Slides to be used with Ullman/Widom *A First Course in Database Systems*, Prentice-Hall, 1997

Plan

1. Information integration: important new application that motivates what follows.
2. Semistructured data: a new data model designed to cope with problems of information integration.
3. XML: a new Web standard that is essentially semistructured data.

Information Integration

Problem: related data exists in many places. They talk about the same things, but differ in model, schema, conventions (e.g., terminology).

Example

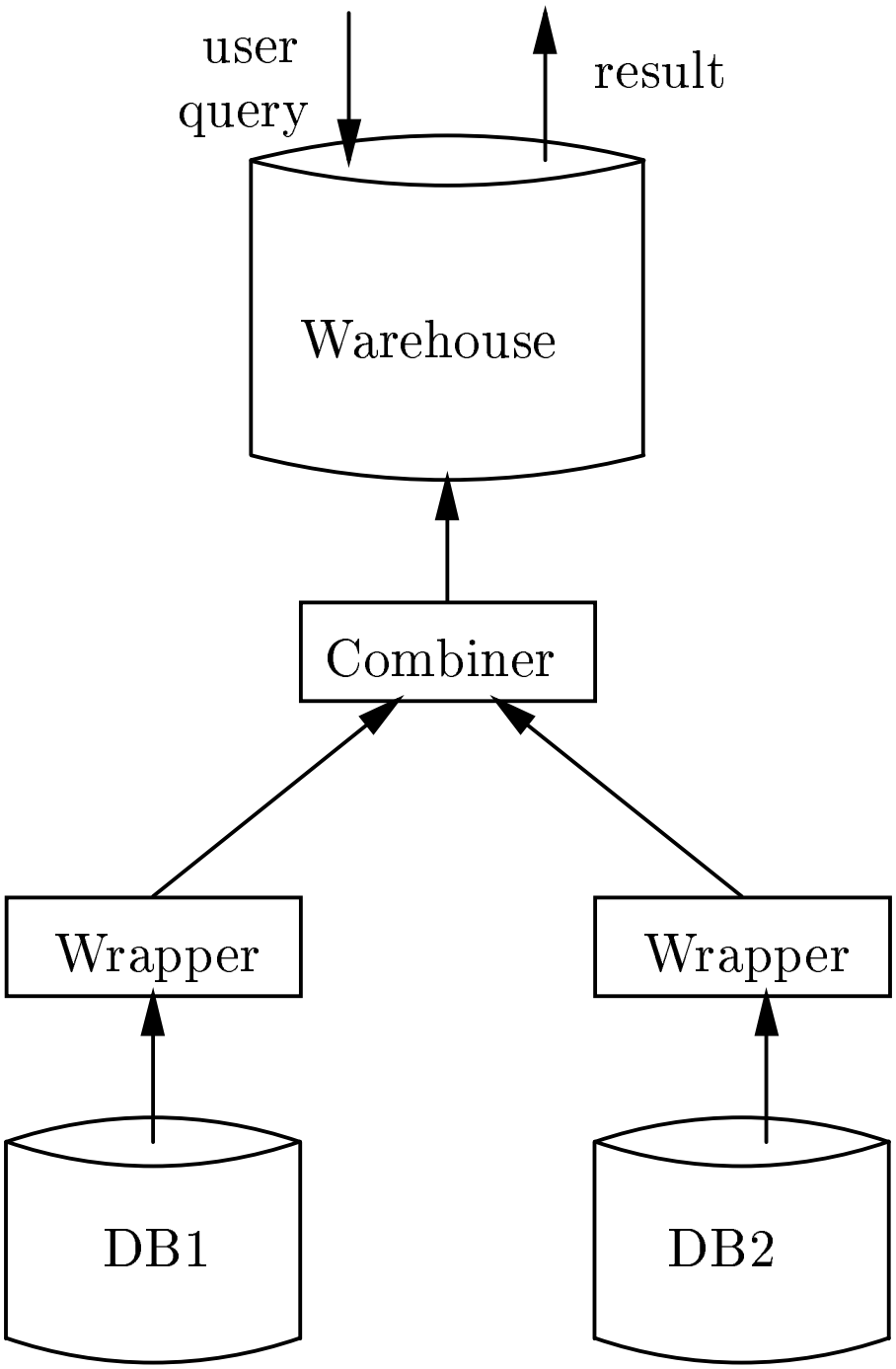
In the real world, every bar has its own database.

- Some may have relations like beer-price; others have an MS-word file from which the menu is printed.
- Some keep phones of manufacturers but not addresses.
- Some distinguish beers and ales; others do not.

