause
  – A WHERE clause

• Rule of thumb: avoid nested queries when possible
  – But sometimes it’s impossible, as we will see
Subqueries…

• Can return a single value to be included in a SELECT clause
• Can return a relation to be included in the FROM clause, aliased using a tuple variable
• Can return a single value to be compared with another value in a WHERE clause
• Can return a relation to be used in the WHERE or HAVING clause under an existential quantifier
1. Subqueries in SELECT

Product (pname, price, cid)
Company (cid, cname, city)

For each product return the city where it is manufactured

```
SELECT X.pname, (SELECT Y.city
               FROM Company Y
               WHERE Y.cid=X.cid) as City
FROM Product X
```

What happens if the subquery returns more than one city?
We get a runtime error
(and SQLite simply ignores the extra values…)

“correlated subquery”
1. Subqueries in SELECT

Whenever possible, don’t use a nested queries:

```sql
SELECT X.pname, (SELECT Y.city
                 FROM Company Y
                 WHERE Y.cid=X.cid) as City
FROM   Product X
```

We have “unnested” the query.
1. Subqueries in SELECT

Compute the number of products made by each company

```
SELECT DISTINCT C.cname, (SELECT count(*)
   FROM Product P
   WHERE P.cid=C.cid)
FROM Company C
```
1. Subqueries in SELECT

Compute the number of products made by each company

```
SELECT DISTINCT C.cname, (SELECT count(*)
    FROM Product P
    WHERE P.cid=C.cid)
FROM Company C
```

Better: we can unnest using a GROUP BY

```
SELECT C.cname, count(*)
FROM Company C, Product P
WHERE C.cid=P.cid
GROUP BY C.cname
```
1. Subqueries in SELECT

But are these really equivalent?

```sql
SELECT DISTINCT C.cname, (SELECT count(*)
FROM Product P
WHERE P.cid=C.cid)
FROM Company C
```

```sql
SELECT C.cname, count(*)
FROM Company C, Product P
WHERE C.cid=P.cid
GROUP BY C.cname
```
1. Subqueries in SELECT

But are these really equivalent?

```sql
SELECT DISTINCT C.cname, (SELECT count(*)
FROM Product P
WHERE P.cid=C.cid)
FROM Company C
```

No! Different results if a company has no products

```sql
SELECT C.cname, count(*)
FROM Company C, Product P
WHERE C.cid=P.cid
GROUP BY C.cname
```

Recall: count of an empty table is 0

```sql
SELECT C.cname, count(pname)
FROM Company C LEFT OUTER JOIN Product P
ON C.cid=P.cid
GROUP BY C.cname
```
2. Subqueries in FROM

Find all products whose prices is > 20 and < 500

```
SELECT X.pname
FROM (SELECT *
      FROM Product AS Y
      WHERE price > 20) as X
WHERE X.price < 500
```
Product (pname, price, cid)  
Company (cid, cname, city)

2. Subqueries in FROM

Find all products whose prices is > 20 and < 500

```
SELECT X.pname  
FROM (SELECT *  
      FROM Product AS Y  
      WHERE price > 20) as X  
WHERE X.price < 500
```

Try unnest this query!
Find all products whose prices is > 20 and < 500

```sql
SELECT X.pname
FROM (SELECT *
      FROM Product AS Y
      WHERE price > 20) as X
WHERE X.price < 500
```

Try unnest this query!

Side note: This is not a correlated subquery. (why?)
2. Subqueries in FROM

Sometimes we need to compute an intermediate table only to use it later in a SELECT-FROM-WHERE

• Option 1: use a subquery in the FROM clause
• Option 2: use the WITH clause
2. Subqueries in FROM

```
SELECT X.pname
FROM (SELECT *
     FROM Product AS Y
     WHERE price > 20) as X
WHERE X.price < 500
```

A subquery whose result we called myTable

```
WITH myTable AS (SELECT *
                FROM Product AS Y
                WHERE price > 20)
SELECT X.pname
FROM myTable as X
WHERE X.price < 500
```
3. Subqueries in WHERE

Find all companies that make some products with price < 200
3. Subqueries in WHERE

Find all companies that make some products with price < 200

Product (pname, price, cid)
Company (cid, cname, city)
3. Subqueries in WHERE

Find all companies that make some products with price < 200

Using **EXISTS**: 

```sql
SELECT DISTINCT C.cname
FROM Company C
WHERE EXISTS (SELECT *
              FROM Product P
              WHERE C.cid = P.cid AND P.price < 200)
```
3. Subqueries in WHERE

Find all companies that make some products with price < 200

Using \texttt{IN}

\begin{verbatim}
SELECT DISTINCT C.cname
FROM Company C
WHERE C.cid IN (SELECT P.cid
                 FROM Product P
                 WHERE P.price < 200)
\end{verbatim}
3. Subqueries in WHERE

Find all companies that make some products with price < 200

Using ANY:

\[
\begin{align*}
&\text{SELECT DISTINCT C.cname} \\
&\text{FROM Company C} \\
&\text{WHERE 200 > ANY (SELECT price} \\
&\quad\text{FROM Product P} \\
&\quad\text{WHERE P.cid = C.cid)}
\end{align*}
\]
3. Subqueries in WHERE

Find all companies that make some products with price < 200

Using **ANY**:

```
SELECT DISTINCT C.cname
FROM Company C
WHERE 200 > ANY (SELECT price
                   FROM Product P
                   WHERE P.cid = C.cid)
```

Existential quantifiers

Product (**pname**, **price**, **cid**)  
Company (**cid**, **cname**, **city**)  

Not supported in sqlite
3. Subqueries in WHERE

Find all companies that make some products with price < 200

Now let’s unnest it:

```
SELECT DISTINCT C.cname
FROM Company C, Product P
WHERE C.cid = P.cid and P.price < 200
```
3. Subqueries in WHERE

Find all companies that make some products with price < 200

Existential quantifiers

Now let’s unnest it:

```sql
SELECT DISTINCT C.cname
FROM Company C, Product P
WHERE C.cid = P.cid and P.price < 200
```

Existential quantifiers are easy! 😊
3. Subqueries in WHERE

Find all companies s.t. all their products have price < 200

same as:

Find all companies that make only products with price < 200
3. Subqueries in WHERE

Find all companies s.t. all their products have price < 200

same as:

Find all companies that make only products with price < 200
3. Subqueries in WHERE

Find all companies s.t. all their products have price < 200

same as:

Find all companies that make only products with price < 200

Universal quantifiers are hard! 😞
Product (\textit{pname}, \textit{price}, \textit{cid})
Company (\textit{cid}, \textit{cname}, \textit{city})

3. Subqueries in WHERE

Find all companies s.t. all their products have price < 200

1. Find \textit{the other} companies ... ...which ones?
Product (pname, price, cid)
Company (cid, cname, city)

3. Subqueries in WHERE

Find all companies s.t. all their products have price < 200

1. Find the other companies that make some product ≥ 200
3. Subqueries in WHERE

Find all companies s.t. all their products have price < 200

1. Find the other companies that make some product ≥ 200

```sql
SELECT DISTINCT C.cname
FROM Company C
WHERE C.cid IN (SELECT P.cid
                   FROM Product P
                   WHERE P.price >= 200)
```
Product (pname, price, cid)
Company (cid, cname, city)

3. Subqueries in WHERE

Find all companies s.t. all their products have price < 200

1. Find the other companies that make some product ≥ 200

```
SELECT DISTINCT C.cname
FROM Company C
WHERE C.cid IN (SELECT P.cid
                 FROM Product P
                 WHERE P.price >= 200)
```

2. Find all companies s.t. all their products have price < 200

```
SELECT DISTINCT C.cname
FROM Company C
WHERE C.cid NOT IN (SELECT P.cid
                    FROM Product P
                    WHERE P.price >= 200)
```
3. Subqueries in WHERE

Find all companies s.t. all their products have price < 200

Using `EXISTS`:

```
SELECT DISTINCT C.cname
FROM Company C
WHERE NOT EXISTS (SELECT *
    FROM Product P
    WHERE P.cid = C.cid and P.price >= 200)
```
3. Subqueries in WHERE

Find all companies s.t. all their products have price < 200

Using **ALL**:

```sql
SELECT DISTINCT C.cname
FROM Company C
WHERE 200 >= ALL (SELECT price
                    FROM Product P
                    WHERE P.cid = C.cid)
```
3. Subqueries in WHERE

Find all companies s.t. all their products have price < 200

Using **ALL**:  

```sql
SELECT DISTINCT C.cname  
FROM Company C  
WHERE 200 >= ALL (SELECT price  
FROM Product P  
WHERE P.cid = C.cid)
```

**Not supported in sqlite**
Question for Database Theory
Fans and their Friends

• Can we unnest the universal quantifier query?

