Advanced Databases: Course Overview

Jerome Simeon

Course Staff and Information

- Instructor:
  - Jerome Simeon, IBM Research, T.J. Watson Lab
- Reach me at js1491@nyu.edu
- In our wiki you will find:

Tentative schedule

News and announcements

Reading list

Assignments

http://www.vistrails.org/index.php/Course:_Advanced_Databases

Check it often!!!

What we will cover

- Query compilation (beyond DB 101): Architecture, indices, query rewritings, new data models, distribution
- Data Integration: Architecture**s**, Schema and data mappings, addressing heterogeneity, wrappers/mediators, query decomposition
- Tentative schedule in:

http://www.vistrails.org/index.php/Course:_Advanced_Databases

Pre-Requisites

- A course in database systems, covering application programming in SQL (and other database-related languages such as OQL or XQuery)
- A course on algorithms and data structures
- Good programming skills
  - *Useful but not required: course on compilers (e.g., for programming languages)*

Readings

- Scientific papers, specific per class
- Textbooks for further study:

"Database Management Systems", by Ramakrishnan and Gehrke, McGraw-Hill, 2002. This book is for background, but we will go beyond its content.

"Query Compilers" by Guido Moerkoette at:
http://pi3.informatik.uni-mannheim.de/~moer/querycompiler.pdf
A bible of query compilation techniques, notably covering techniques that go beyond relational stores.

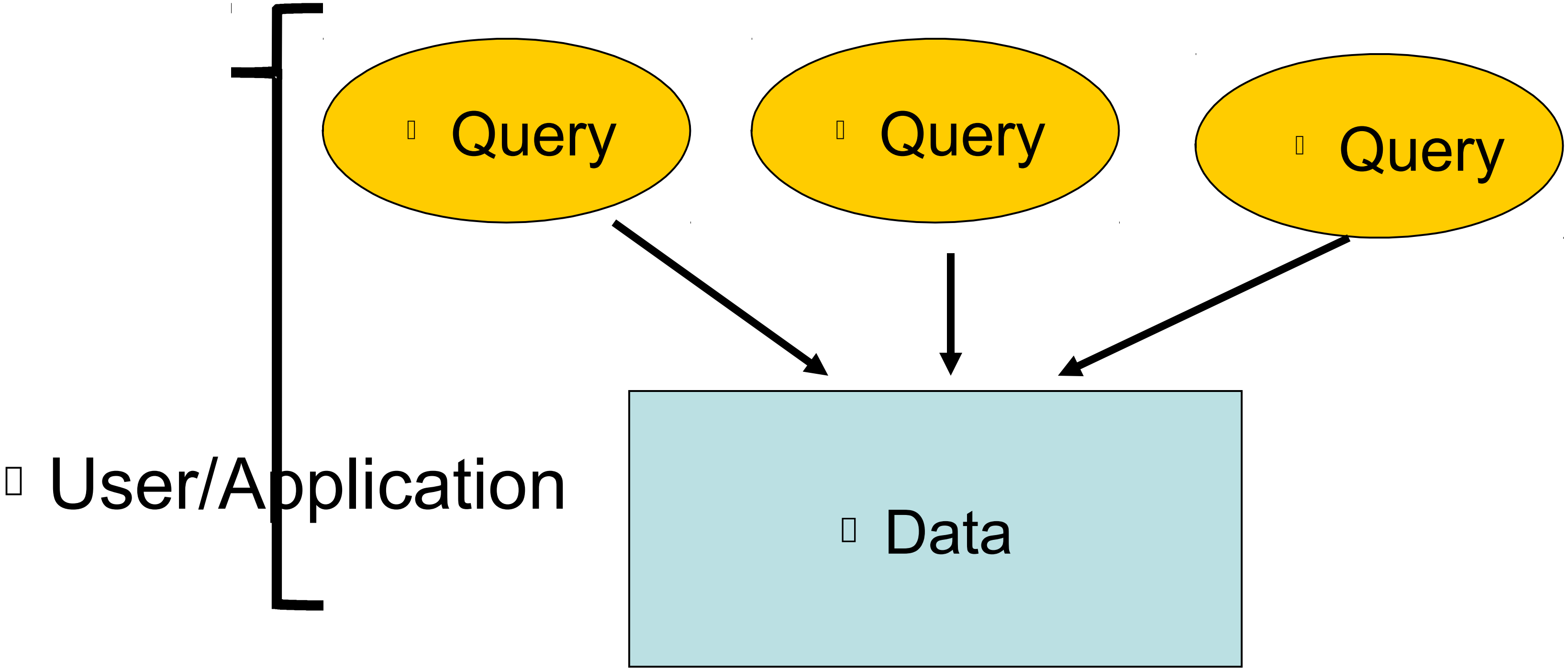
"Principles of Data Integration" by Anhai Doan, Alon Halevy, Zachary Ives. Morgan Kaufman, 2012.
A recent book on data integration that covers most technical aspects important to this area.

What you will do

- Reading assignments and review (33.3%) *done in pairs*
  - *You will need TIME and WORK*
- Quizzes (33.3%): you will use Gradiance
  - Register at http://www.newgradiance.com/services
  - Use token ***(tbd)***
- Final exam (33.3%)

Overview:
Query Compilation – Data Integration

Data Management



Query

Query

Query

User/Application

Data

- DataBase Management System (DBMS)

Example: At a Company

Query 1: Is there an employee named “Nemo”?

Query 2: What is “Nemo’s” salary?