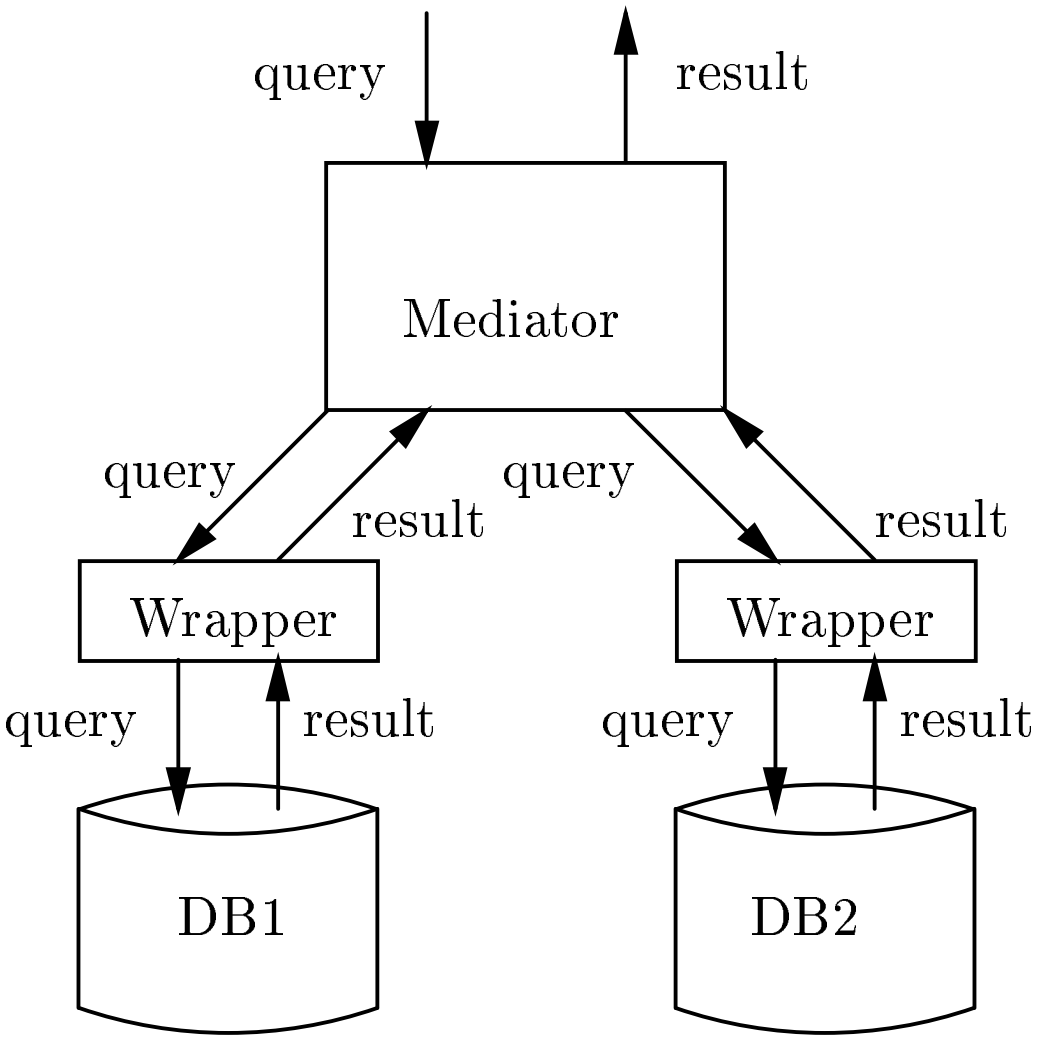
Two approaches

1. *Warehousing*: Make copies of information at each data source centrally.
   - ❖ Reconstruct data daily/weekly/monthly, but do not try to keep it up-to-date.
2. *Mediation*: Create a view of all information, but do not make copies.
   - ❖ Answer queries by sending appropriate queries to sources.

Warehousing



Mediation


query
result
Mediator
query
result
query
result
Wrapper
Wrapper
query
result
query
result
DB1
DB2