• We need to first discuss the concept of monotonicity
Monotone Queries

- Definition A query Q is monotone if:
  - Whenever we add tuples to one or more input tables, the answer to the query will not lose any of the tuples

Product (pname, price, cid)
Company (cid, cname, city)
Monotone Queries

- Definition A query $Q$ is **monotone** if:
  - Whenever we add tuples to one or more input tables, the answer to the query will not lose any of the tuples

<table>
<thead>
<tr>
<th>Product</th>
<th>Company</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>pname</strong></td>
<td><strong>cid</strong></td>
</tr>
<tr>
<td>Gizmo</td>
<td>c001</td>
</tr>
<tr>
<td>Gadget</td>
<td>c004</td>
</tr>
<tr>
<td>Camera</td>
<td>c003</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>pname</strong></th>
<th><strong>city</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Gizmo</td>
<td>Lyon</td>
</tr>
<tr>
<td>Camera</td>
<td>Lodtz</td>
</tr>
</tbody>
</table>
Monotone Queries

- Definition A query Q is **monotone** if:
  - Whenever we add tuples to one or more input tables, the answer to the query will not lose any of the tuples.

Product (pname, price, cid)
Company (cid, cname, city)

So far it looks monotone...
Monotone Queries

- **Definition** A query $Q$ is **monotone** if:
  - Whenever we add tuples to one or more input tables, the answer to the query will not lose any of the tuples.

<table>
<thead>
<tr>
<th>Product</th>
<th>Company</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>pname</strong></td>
<td><strong>company</strong></td>
</tr>
<tr>
<td>Gizmo</td>
<td>c001</td>
</tr>
<tr>
<td>Gadget</td>
<td>c004</td>
</tr>
<tr>
<td>Camera</td>
<td>c003</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Product</th>
<th>Company</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>pname</strong></td>
<td><strong>company</strong></td>
</tr>
<tr>
<td>Gizmo</td>
<td>c001</td>
</tr>
<tr>
<td>Camera</td>
<td>c003</td>
</tr>
</tbody>
</table>

$Q$ is not monotone!
Monotone Queries

• **Theorem**: If Q is a SELECT-FROM-WHERE query that does not have subqueries, and no aggregates, then it is monotone.
Monotone Queries

• **Theorem**: If Q is a SELECT-FROM-WHERE query that does not have subqueries, and no aggregates, then it is monotone.

• **Proof.** We use the nested loop semantics: if we insert a tuple in a relation $R_i$, this will not remove any tuples from the answer.

```sql
SELECT a_1, a_2, ..., a_k
FROM  R_1 AS x_1, R_2 AS x_2, ..., R_n AS x_n
WHERE  Conditions
```

```plaintext
for x_1 in R_1 do
  for x_2 in R_2 do
    ...
    for x_n in R_n do
      if Conditions
        output (a_1,...,a_k)
```
Monotone Queries

• The query:

Find all companies s.t. all their products have price < 200 is not monotone
Monotone Queries

• The query:

Find all companies s.t. all their products have price < 200 is not monotone
Monotone Queries

• The query:

Find all companies s.t. all their products have price < 200 is not monotone

- Consequence: If a query is not monotonic, then we cannot write it as a SELECT-FROM-WHERE query without nested subqueries.
Queries that must be nested

• Queries with universal quantifiers or with negation
Queries that must be nested

- Queries with universal quantifiers or with negation

- Queries with aggregates are usually not monotone
  - `sum(..)` and `count(*)` are NOT monotone, because they do not satisfy set containment
  - `select count(*) from R` is not monotone!
Introduction to Data Management
CSE 344

Lecture 7-8: SQL Wrap-up
Relational Algebra
Announcements

• You received invitation email to @cs
• You will be prompted to choose passwd
  – Problems with existing account?
  – In the worst case we will ask you to create a new @outlook account just for this class
• If OK, create the database server
  – Choose cheapest pricing tier!
• Remember: WQ2 due on Friday
GROUP BY v.s. Nested Queries

```
SELECT product, Sum(quantity) AS TotalSales
FROM Purchase
WHERE price > 1
GROUP BY product
```

```
SELECT DISTINCT x.product, (SELECT Sum(y.quantity)
    FROM Purchase y
    WHERE x.product = y.product
    AND y.price > 1)
    AS TotalSales
FROM Purchase x
WHERE x.price > 1
```

Why twice?
Author(login, name)
Wrote(login, url)

More Unnesting

Find authors who wrote ≥ 10 documents:
More Unnesting

Find authors who wrote ≥ 10 documents:
Attempt 1: with nested queries

```
SELECT DISTINCT Author.name
FROM Author
WHERE (SELECT count(Wrote.url)
FROM Wrote
WHERE Author.login=Wrote.login)
>= 10
```

This is SQL by a novice
More Unnesting

Find authors who wrote ≥ 10 documents:

Attempt 1: with nested queries

Attempt 2: using GROUP BY and HAVING

```
SELECT Author.name
FROM Author, Wrote
WHERE Author.login=Wrote.login
GROUP BY Author.name
HAVING count(wrote.url) >= 10
```
Finding Witnesses

For each city, find the most expensive product made in that city
Finding Witnesses

For each city, find the most expensive product made in that city
Finding the maximum price is easy...

```
SELECT x.city, max(y.price)
FROM Company x, Product y
WHERE x.cid = y.cid
GROUP BY x.city;
```

But we need the *witnesses*, i.e., the products with max price
To find the witnesses, compute the maximum price in a subquery (in FROM or in WITH)

WITH CityMax AS
  (SELECT x.city, max(y.price) as maxprice
   FROM Company x, Product y
   WHERE x.cid = y.cid
   GROUP BY x.city)
SELECT DISTINCT u.city, v.pname, v.price
FROM Company u, Product v, CityMax w
WHERE u.cid = v.cid
  and u.city = w.city
  and v.price = w.maxprice;
Finding Witnesses

To find the witnesses, compute the maximum price in a subquery (in FROM or in WITH)

```
SELECT DISTINCT u.city, v.pname, v.price
FROM Company u, Product v,
    (SELECT x.city, max(y.price) as maxprice
    FROM Company x, Product y
    WHERE x.cid = y.cid
    GROUP BY x.city) w
WHERE u.cid = v.cid
    and u.city = w.city
    and v.price = w.maxprice;
```
Finding Witnesses

Or we can use a subquery in where clause

```
SELECT u.city, v.pname, v.price
FROM Company u, Product v
WHERE u.cid = v.cid
  and v.price >= ALL (SELECT y.price
                      FROM Company x, Product y
                      WHERE u.city=x.city
                      and x.cid=y.cid);
```
There is a more concise solution here:

```sql
SELECT u.city, v.pname, v.price
FROM Company u, Product v, Company x, Product y
WHERE u.cid = v.cid and u.city = x.city
    and x.cid = y.cid
GROUP BY u.city, v.pname, v.price
HAVING v.price = max(y.price)
```
SQL: Our first language for the relational model

- Projections
- Selections
- Joins (inner and outer)
- Inserts, updates, and deletes
- Aggregates
- Grouping
- Ordering
- Nested queries
Relational Algebra
Relational Algebra

- In SQL we say *what* we want
- In RA we can express *how* to get it
- RA = set-at-a-time algebra for relations

- Every DBMS implementations converts a SQL query to RA in order to execute it
- An RA expression is called a *query plan*
Why study another relational query language?