Query 3: How many departments are there in the company?

Query 4: What is the name of “Nemo’s” department?

Query 5: How many employees are there in the “Accounts” department?

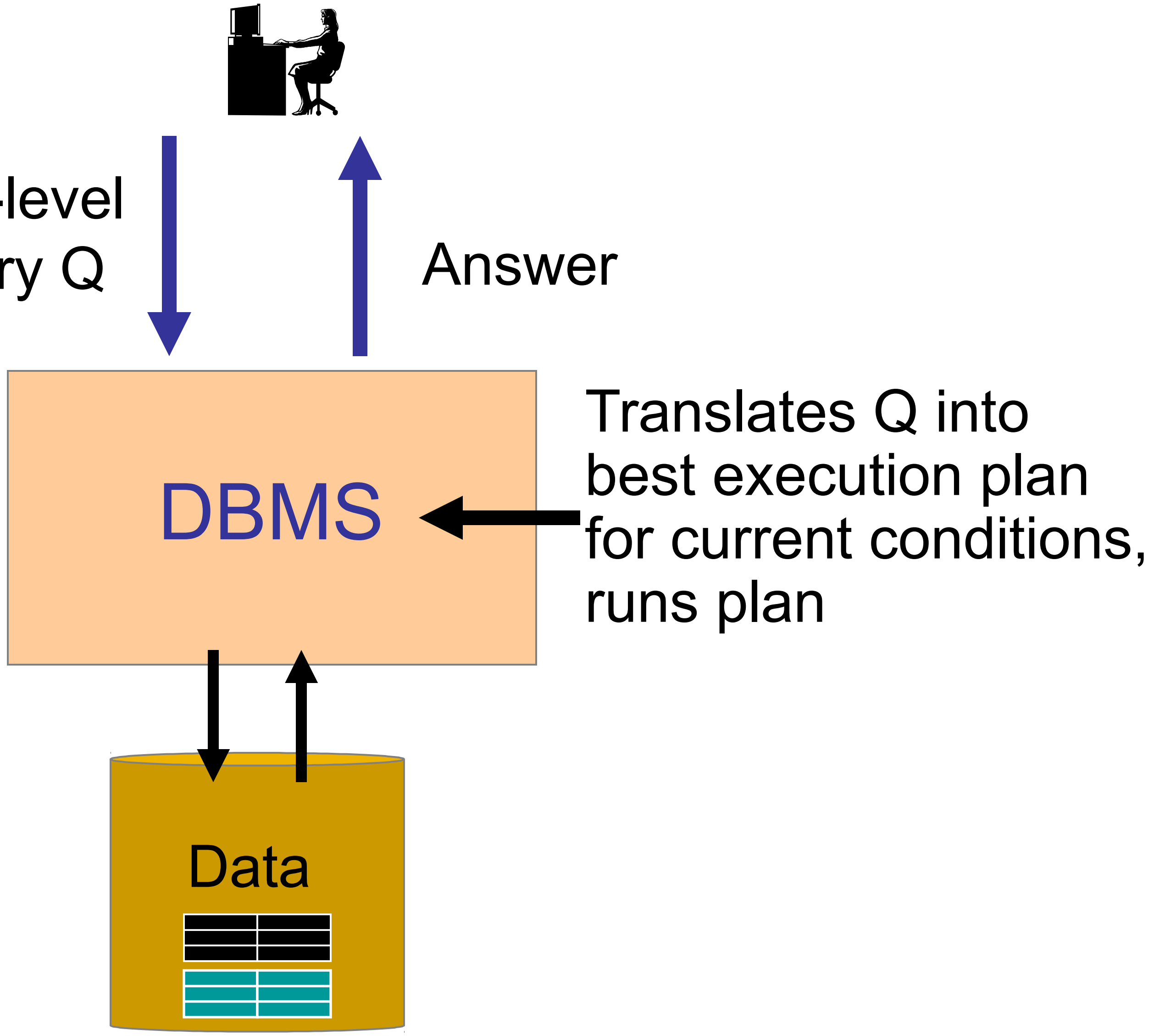
Employee

ID	Name	DeptID	Salary	…
10	Nemo	12	**120K**	…
20	Dory	156	**79K**	…
40	Gill	89	**76K**	…
52	Ray	34	**85K**	…
…	…	…	…	…