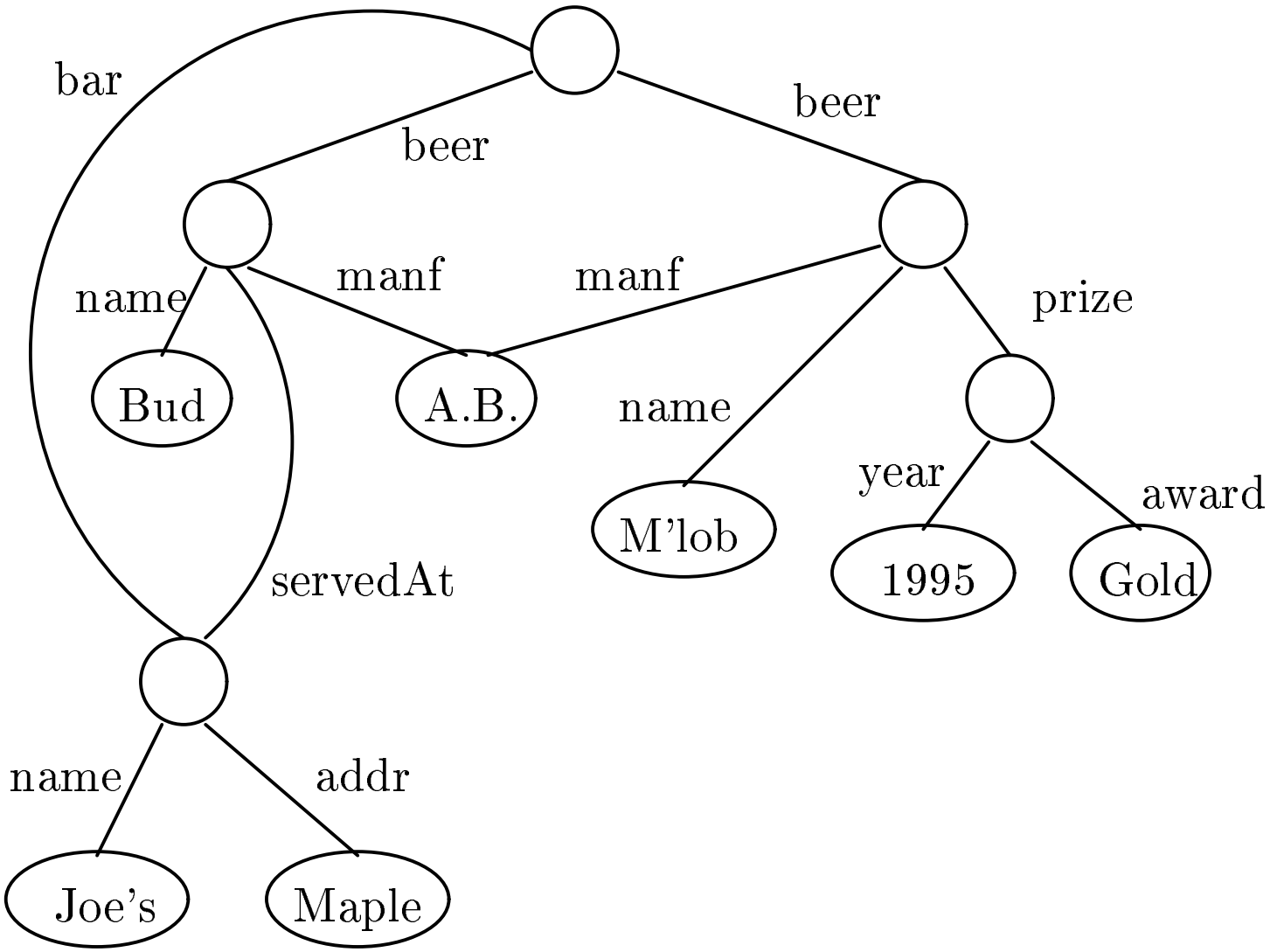
Semistructured Data

- A different kind of data model, more suited to information-integration applications than either relational or OO.
  - ❖ Think of "objects," but with the type of an object its own business rather than the business of the class to which it belongs.
  - ❖ Allows information from several sources, with related but different properties, to be fit together in one whole.
- Up-and-coming application: XML documents.

Object-Exchange Model (OEM)

- Stanford's model for semistructured data.
- Nodes = objects.
- Nodes connected in a general rooted graph structure.
- Two variations: labels on nodes or on arcs.
  - ❖ We'll use the latter.
  - ❖ Atomic values on leaf nodes in either variation.
- Big deal: no restriction on labels (roughly = attributes).
  - ❖ Zero, one, or many children of a given label type are all OK.

Example


bar
beer
beer
name
manf
manf
prize
Bud
A.B.
name
year
award
M'lob
1995
Gold
servedAt
name
addr
Joe's
Maple

Querying Semistructured Data

- LOREL = "LORE Language."
  - ❖ LORE = "Lightweight Object Repository," Prof. Widom's main research project.
- Similar to OQL.
  - ❖ But typeless; a path binds to part of the query exactly when there are labels that match.