• RA is how SQL is implemented in DBMS
  – We will see more of this in a few weeks

• RA opens up opportunities for *query optimization*
Basics

- Relations and attributes
- Functions that are applied to relations
  - Return relations
  - Can be composed together
  - Often displayed using a tree rather than linearly
  - Use Greek symbols: $\sigma$, $\pi$, $\delta$, etc
Sets v.s. Bags

- Sets: \{a,b,c\}, \{a,d,e,f\}, \\}, \ldots
- Bags: \{a, a, b, c\}, \{b, b, b, b, b\}, \ldots

Relational Algebra has two flavors:
- Set semantics = standard Relational Algebra
- Bag semantics = extended Relational Algebra

DB systems implement bag semantics (Why?)
Relational Algebra Operators

• Union $\cup$, intersection $\cap$, difference $-$
• Selection $\sigma$
• Projection $\pi$
• Cartesian product $\times$, join $\Join$
• (Rename $\rho$)
• Duplicate elimination $\delta$
• Grouping and aggregation $\gamma$
• Sorting $\tau$

All operators take in 1 or 2 relations as inputs and return another relation
Union and Difference

\[ R1 \cup R2 \]
\[ R1 - R2 \]

Only make sense if R1, R2 have the same schema

What do they mean over bags?
What about Intersection?

• Derived operator using minus

\[ R_1 \cap R_2 = R_1 - (R_1 - R_2) \]

• Derived using join

\[ R_1 \cap R_2 = R_1 \bowtie R_2 \]
Selection

• Returns all tuples which satisfy a condition

\[ \sigma_c(R) \]

• Examples
  
  – \( \sigma_{\text{Salary} > 40000} \) (Employee)
  
  – \( \sigma_{\text{name} = \text{"Smith"}} \) (Employee)

• The condition c can be =, <, <=, >, >=, <> combined with AND, OR, NOT
### Employee Table

<table>
<thead>
<tr>
<th>SSN</th>
<th>Name</th>
<th>Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>1234545</td>
<td>John</td>
<td>20000</td>
</tr>
<tr>
<td>5423341</td>
<td>Smith</td>
<td>60000</td>
</tr>
<tr>
<td>4352342</td>
<td>Fred</td>
<td>50000</td>
</tr>
</tbody>
</table>

\[ \sigma_{\text{Salary} > 40000} \ (\text{Employee}) \]

### Sub-Table

<table>
<thead>
<tr>
<th>SSN</th>
<th>Name</th>
<th>Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>5423341</td>
<td>Smith</td>
<td>60000</td>
</tr>
<tr>
<td>4352342</td>
<td>Fred</td>
<td>50000</td>
</tr>
</tbody>
</table>
Projection

• Eliminates columns

\[ \pi_{A_1, \ldots, A_n} (R) \]

• Example: project social-security number and names:

\[ \pi_{\text{SSN, Name}} (\text{Employee}) \rightarrow \text{Answer(SSN, Name)} \]

Different semantics over sets or bags! Why?
### Employee

<table>
<thead>
<tr>
<th>SSN</th>
<th>Name</th>
<th>Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>1234545</td>
<td>John</td>
<td>20000</td>
</tr>
<tr>
<td>5423341</td>
<td>John</td>
<td>60000</td>
</tr>
<tr>
<td>4352342</td>
<td>John</td>
<td>20000</td>
</tr>
</tbody>
</table>

\[
\pi_{\text{Name, Salary}}(\text{Employee})
\]

<table>
<thead>
<tr>
<th>Name</th>
<th>Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>20000</td>
</tr>
<tr>
<td>John</td>
<td>60000</td>
</tr>
<tr>
<td>John</td>
<td>20000</td>
</tr>
</tbody>
</table>

### Bag semantics

<table>
<thead>
<tr>
<th>Name</th>
<th>Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>20000</td>
</tr>
<tr>
<td>John</td>
<td>60000</td>
</tr>
</tbody>
</table>

### Set semantics

<table>
<thead>
<tr>
<th>Name</th>
<th>Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>20000</td>
</tr>
</tbody>
</table>

Which is more efficient?
Composing RA Operators

**Patient**

<table>
<thead>
<tr>
<th>no</th>
<th>name</th>
<th>zip</th>
<th>disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>p1</td>
<td>98125</td>
<td>flu</td>
</tr>
<tr>
<td>2</td>
<td>p2</td>
<td>98125</td>
<td>heart</td>
</tr>
<tr>
<td>3</td>
<td>p3</td>
<td>98120</td>
<td>lung</td>
</tr>
<tr>
<td>4</td>
<td>p4</td>
<td>98120</td>
<td>heart</td>
</tr>
</tbody>
</table>

\[\sigma_{\text{disease}='\text{heart'}}(\text{Patient})\]

<table>
<thead>
<tr>
<th>no</th>
<th>name</th>
<th>zip</th>
<th>disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>p2</td>
<td>98125</td>
<td>heart</td>
</tr>
<tr>
<td>4</td>
<td>p4</td>
<td>98120</td>
<td>heart</td>
</tr>
</tbody>
</table>

**\(\pi_{\text{zip, disease}}(\text{Patient})\)**

<table>
<thead>
<tr>
<th>zip</th>
<th>disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>98125</td>
<td>flu</td>
</tr>
<tr>
<td>98125</td>
<td>heart</td>
</tr>
<tr>
<td>98120</td>
<td>lung</td>
</tr>
<tr>
<td>98120</td>
<td>heart</td>
</tr>
</tbody>
</table>

**\(\pi_{\text{zip, disease}}(\sigma_{\text{disease}='\text{heart'}}(\text{Patient}))\)**

<table>
<thead>
<tr>
<th>zip</th>
<th>disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>98125</td>
<td>heart</td>
</tr>
<tr>
<td>98120</td>
<td>heart</td>
</tr>
</tbody>
</table>
Cartesian Product

- Each tuple in R1 with each tuple in R2

\[ R1 \times R2 \]

- Rare in practice; mainly used to express joins
**Cross-Product Example**

<table>
<thead>
<tr>
<th>Employee</th>
<th>Dependent</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Name</strong></td>
<td><strong>EmpSSN</strong></td>
</tr>
<tr>
<td>John</td>
<td>9999999999</td>
</tr>
<tr>
<td>Tony</td>
<td>7777777777</td>
</tr>
</tbody>
</table>

**Employee × Dependent**

<table>
<thead>
<tr>
<th>Name</th>
<th>SSN</th>
<th>EmpSSN</th>
<th>DepName</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>9999999999</td>
<td>9999999999</td>
<td>Emily</td>
</tr>
<tr>
<td>John</td>
<td>9999999999</td>
<td>7777777777</td>
<td>Joe</td>
</tr>
<tr>
<td>Tony</td>
<td>7777777777</td>
<td>9999999999</td>
<td>Emily</td>
</tr>
<tr>
<td>Tony</td>
<td>7777777777</td>
<td>7777777777</td>
<td>Joe</td>
</tr>
</tbody>
</table>
Renaming