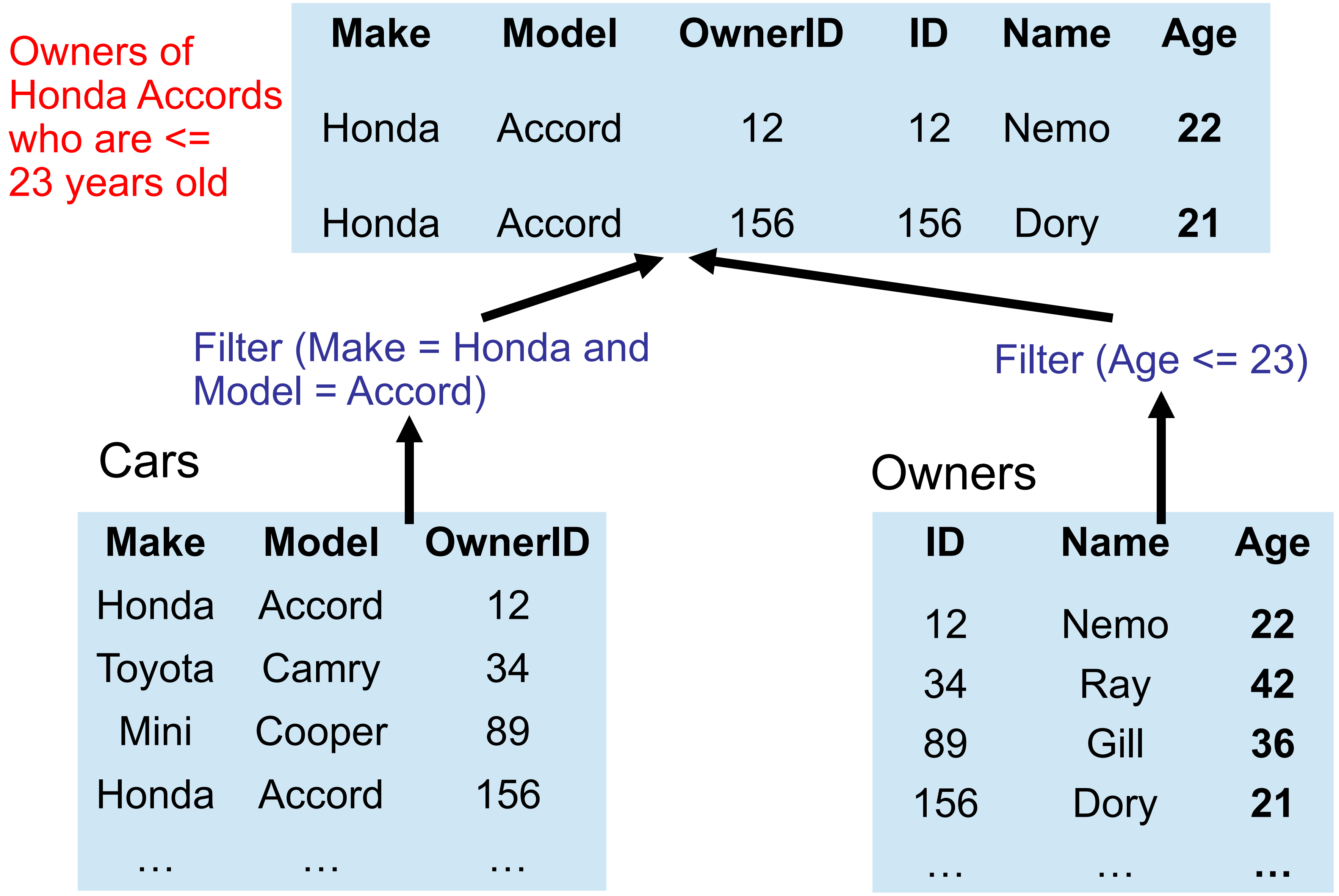
Department

ID	Name	…
12	IT	…
34	Accounts	…
89	HR	…
156	Marketing	…
…	…	…

DataBase Management System (DBMS)



Example: Store that Sells Cars



Owners of Honda Accords who are <= 23 years old

Make	Model	OwnerID	ID	Name	Age
Honda	Accord	12	12	Nemo	**22**
Honda	Accord	156	156	Dory	**21**

Filter (Make = Honda and Model = Accord)

Filter (Age <= 23)