Example

Find the bars that serve beers that won gold medals in 1995.

```
SELECT B.servedAt.name
FROM beer B, B.prize P
WHERE P.award = "Gold" AND
    P.year = 1995;
```

Example

Find the manufacturers of beers served by Joe's bar.

```
SELECT beer.manf
WHERE beer.servedAt.name =
    "Joe's Bar";
```

- Notice that the from-clause is unnecessary sometimes.
  - ❖ Label sequences with a common prefix must have that prefix bound to the same path in any solution to the query.

XML (Extensible Markup Language)

- Next generation of HTML.
- Two modes:
  1. *Well-formed* XML allows you to invent your own tags, much like labels in semistructured data.
  2. *Valid* XML involves a DTD (Document Type Definition) that tells the labels and gives a grammar for how they may be nested.
- As with LORE, a query on well-formed XML specifies "interesting" tags and ignores others.
  - ❖ In a sense, it becomes the document creator's job to anticipate the queries.

Well-Formed XML

1. Declaration = `<? ... ?>`.
   - ❖ Normal declaration is `<? XML VERSION = "1.0" STANDALONE = "yes" ?>`
   - ❖ "Standalone" means that there is no DTD specified.
2. *Root* tag surrounds the entire balance of the document.
   - ❖ `<FOO>` is balanced by `</FOO>`, as in HTML.
3. Any balanced structure of tags OK.
   - ❖ Option of tags that don't *require* balance, like `<P>` in HTML.

Example

```
<?XML VERSION = "1.0" STANDALONE = "yes"?>
<BARS>
    <BAR><NAME>Joe's Bar</NAME>
        <BEER><NAME>Bud</NAME>
            <PRICE>2.50</PRICE></BEER>
        <BEER><NAME>Miller</NAME>
            <PRICE>3.00</PRICE></BEER>
    </BAR>
    <BAR> ...
</BARS>
```

Document Type Definitions (DTD)

Essentially a grammar describing the legal nesting of tags.

- Intention is that DTD's will be standards for a domain, used by everyone preparing or using data in that domain.
  - ❖ Example: a DTD for describing protein structure; a DTD for describing bar menus, etc.

Gross Structure of a DTD

```
<!DOCTYPE root tag [
    <!ELEMENT name (components)>
    more elements
]>
```

Elements of a DTD

An *element* is a name (its tag) and a parenthesized description of tags within an element.

- Special case: `(#PCDATA)` after an element name means it is text.

Example

```
<!DOCTYPE Bars [
    <!ELEMENT BARS (BAR*)>
    <!ELEMENT BAR (NAME, BEER+)>
    <!ELEMENT NAME (#PCDATA)>
    <!ELEMENT BEER (NAME, PRICE)>
    <!ELEMENT PRICE (#PCDATA)>
]>
```

Components

- Each element name is a tag.
- Its components are the tags that may appear nested within.
  - ❖ But order is not required, even if there is a list of components.
- Multiplicity of a tag is controlled by:
  a) `*` = zero or more of.
  b) `+` = one or more of.
  c) `?` = zero or one of.
- In addition, `|` = "or."

Using a DTD

1. Set `STANDALONE = "no"`.
2. Either
   a) Include the DTD as a preamble, or
   b) Follow the `XML` tag by a `DOCTYPE` declaration with the root tag, the keyword `SYSTEM`, and a file where the DTD can be found.

Example of (a)

```
<?XML VERSION = "1.0" STANDALONE = "no"?>

<!DOCTYPE Bars [
    <!ELEMENT BARS (BAR*)>
    <!ELEMENT BAR (NAME, BEER+)>
    <!ELEMENT NAME (#PCDATA)>
    <!ELEMENT BEER (NAME, PRICE)>
    <!ELEMENT PRICE (#PCDATA)>
]>

<BARS>
    <BAR><NAME>Joe's Bar</NAME>
        <BEER><NAME>Bud</NAME>
            <PRICE>2.50</PRICE></BEER>
        <BEER><NAME>Miller</NAME>
            <PRICE>3.00</PRICE></BEER>
    </BAR>
    <BAR> ...
</BARS>
```

Example of (b)

Suppose our bars DTD is in file `bar.dtd` .

```
<?XML VERSION = "1.0" STANDALONE = "no"?>

<!DOCTYPE Bars SYSTEM "bar.dtd">

<BARS>
    <BAR><NAME>Joe's Bar</NAME>
        <BEER><NAME>Bud</NAME>
            <PRICE>2.50</PRICE></BEER>
        <BEER><NAME>Miller</NAME>
            <PRICE>3.00</PRICE></BEER>
    </BAR>
    <BAR> ...
</BARS>
```

That's It Folks

That's all there is.

- CS245 awaits in January.