• Changes the schema, not the instance

\[ \rho_{B_1, \ldots, B_n} (R) \]

• Example:
  – Given Employee(Name, SSN)
  – \[ \rho_{N, S}(\text{Employee}) \rightarrow \text{Answer}(N, S) \]
Natural Join

\[ R_1 \bowtie R_2 \]

• Meaning: \( R_1 \bowtie R_2 = \Pi_A(\sigma_\theta (R_1 \times R_2)) \)

• Where:
  – Selection \( \sigma_\theta \) checks equality of all common attributes (i.e., attributes with same names)
  – Projection \( \Pi_A \) eliminates duplicate common attributes
Natural Join Example

\[ R \bowtie S = \Pi_{ABC}(\sigma_{R.B=S.B}(R \times S)) \]

\[ R \bowtie S = \]

\[ \begin{array}{lll}
A & B & C \\
X & Z & U \\
X & Z & V \\
Y & Z & U \\
Y & Z & V \\
Z & V & W \\
\end{array} \]
Natural Join Example 2

AnonPatient P

<table>
<thead>
<tr>
<th>age</th>
<th>zip</th>
<th>disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>54</td>
<td>98125</td>
<td>heart</td>
</tr>
<tr>
<td>20</td>
<td>98120</td>
<td>flu</td>
</tr>
</tbody>
</table>

Voters V

<table>
<thead>
<tr>
<th>name</th>
<th>age</th>
<th>zip</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>54</td>
<td>98125</td>
</tr>
<tr>
<td>Bob</td>
<td>20</td>
<td>98120</td>
</tr>
</tbody>
</table>

P \bowtie V

<table>
<thead>
<tr>
<th>age</th>
<th>zip</th>
<th>disease</th>
<th>name</th>
</tr>
</thead>
<tbody>
<tr>
<td>54</td>
<td>98125</td>
<td>heart</td>
<td>Alice</td>
</tr>
<tr>
<td>20</td>
<td>98120</td>
<td>flu</td>
<td>Bob</td>
</tr>
</tbody>
</table>
Natural Join

• Given schemas $R(A, B, C, D)$, $S(A, C, E)$, what is the schema of $R \bowtie S$?

• Given $R(A, B, C)$, $S(D, E)$, what is $R \bowtie S$?

• Given $R(A, B)$, $S(A, B)$, what is $R \bowtie S$?
Theta Join

- A join that involves a predicate

\[ R1 \bowtie_\theta R2 = \sigma_\theta (R1 \times R2) \]

- Here \( \theta \) can be any condition
- No projection in this case!
- For our voters/patients example:

\[ P \bowtie \quad \text{P.zip} = \text{V.zip and P.age} \geq \text{V.age} - 1 \quad \text{and P.age} \leq \text{V.age} + 1 \]

AnonPatient (age, zip, disease)
Voters (name, age, zip)
Equijoin

- A theta join where $\theta$ is an equality predicate

\[
R_1 \bowtie_{\theta} R_2 = \sigma_{\theta} (R_1 \times R_2)
\]

- By far the most used variant of join in practice
- What is the relationship with natural join?
**Equijoin Example**

**AnonPatient P**

<table>
<thead>
<tr>
<th>age</th>
<th>zip</th>
<th>disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>54</td>
<td>98125</td>
<td>heart</td>
</tr>
<tr>
<td>20</td>
<td>98120</td>
<td>flu</td>
</tr>
</tbody>
</table>

**Voters V**

<table>
<thead>
<tr>
<th>name</th>
<th>age</th>
<th>zip</th>
</tr>
</thead>
<tbody>
<tr>
<td>p1</td>
<td>54</td>
<td>98125</td>
</tr>
<tr>
<td>p2</td>
<td>20</td>
<td>98120</td>
</tr>
</tbody>
</table>

**P \( \bowtie_{P.age=V.age} \) V**

<table>
<thead>
<tr>
<th>P.age</th>
<th>P.zip</th>
<th>P.disease</th>
<th>V.name</th>
<th>V.age</th>
<th>V.zip</th>
</tr>
</thead>
<tbody>
<tr>
<td>54</td>
<td>98125</td>
<td>heart</td>
<td>p1</td>
<td>54</td>
<td>98125</td>
</tr>
<tr>
<td>20</td>
<td>98120</td>
<td>flu</td>
<td>p2</td>
<td>20</td>
<td>98120</td>
</tr>
</tbody>
</table>
Join Summary

- **Theta-join**: \( R \bowtie_\theta S = \sigma_\theta (R \times S) \)
  - Join of \( R \) and \( S \) with a join condition \( \theta \)
  - Cross-product followed by selection \( \theta \)
  - No projection

- **Equijoin**: \( R \bowtie_\theta S = \sigma_\theta (R \times S) \)
  - Join condition \( \theta \) consists only of equalities
  - No projection

- **Natural join**: \( R \bowtie S = \pi_A (\sigma_\theta (R \times S)) \)
  - Equality on **all** fields with same name in \( R \) and in \( S \)
  - Projection \( \pi_A \) drops all redundant attributes
So Which Join Is It?

When we write $R \bowtie S$ we usually mean an equijoin, but we often omit the equality predicate when it is clear from the context.
More Joins

• **Outer join**
  – Include tuples with no matches in the output
  – Use NULL values for missing attributes
  – Does not eliminate duplicate columns

• **Variants**
  – Left outer join
  – Right outer join
  – Full outer join
### Outer Join Example

#### AnonPatient P

<table>
<thead>
<tr>
<th>age</th>
<th>zip</th>
<th>disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>54</td>
<td>98125</td>
<td>heart</td>
</tr>
<tr>
<td>20</td>
<td>98120</td>
<td>flu</td>
</tr>
<tr>
<td>33</td>
<td>98120</td>
<td>lung</td>
</tr>
</tbody>
</table>

#### AnonJob J

<table>
<thead>
<tr>
<th>job</th>
<th>age</th>
<th>zip</th>
</tr>
</thead>
<tbody>
<tr>
<td>lawyer</td>
<td>54</td>
<td>98125</td>
</tr>
<tr>
<td>cashier</td>
<td>20</td>
<td>98120</td>
</tr>
</tbody>
</table>

#### Join Operation

![Join Operation Diagram]

<table>
<thead>
<tr>
<th>P.age</th>
<th>P.zip</th>
<th>P.disease</th>
<th>J.job</th>
<th>J.age</th>
<th>J.zip</th>
</tr>
</thead>
<tbody>
<tr>
<td>54</td>
<td>98125</td>
<td>heart</td>
<td>lawyer</td>
<td>54</td>
<td>98125</td>
</tr>
<tr>
<td>20</td>
<td>98120</td>
<td>flu</td>
<td>cashier</td>
<td>20</td>
<td>98120</td>
</tr>
<tr>
<td>33</td>
<td>98120</td>
<td>lung</td>
<td>null</td>
<td>null</td>
<td>null</td>
</tr>
</tbody>
</table>
Some Examples

Supplier(sno, surname, scity, sstate)
Part(pno, pname, psize, pcolor)
Supply(sno, pno, qty, price)

Name of supplier of parts with size greater than 10
\[ \pi_{\text{sname}} (\text{Supplier} \bowtie (\text{Supply} \bowtie (\sigma_{\text{psize}>10} (\text{Part}))) \)