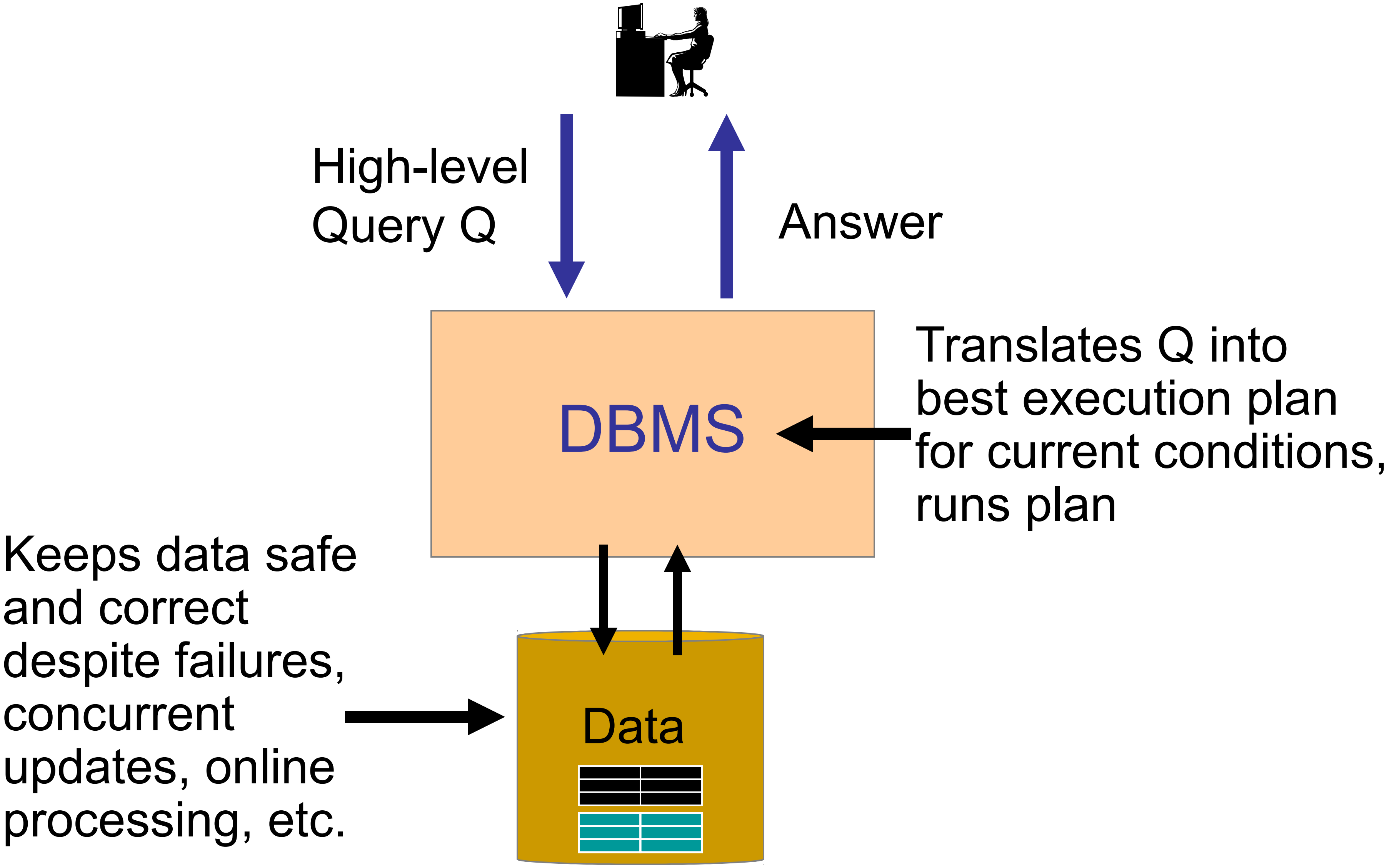
Cars

Make	Model	OwnerID
Honda	Accord	12
Toyota	Camry	34
Mini	Cooper	89
Honda	Accord	156
…	…	…

Owners

ID	Name	Age
12	Nemo	**22**
34	Ray	**42**
89	Gill	**36**
156	Dory	**21**
…	…	**…**

DataBase Management System (DBMS)



High-level Query Q

Answer

DBMS

Translates Q into best execution plan for current conditions, runs plan

Keeps data safe and correct despite failures, concurrent updates, online processing, etc.

Data

DBMS is multi-user

Example

```
Get account balance from database;
If balance > amount of withdrawal then
    balance = balance - amount of withdrawal;
    dispense cash;
    store new balance into database;
```

- Homer at ATM1 withdraws $100
- Marge at ATM2 withdraws $50
- Initial balance = $400, final balance = ?
  - Should be $250 no matter who goes first

Final balance = $250

Homer withdraws $100:

```
read balance; $400
if balance > amount then
    balance = balance - amount; $300
    write balance; $300
```

Marge withdraws $50:

```
read balance; $300
if balance > amount then
    balance = balance - amount; $250
    write balance; $250
```

Final balance = $300

```
Homer withdraws $100:              Marge withdraws $50:

read balance; $400
                                   read balance; $400
                                   If balance > amount then
                                    balance = balance - amount; $350
                                    write balance; $350


if balance > amount then
  balance = balance - amount; $300
  write balance; $300
```

Final balance = $350

```
Homer withdraws $100:                  Marge withdraws $50:

  read balance; $400
                                       read balance; $400

  if balance > amount then
   balance = balance - amount; $300
   write balance; $300

                                       if balance > amount then
                                        balance = balance - amount; $350
                                        write balance; $350
```

Concurrency control in DBMS

- Similar to concurrent programming problems
  - But data is not all in main-memory
- Appears similar to file system concurrent access?
  - Approach taken by MySQL initially; now MySQL offers better alternatives
- But want to control at much finer granularity
  - Or else one withdrawal would lock up all accounts!

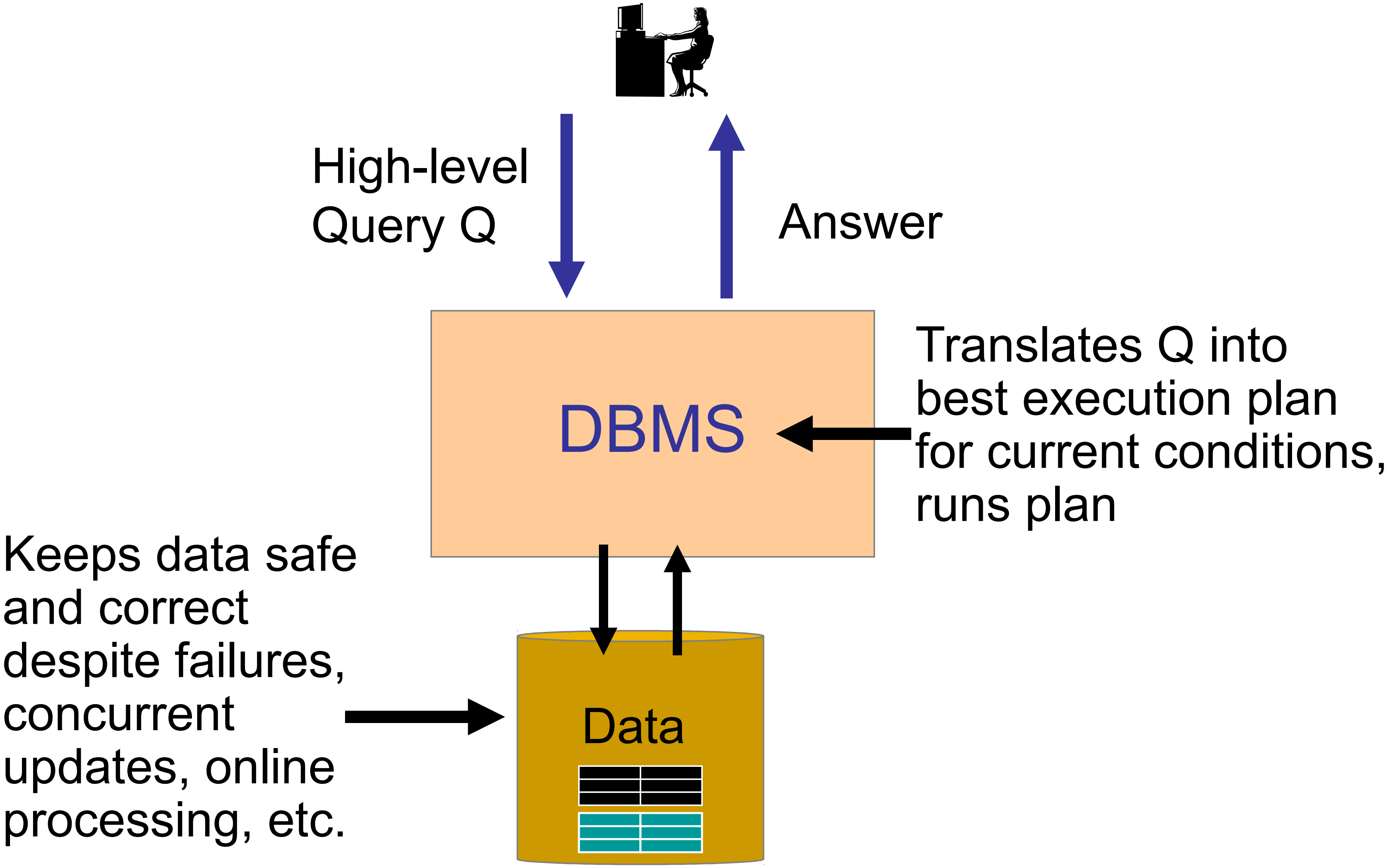
Recovery in DBMS

Example: balance transfer

decrement the balance of account X by $100;

- increment the balance of account Y by $100;
- Scenario 1: Power goes out after the first instruction
- Scenario 2: DBMS buffers and updates data in memory (for efficiency); before they are written back to disk, power goes out
- Log updates; undo/redo during recovery

DataBase Management System (DBMS)



High-level Query Q

Answer

DBMS

Translates Q into best execution plan for current conditions, runs plan

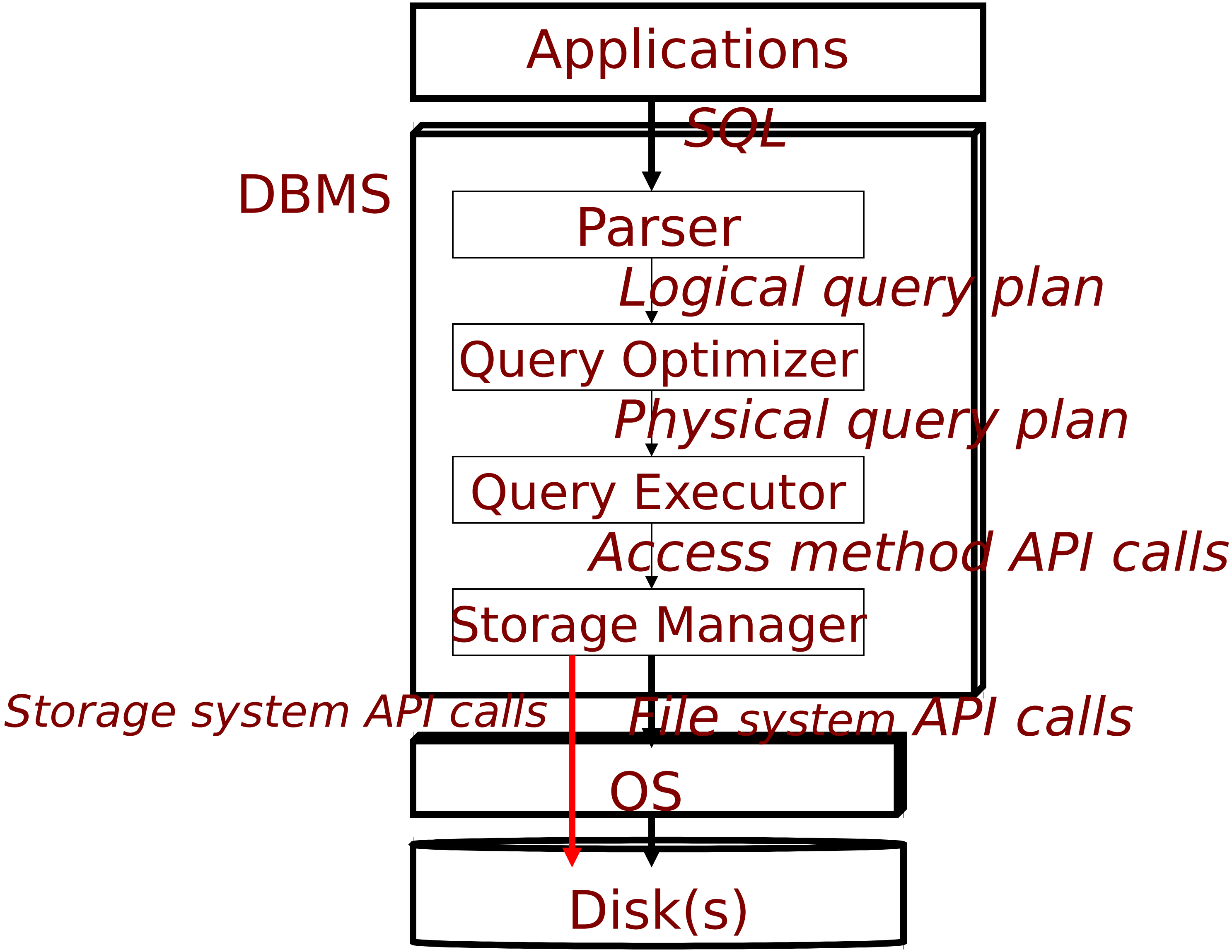
Keeps data safe and correct despite failures, concurrent updates, online processing, etc.