Name of supplier of red parts or parts with size greater than 10
\[ \pi_{\text{sname}} (\text{Supplier} \bowtie (\text{Supply} \bowtie (\sigma_{\text{psize}>10} (\text{Part}) \cup \sigma_{\text{pcolor}='red'} (\text{Part})))) \]
\[ \pi_{\text{sname}} (\text{Supplier} \bowtie (\text{Supply} \bowtie (\sigma_{\text{psize}>10} \lor \text{pcolor}='red' (\text{Part})))) \]

Can be represented as trees as well
Some Examples

Supplier(sno,sname,scity,sstate)
Part(pno,pname,psize,pcolor)
Supply(sno,pno,qty,price)

Name of supplier of parts with size greater than 10
Project[sname](Supplier Join[sno=sno]
  (Supply Join[pno=pno] (Select[psize>10](Part))))

Name of supplier of red parts or parts with size greater than 10
Project[sname](Supplier Join[sno=sno]
  (Supply Join[pno=pno]
   ((Select[psize>10](Part)) Union
    (Select[pcolor='red'](Part))))

Project[sname](Supplier Join[sno=sno] (Supply Join[pno=pno]
  (Select[psize>10 OR pcolor='red'](Part))))

Can be represented as trees as well
Representing RA Queries as Trees

\[ \pi_{\text{sname}} (\text{Supplier} \bowtie \text{Supply} \bowtie (\sigma_{\text{psize} > 10} (\text{Part}))) \]
Relational Algebra Operators

- Union $\cup$, intersection $\cap$, difference $-$
- Selection $\sigma$
- Projection $\pi$
- Cartesian product $\times$, join $\Join$
- (Rename $\rho$)
- Duplicate elimination $\delta$
- Grouping and aggregation $\gamma$
- Sorting $\tau$

All operators take in 1 or 2 relations as inputs and return another relation
Extended RA: Operators on Bags

• Duplicate elimination $\delta$

• Grouping $\gamma$
  – Takes in relation and a list of grouping operations (e.g., aggregates). Returns a new relation.

• Sorting $\tau$
  – Takes in a relation, a list of attributes to sort on, and an order. Returns a new relation.
Using Extended RA Operators

\[
\begin{align*}
\text{Answer} & \quad \Pi_{\text{city, q}} \\
& \quad \sigma_{c > 100} \quad T2(\text{city, q, c}) \\
& \quad \gamma_{\text{city, sum(quantity)} \rightarrow q, \text{count(*)} \rightarrow c} \quad T1(\text{city, q, c}) \\
& \quad \gamma_{\text{city, sum(quantity)} \rightarrow q, \text{count(*)} \rightarrow c} \quad \text{sales(product, city, quantity)}
\end{align*}
\]

\[
\begin{align*}
\text{SELECT} & \quad \text{city, sum(quantity)} \\
\text{FROM} & \quad \text{sales} \\
\text{GROUP BY} & \quad \text{city} \\
\text{HAVING} & \quad \text{count(*)} > 100
\end{align*}
\]

\(T1, T2\) = temporary tables
Typical Plan for a Query (1/2)

Answer

\[ \pi_{\text{fields}} \]

\[ \sigma_{\text{selection condition}} \]

join condition

join condition

R

S

SELECT fields
FROM R, S, ...
WHERE condition

SELECT-PROJECT-JOIN
Query
Typical Plan for a Query (1/2)

\[ \sigma_{\text{having condition}} \]
\[ \gamma_{\text{fields, sum/count/min/max(fields)}} \]
\[ \pi_{\text{fields}} \]
\[ \sigma_{\text{where condition}} \]
\[ \text{join condition} \]

SELECT fields
FROM R, S, ...
WHERE condition
GROUP BY fields
HAVING condition
How about Subqueries?

```
SELECT  Q.sno
FROM    Supplier Q
WHERE   Q.sstate = 'WA'
        and not exists
        (SELECT  *
         FROM    Supply P
         WHERE   P.sno = Q.sno
                 and P.price > 100)
```
How about Subqueries?

```
SELECT  Q.sno
FROM    Supplier Q
WHERE   Q.sstate = 'WA'
        and not exists
          (SELECT  *
           FROM    Supply P
           WHERE   P.sno = Q.sno
                   and P.price > 100)
```
How about Subqueries?

```
SELECT  Q.sno
FROM    Supplier Q
WHERE   Q.sstate = 'WA'
        and not exists
        (SELECT  *
         FROM    Supply P
         WHERE   P.sno = Q.sno
                 and P.price > 100)
```

De-Correlation

```
SELECT  Q.sno
FROM    Supplier Q
WHERE   Q.sstate = 'WA'
        and Q.sno not in
        (SELECT  P.sno
         FROM    Supply P
         WHERE   P.price > 100)
```
(SELECT Q.sno
FROM Supplier Q
WHERE Q.sstate = 'WA')
EXCEPT
(SELECT P.sno
FROM Supply P
WHERE P.price > 100)
EXCEPT = set difference

How about Subqueries?

SELECT Q.sno
FROM Supplier Q
WHERE Q.sstate = 'WA'
EXCEPT
(SELECT P.sno
FROM Supply P
WHERE P.price > 100
AND Q.sno not in
(SELECT P.sno
FROM Supply P
WHERE P.price > 100))
How about Subqueries?

\[
(\text{SELECT Q.sno FROM Supplier Q WHERE Q.sstate = 'WA'})
\text{EXCEPT}
(\text{SELECT P.sno FROM Supply P WHERE P.price > 100})
\]

Finally…

\[
\begin{align*}
\Pi_{sno} & \quad \Pi_{sno} \\
\sigma_{\text{sstate}='WA'} & \quad \sigma_{\text{Price} > 100} \\
\text{Supplier} & \quad \text{Supply}
\end{align*}
\]
Summary of RA and SQL

- **SQL** = a declarative language where we say *what* data we want to retrieve
- **RA** = an algebra where we say *how* we want to retrieve the data
- **Theorem**: SQL and RA can express exactly the same class of queries

RDBMS translates SQL $\rightarrow$ RA, then optimizes RA
Summary of RA and SQL

• SQL (and RA) cannot express ALL queries that we could write in, say, Java

• Example:
  – Parent(p,c): find all descendants of ‘Alice’
  – No RA query can compute this!
  – This is called a recursive query

• Next lecture: Datalog is an extension that can compute recursive queries
Introduction to Data Management
CSE 344

Lectures 9-10: Datalog
Class Overview

- Unit 1: Intro
- Unit 2: Relational Data Models and Query Languages
  - Data models, SQL, Datalog, Relational Algebra
- Unit 3: Non-relational data
- Unit 4: RDMBS internals and query optimization
- Unit 5: Parallel query processing
- Unit 6: DBMS usability, conceptual design
- Unit 7: Transactions
What is Datalog?

• Another query language for relational model
  – Designed in the 80’s
  – Simple, concise, elegant
  – Extends relational queries with recursion

• Today is a hot topic:
  – Souffle (we will use in HW4)
  – Eve http://witheve.com/
  – Differential datalog
    https://github.com/frankmcsherry/differential-dataflow
  – Beyond databases in many research projects:
    network protocols, static program analysis
• Open-source implementation of Datalog DBMS
• Under active development
• Commercial implementations are available
  – More difficult to set up and use
• “sqlite” of Datalog
  – Set-based rather than bag-based

• Install in your VM
  – Run `sudo yum install souffle` in terminal
  – More details in upcoming HW4
Why bother with *yet* another relational query language?
Example: storing FB friends

As a graph

As a relation

Peter

Mary

John

Phil

<table>
<thead>
<tr>
<th>Person1</th>
<th>Person2</th>
<th>is_friend</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peter</td>
<td>John</td>
<td>1</td>
</tr>
<tr>
<td>John</td>
<td>Mary</td>
<td>0</td>
</tr>
<tr>
<td>Mary</td>
<td>Phil</td>
<td>1</td>
</tr>
<tr>
<td>Phil</td>
<td>Peter</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

We will learn the tradeoffs of different data models later this quarter
Compute your friends graph

Datalog allows us to write recursive queries easily

When does it end???
Datalog: Facts and Rules

Facts = tuples in the database

Rules = queries

Schema

Actor(id, fname, lname)
Casts(pid, mid)
Movie(id, name, year)
Datalog: Facts and Rules

Facts = tuples in the database

Rules = queries

Actor(id, fname, lname)
Casts(pid, mid)
Movie(id, name, year)

.decl Actor(id:number, fname:symbol, lname:symbol)
.decl Casts(id:number, mid:number)
.decl Movie(id:number, name:symbol, year:number)

Actor(344759, ‘Douglas’, ‘Fowley’).
Casts(344759, 29851).
Casts(355713, 29000).
Movie(29445, ‘Ave Maria’, 1940).