Data

Summary of modern DBMS features

- Persistent storage of data
- Logical data model; declarative queries and updates ! physical data independence
- Multi-user concurrent access
- Safety from system failures
- Performance, performance, performance
  - Massive amounts of data (terabytes ~ petabytes)
  - High throughput (thousands ~ millions transactions per minute)
  - High availability (¸ 99.999% uptime)

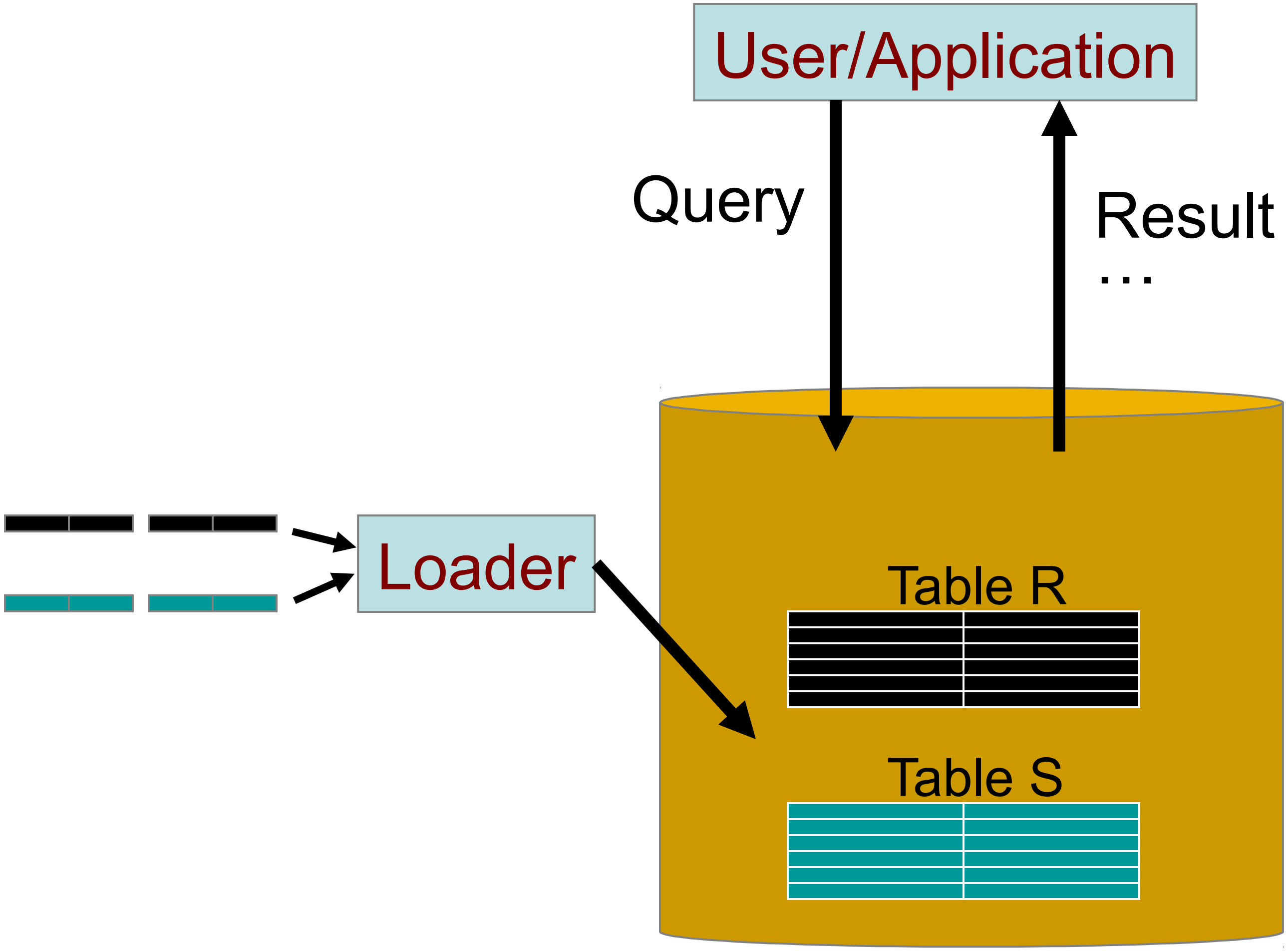
Modern DBMS Architecture



Course Outline

- 50% is about modern DBMS technology
  - Query execution, query optimization, transactions, recovery, etc.
  - Textbook material is starting point, but we will go beyond
- 50% is about one important class of “what is happening today in data management”: Data Integration
  - Structured vs unstructured data
  - Data is not locally stored
  - Data is in many places
  - Data is heterogeneous
  - Data is inconsistent

Using a Traditional DBMS



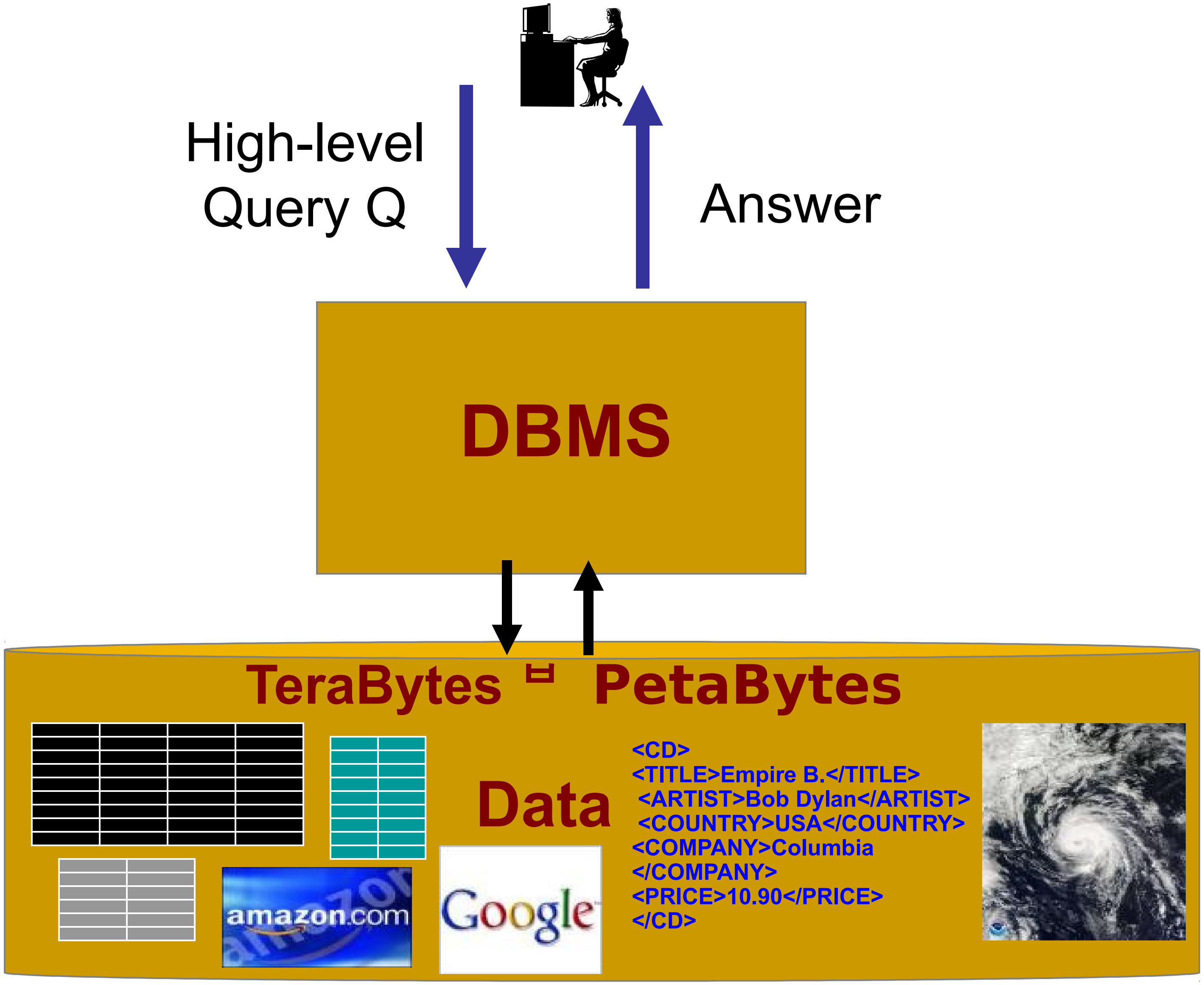
Query Processing

Declarative Query → Query Plan

- NOTE: You will need to be familiar with SQL. We will
- Look at other query languages in class. A SQL refresher is available on the Wiki.

- Focus: Relational System and approach (i.e., data is organized as tables, or relations) is still central for query optimization. We will look at extensions, in particular nested relational algebra, and limited schema which are common in modern scenarios.

New Challenges in DBMSs



```
<CD>
<TITLE>Empire B.</TITLE>
 <ARTIST>Bob Dylan</ARTIST>
 <COUNTRY>USA</COUNTRY>
<COMPANY>Columbia
</COMPANY>
<PRICE>10.90</PRICE>
</CD>
```

Data Integration: Overview

- ➢ Introduction: data integration as a new abstraction
- Examples of data integration applications
- Schema heterogeneity
- Goal of data integration, why it’s a hard problem
- Data integration architectures

Data Integration

- Databases are great: they let us manage huge amounts of data
  - Assuming you’ve put it all into your schema.
- In reality, data sets are often created independently
  - Only to discover later that they need to combine their data!
  - At that point, they’re using different systems, different schemata and have limited interfaces to their data.
- The goal of data integration: tie together different sources, controlled by many people, under a common schema.