Table declaration
Types in Souffle:
number
symbol (aka varchar)
Insert data
Datalog: Facts and Rules

Facts = tuples in the database

Rules = queries

Actor(344759, ‘Douglas’, ‘Fowley’).
Casts(344759, 29851).
Casts(355713, 29000).
Movie(29445, ‘Ave Maria’, 1940).

Q1(y) :- Movie(x,y,z), z=1940.
Datalog: Facts and Rules

Facts = tuples in the database

Actor(344759, ‘Douglas’, ‘Fowley’).
Casts(344759, 29851).
Casts(355713, 29000).
Movie(29445, ‘Ave Maria’, 1940).

Rules = queries

Q1(y) :- Movie(x,y,z), z=1940.

Find Movies made in 1940
Datalog: Facts and Rules

Facts = tuples in the database

Rules = queries

Actor(344759,'Douglas', 'Fowley').
Casts(344759, 29851).
Casts(355713, 29000).
Movie(7909, 'A Night in Armour', 1910).
Movie(29000, 'Arizona', 1940).
Movie(29445, 'Ave Maria', 1940).

Q1(y) :- Movie(x,y,z), z=1940.

SQL
SELECT name
FROM Movie
WHERE year = 1940

Find Movies made in 1940
Datalog: Facts and Rules

**Facts** = tuples in the database

- Actor(344759, ‘Douglas’, ‘Fowley’).
- Casts(344759, 29851).
- Casts(355713, 29000).

**Rules** = queries

Q1(y) :- Movie(x, y, z), z=1940.

Order of variable matters!

Find Movies made in 1940
Datalog: Facts and Rules

**Facts** = tuples in the database

Actor(344759, ‘Douglas’, ‘Fowley’).
Casts(344759, 29851).
Casts(355713, 29000).
Movie(29445, ‘Ave Maria’, 1940).

**Rules** = queries

Q1(y) :- Movie(iDontCare, y, z), z=1940.

Find Movies made in 1940
Facts = tuples in the database

Rules = queries

Actor(344759, ‘Douglas’, ‘Fowley’).
Casts(344759, 29851).
Casts(355713, 29000).
Movie(29445, ‘Ave Maria’, 1940).

Q1(y) :- Movie(_, y, z), z=1940.

_ = “don’t care” variables

Find Movies made in 1940
Datalog: Facts and Rules

Facts = tuples in the database

Actor(344759, ‘Douglas’, ‘Fowley’).
Casts(344759, 29851).
Casts(355713, 29000).
Movie(29445, ‘Ave Maria’, 1940).

Rules = queries

Q1(y) :- Movie(x, y, z), z = 1940.
Q2(f, l) :- Actor(z, f, l), Casts(z, x), Movie(x, y, 1940).
Datalog: Facts and Rules

Facts = tuples in the database

Actor(344759, ‘Douglas’, ‘Fowley’).
Casts(344759, 29851).
Casts(355713, 29000).
Movie(29445, ‘Ave Maria’, 1940).

Rules = queries

Q1(y) :- Movie(x, y, z), z = 1940.
Q2(f, l) :- Actor(z, f, l), Casts(z, x), Movie(x, y, 1940).
Datalog: Facts and Rules

**Facts** = tuples in the database

- Actor(344759, ‘Douglas’, ‘Fowley’).
- Casts(344759, 29851).
- Casts(355713, 29000).

**Rules** = queries

- Q1(y) :- Movie(x, y, z), z = 1940.
- Q2(f, l) :- Actor(z, f, l), Casts(z, x), Movie(x, y, 1940).

Find Actors who acted in Movies made in 1940
**Facts** = tuples in the database

- **Actor** (id, fname, lname)
- **Casts** (pid, mid)
- **Movie** (id, name, year)

**Rules** = queries

- **Q1(y)** :- Movie(x, y, z), z=1940.
- **Q2(f,l)** :- Actor(z, f, l), Casts(z, x), Movie(x, y, 1940).
- **Q3(f,l)** :- Actor(z, f, l), Casts(z, x1), Movie(x1, y1, 1910), Casts(z, x2), Movie(x2, y2, 1940).

Sample facts:
- Actor(344759, ‘Douglas’, ‘Fowley’).
- Casts(344759, 29851).
- Casts(355713, 29000).
Datalog: Facts and Rules

**Facts** = tuples in the database

- Actor(344759, ‘Douglas’, ‘Fowley’).
- Casts(344759, 29851).
- Casts(355713, 29000).

**Rules** = queries

- **Q1(y)**: - Movie(x, y, z), z=1940.
- **Q2(f,l)**: - Actor(z, f, l), Casts(z, x), Movie(x, y, 1940).
- **Q3(f,l)**: - Actor(z, f, l), Casts(z, x1), Movie(x1, y1, 1910), Casts(z, x2), Movie(x2, y2, 1940).

Find Actors who acted in a Movie in 1940 and in one in 1910
Datalog: Facts and Rules

**Facts** = tuples in the database

- Actor(344759, ‘Douglas’, ‘Fowley’).
- Casts(344759, 29851).
- Casts(355713, 29000).

**Rules** = queries

- Q1(y) :- Movie(x,y,z), z=1940.
- Q2(f,l) :- Actor(z,f,l), Casts(z,x), Movie(x,y,1940).
- Q3(f,l) :- Actor(z,f,l), Casts(z,x1), Movie(x1,y1,1910), Casts(z,x2), Movie(x2,y2,1940).

**Extensional Database Predicates** = EDB = Actor, Casts, Movie

**Intensional Database Predicates** = IDB = Q1, Q2, Q3
Datalog: Terminology

Q2(f, l) :- Actor(z, f, l), Casts(z, x), Movie(x, y, 1940).

f, l = head variables
x, y, z = existential variables
More Datalog Terminology

Q(args) :- R1(args), R2(args), ...

- $R_i(args_i)$ called an **atom**, or a **relational predicate**
- $R_i(args_i)$ evaluates to true when relation $R_i$ contains the tuple described by $args_i$.
  - Example: $\text{Actor}(344759, 'Douglas', 'Fowley')$ is true
- In addition we can also have arithmetic predicates
  - Example: $z > 1940$.
- Book uses AND instead of ,

Q(args) :- R1(args) AND R2(args) ...
Datalog program

• A Datalog program consists of several rules
• Importantly, rules may be recursive!
  – Recall CSE 143!
• Usually there is one distinguished predicate that’s the output
• We will show an example first, then give the general semantics.
R encodes a graph
e.g., connected cities

$R = \begin{array}{c|c}
1 & 2 \\
2 & 1 \\
2 & 3 \\
1 & 4 \\
3 & 4 \\
4 & 5 \\
\end{array}$
Example

\[ R = T(x,y) : R(x,y). \]
\[ T(x,y) : R(x,z), T(z,y). \]

R encodes a graph e.g., connected cities

Multiple rules for the same IDB means OR

What does it compute?
R encodes a graph e.g., connected cities

Example

\[ T(x,y) :- R(x,y). \]
\[ T(x,y) :- R(x,z), T(z,y). \]

Initially: T is empty.

\[
\begin{array}{c|c}
1 & 2 \\
2 & 1 \\
2 & 3 \\
1 & 4 \\
3 & 4 \\
4 & 5 \\
\end{array}
\]

What does it compute?
R encodes a graph e.g., connected cities

Initially:
T is empty.

First iteration:
T =

First rule generates this

Second rule generates nothing (because T is empty)

Example

T(x,y) :- R(x,y).
T(x,y) :- R(x,z), T(z,y).

What does it compute?

R =

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

Example

T(x,y) :- R(x,y).
T(x,y) :- R(x,z), T(z,y).

What does it compute?

R encodes a graph e.g., connected cities

Initially:
T is empty.

First iteration:
T =

First rule generates this

Second rule generates nothing (because T is empty)
R encodes a graph e.g., connected cities

Example

\[
\begin{align*}
T(x, y) & :- R(x, y). \\
T(x, y) & :- R(x, z), T(z, y).
\end{align*}
\]

First iteration:
\[
T = \begin{array}{cccc}
1 & 2 \\
2 & 1 \\
2 & 3 \\
1 & 4 \\
3 & 4 \\
4 & 5 \\
\end{array}
\]

Second iteration:
\[
T = \begin{array}{cccc}
1 & 2 \\
2 & 1 \\
2 & 3 \\
1 & 4 \\
3 & 4 \\
4 & 5 \\
1 & 1 \\
2 & 2 \\
1 & 3 \\
2 & 4 \\
1 & 5 \\
3 & 5 \\
\end{array}
\]

What does it compute?
Example

\[ T(x, y) \text{ :- } R(x, y) \]
\[ T(x, y) \text{ :- } R(x, z), T(z, y) \]

What does it compute?

Initially:
\[ T \text{ is empty.} \]

First iteration:
\[ T = \]

Second iteration:
\[ T = \]

Third iteration:
\[ T = \]

Both rules

First rule

Second rule

New fact

The rule \[ R \text{ encodes a graph e.g., connected cities} \]
Example

R encodes a graph e.g., connected cities

\[ R = \]

1 2
2 1
2 3
1 4
3 4
4 5

Initially: T is empty.

\[ T(1,2) \]
\[ T(2,1) \]
\[ T(2,3) \]
\[ T(1,4) \]
\[ T(3,4) \]
\[ T(4,5) \]

What does it compute?

\[ T(x,y) : R(x,y). \]
\[ T(x,y) : R(x,z), T(z,y). \]

First iteration: T =

Second iteration: T =

Third iteration: T =

Fourth iteration:

T = (same)

No new facts.

DONE
Datalog Semantics

Fixpoint semantics

• Start:
  \[ \text{IDB}_0 = \text{empty relations} \]
  \[ t = 0 \]

Repeat:
  \[ \text{IDB}_{t+1} = \text{Compute Rules(EDB, IDB}_t) \]
  \[ t = t+1 \]

Until \[ \text{IDB}_t = \text{IDB}_{t-1} \]

• Remark: since rules are monotone:
  \[ \emptyset = \text{IDB}_0 \subseteq \text{IDB}_1 \subseteq \text{IDB}_2 \subseteq \ldots \]

• It follows that a datalog program w/o functions (+, *, ...) always terminates. (Why?)
Three Equivalent Programs

R encodes a graph
e.g., connected cities

\[ T(x,y) :\neg R(x,y). \]
\[ T(x,y) :\neg R(x,z), T(z,y). \]
\[ T(x,y) :\neg R(x,y). \]
\[ T(x,y) :\neg T(x,z), R(z,y). \]
\[ T(x,y) :\neg R(x,y). \]
\[ T(x,y) :\neg T(x,z), T(z,y). \]

Question: which terminates in fewest iterations?
More Features

• Aggregates

• Grouping

• Negation
Aggregates

[aggregate name] <var> : { [relation to compute aggregate on] }

\[ \text{min} \ x : \{ \text{Actor}(x, y, _), \ y = 'John' \} \]

\[ Q(\text{minId}) : \text{minId} = \text{min} \ x : \{ \text{Actor}(x, y, _), \ y = 'John' \} \]

Assign variable to the value of the aggregate

Meaning (in SQL)

\[
\text{SELECT} \ \text{min}(id) \ \text{as minId} \\
\text{FROM} \ \text{Actor} \ \text{as a} \\
\text{WHERE} \ a.\text{name} = 'John'
\]
Counting

Q(c) :- c = \texttt{count} : \{ \texttt{Actor}(\_ , y , \_), y = ‘John’ \}

Meaning (in SQL, assuming no NULLs)

```sql
SELECT count(*) as c
FROM Actor as a
WHERE a.name = ‘John’
```
Grouping

\[ Q(y,c) :\quad \text{Movie}(\_,\_,y), \ c = \text{count} : \{ \text{Movie}(\_,\_,y) \} \]

**Meaning (in SQL)**

```sql
SELECT m.year, count(*)
FROM Movie as m
GROUP BY m.year
```
Examples

A genealogy database (parent/child)

ParentChild

<table>
<thead>
<tr>
<th>p</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>Carol</td>
</tr>
<tr>
<td>Bob</td>
<td>Carol</td>
</tr>
<tr>
<td>Bob</td>
<td>David</td>
</tr>
<tr>
<td>Carol</td>
<td>Eve</td>
</tr>
</tbody>
</table>

...
Count Descendants

For each person, count his/her descendants

ParentChild(p,c)
Count Descendants

For each person, count his/her descendants

Answer

<table>
<thead>
<tr>
<th>p</th>
<th>cnt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>4</td>
</tr>
<tr>
<td>Bob</td>
<td>5</td>
</tr>
<tr>
<td>Carol</td>
<td>3</td>
</tr>
<tr>
<td>David</td>
<td>2</td>
</tr>
<tr>
<td>Fred</td>
<td>1</td>
</tr>
</tbody>
</table>
Count Descendants

For each person, count his/her descendants

Answer

<table>
<thead>
<tr>
<th>p</th>
<th>cnt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>4</td>
</tr>
<tr>
<td>Bob</td>
<td>5</td>
</tr>
<tr>
<td>Carol</td>
<td>3</td>
</tr>
<tr>
<td>David</td>
<td>2</td>
</tr>
<tr>
<td>Fred</td>
<td>1</td>
</tr>
</tbody>
</table>

Note: Eve and George do not appear in the answer (why?)
Count Descendants

For each person, compute the total number of descendants

// for each person, compute his/her descendants
Count Descendants

For each person, compute the total number of descendants

// for each person, compute his/her descendants
D(x,y) :- ParentChild(x,y).
Count Descendants

For each person, compute the total number of descendants

// for each person, compute his/her descendants
D(x,y) :- ParentChild(x,y).
D(x,z) :- D(x,y), ParentChild(y,z).
Count Descendants

For each person, compute the total number of descendants

// for each person, compute his/her descendants
D(x,y) :- ParentChild(x,y).
D(x,z) :- D(x,y), ParentChild(y,z).

// For each person, count the number of descendants
Count Descendants

For each person, compute the total number of descendants

// for each person, compute his/her descendants
D(x,y) :- ParentChild(x,y).
D(x,z) :- D(x,y), ParentChild(y,z).

// For each person, count the number of descendants
T(p,c) :- D(p,_), c = count : { D(p,y) }. 

ParentChild(p,c)
How many descendants does Alice have?

// for each person, compute his/her descendants
D(x,y) :- ParentChild(x,y).
D(x,z) :- D(x,y), ParentChild(y,z).

// For each person, count the number of descendants
T(p,c) :- D(p,:), c = count : { D(p,y) }.
Count Descendants

How many descendants does Alice have?

// for each person, compute his/her descendants
D(x,y) :- ParentChild(x,y).
D(x,z) :- D(x,y), ParentChild(y,z).

// For each person, count the number of descendants
T(p,c) :- D(p,_), c = count : { D(p,y) }.

// Find the number of descendants of Alice

ParentChild(p,c)
Count Descendants

How many descendants does Alice have?

// for each person, compute his/her descendants
D(x,y) :- ParentChild(x,y).
D(x,z) :- D(x,y), ParentChild(y,z).

// For each person, count the number of descendants
T(p,c) :- D(p,_), c = count : { D(p,y) }.

// Find the number of descendants of Alice
Q(d) :- T(p,d), p = “Alice”.
Negation: use “!”

Find all descendants of Bob that are not descendants of Alice
Negation: use “!”

Find all descendants of Bob that are not descendants of Alice

Answer

<table>
<thead>
<tr>
<th>x</th>
</tr>
</thead>
<tbody>
<tr>
<td>David</td>
</tr>
</tbody>
</table>
Negation: use “!”

Find all descendants of Bob that are not descendants of Alice

// for each person, compute his/her descendants
D(x,y) :- ParentChild(x,y).
D(x,z) :- D(x,y), ParentChild(y,z).
Negation: use “!”

Find all descendants of Bob that are not descendants of Alice

// for each person, compute his/her descendants
D(x,y) :- ParentChild(x,y).
D(x,z) :- D(x,y), ParentChild(y,z).

// Compute the answer: notice the negation
Q(x) :- D(“Bob”,x), !D(“Alice”,x).
Two people are in the *same generation* if they are descendants at the same generation of some common ancestor.
Same Generation

Compute pairs of people at the same generation

// common parent
Same Generation

Compute pairs of people at the same generation

```prolog
// common parent
SG(x,y) :- ParentChild(p,x), ParentChild(p,y)
```
Same Generation

Compute pairs of people at the same generation

// common parent
SG(x,y) :- ParentChild(p,x), ParentChild(p,y)

// parents at the same generation
Same Generation

Compute pairs of people at the same generation

// common parent
SG(x,y) :- ParentChild(p,x), ParentChild(p,y)

// parents at the same generation
SG(x,y) :- ParentChild(p,x), ParentChild(q,y), SG(p,q)
Same Generation

Compute pairs of people at the same generation

// common parent
SG(x,y) :- ParentChild(p,x), ParentChild(p,y)

// parents at the same generation
SG(x,y) :- ParentChild(p,x), ParentChild(q,y), SG(p,q)

Problem: this includes answers like SG(Carol, Carol)
And also SG(Eve, George), SG(George, Eve)
Same Generation

Compute pairs of people at the same generation

// common parent
SG(x,y) :- ParentChild(p,x), ParentChild(p,y), x < y

// parents at the same generation
SG(x,y) :- ParentChild(p,x), ParentChild(q,y),
         SG(p,q), x < y
Here are *unsafe* datalog rules. What’s “unsafe” about them?

\[
\begin{align*}
U1(x,y) & : - \text{ParentChild(“Alice”,x), } y \neq \text{“Bob”} \\
U2(x) & : - \text{ParentChild(“Alice”,x), !ParentChild(x,y)} \\
U3(\text{minId, y}) & : - \text{minId} = \text{min } x : \{ \text{Actor(x, y, _)} \}
\end{align*}
\]
Here are *unsafe* datalog rules. What’s “unsafe” about them?

\[
U1(x,y) :- \text{ParentChild(“Alice”,x)}, \ y \neq \text{“Bob”}
\]

\[
U2(x) :- \text{ParentChild(“Alice”,x)}, \neg \text{ParentChild(x,y)}
\]

\[
U3(\text{minId}, y) :- \text{minId} = \min x : \{ \text{Actor(x, y, _)} \}
\]
Here are *unsafe* datalog rules. What’s “unsafe” about them?

\[
\begin{align*}
U1(x,y) & : \text{ParentChild(“Alice”,x), } y \neq \text{“Bob”} \\
U2(x) & : \text{ParentChild(“Alice”,x), !ParentChild(x,y)} \\
U3(minId, y) & : \text{minId = min x : \{ Actor(x, y, _) \}}
\end{align*}
\]

Holds for every y other than “Bob”

\( U1 = \text{infinite!} \)
Here are *unsafe* datalog rules. What’s “unsafe” about them?

\[
\text{U1}(x,y) : - \text{ParentChild(“Alice”,x)}, \ y \neq \text{“Bob”}
\]

\[
\text{U2}(x) : - \text{ParentChild(“Alice”,x)}, \neg \text{ParentChild(x,y)}
\]

Want Alice’s childless children, but we get all children x (because there exists some y that x is not parent of y).

\[
\text{U3}(\text{minId}, y) : - \text{minId} = \min x : \{ \text{Actor(x, y, _)} \}
\]

Unclear what y is.
Here are *unsafe* datalog rules. What’s “unsafe” about them?

\[
\text{U1}(x,y) :- \text{ParentChild}(\text{“Alice”},x), \ y \neq \text{“Bob”}
\]

\[
\text{U2}(x) :- \text{ParentChild}(\text{“Alice”},x), \ \neg \text{ParentChild}(x,y)
\]

A datalog rule is *safe* if every variable appears in some positive, non-aggregated relational atom

\[
\text{U3}(\text{minId}, y) :- \text{minId} = \text{min} \ x : \{ \text{Actor}(x, y, _) \}
\]
Stratified Datalog

• Recursion does not cope well with aggregates or negation
• Example: what does this mean?

\[
\begin{align*}
A() & : - !B(). \\
B() & : - !A().
\end{align*}
\]

• A datalog program is **stratified** if it can be partitioned into *strata*
  – Only IDB predicates defined in strata 1, 2, ..., n may appear under ! or agg in stratum n+1.

• Many Datalog DBMSs (including souffle) accepts only stratified Datalog.
Stratified Datalog

Stratum 1

Stratum 2

May use D in an agg since it was defined in previous stratum
Stratified Datalog

<table>
<thead>
<tr>
<th>Stratum 1</th>
<th>Stratum 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>D(x,y) :- ParentChild(x,y).</td>
<td></td>
</tr>
<tr>
<td>D(x,z) :- D(x,y), ParentChild(y,z).</td>
<td></td>
</tr>
<tr>
<td>T(p,c) :- D(p,_), c = count : { D(p,y) }.</td>
<td></td>
</tr>
<tr>
<td>Q(d) :- T(p,d), p = &quot;Alice&quot;.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stratum 1</th>
<th>Stratum 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>D(x,y) :- ParentChild(x,y).</td>
<td></td>
</tr>
<tr>
<td>D(x,z) :- D(x,y), ParentChild(y,z).</td>
<td></td>
</tr>
<tr>
<td>Q(x) :- D(&quot;Alice&quot;,x), !D(&quot;Bob&quot;,x).</td>
<td></td>
</tr>
</tbody>
</table>

Non-stratified

A() :- !B().
B() :- !A().

May use D in an agg since it was defined in previous stratum

May use !D

Cannot use !A
Stratified Datalog

• If we don’t use aggregates or negation, then the Datalog program is already stratified

• If we do use aggregates or negation, it is usually quite natural to write the program in a stratified way