DBMS: it's all about abstraction

- *Logical* vs. *Physical*; *What* vs. *How.*

Students:

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

Takes:

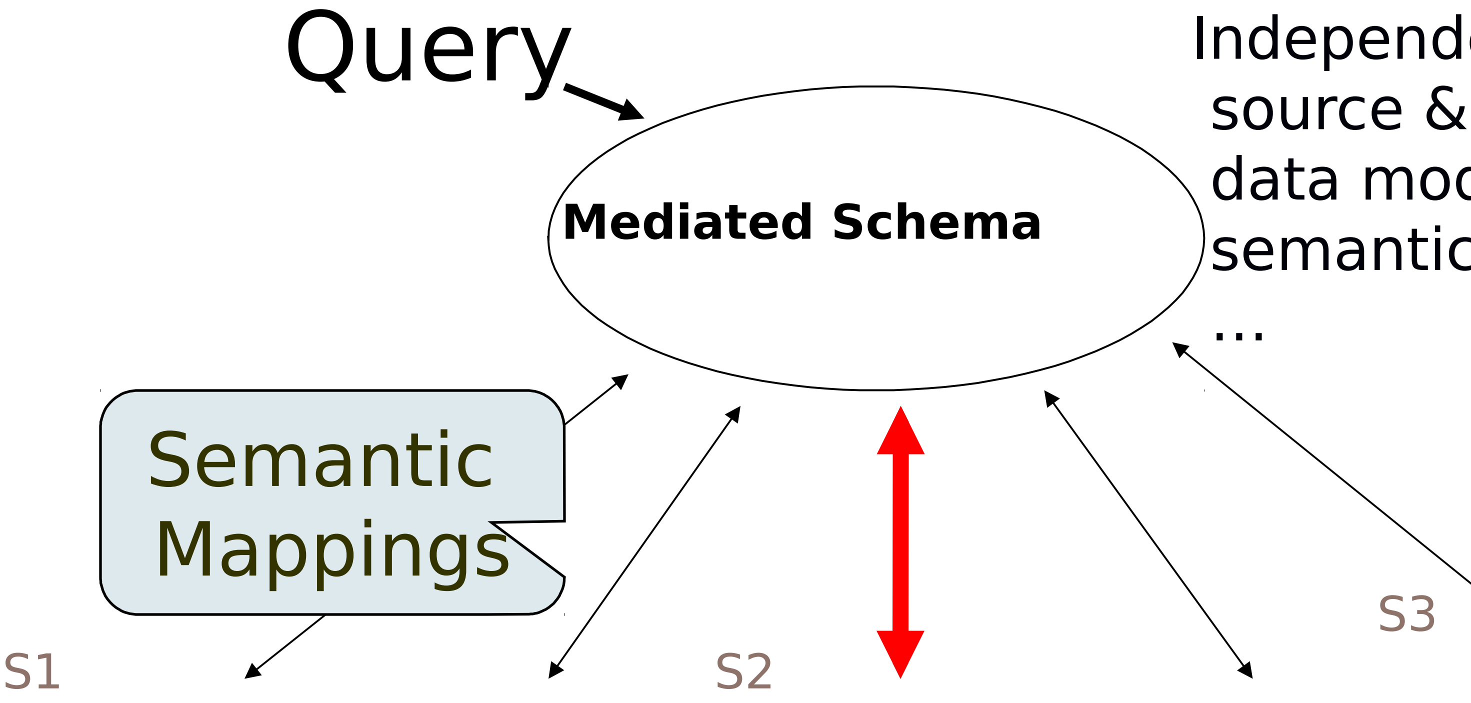
SSN	CID
123-45-6789	CSE444
123-45-6789	CSE444
234-56-7890	CSE142
	…

Courses:

CID	Name	Quarter
CSE444	Databases	fall
CSE541	Operating systems	winter

```
SELECT  C.name
FROM Students S, Takes T, Courses C
WHERE S.name="Mary" and
        S.ssn = T.ssn and T.cid = C.cid
```

Data Integration:
A Higher-level Abstraction



Independence of:
source & location
data model, syntax
semantic variations
…

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

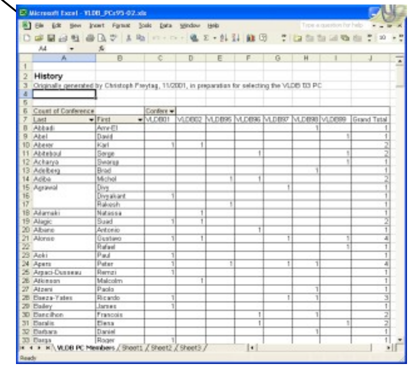
SSN	CID
123-45-6789	CSE444
123-45-6789	CSE444
234-56-7890	CSE142
	…

CID	Name	Quarter
CSE444	Databases	fall
CSE541	Operating systems	winter

. .

- <cd> <title> The best of … </title>
  - <artist> Carreras </artist>
  - <artist> Pavarotti </artist>
  - <artist> Domingo </artist>
  - <price> 19.95 </price>
  - </cd>

. . .


Microsoft Excel - VLDB_PCs95-02.xls
File Edit View Insert Format Tools Data Window Help
Type a question for help
A4

	A	B	C	D	E	F	G	H	I	J
1										
2	History									
3	Originally generated by Christoph Freytag, 11/2001, in preparation for selecting the VLDB 03 PC									
4										
5										
6	Count of Conference		Confere							
7	Last	First	VLDB01	VLDB02	VLDB95	VLDB96	VLDB97	VLDB98	VLDB99	Grand Total
8	Abbadi	Amr El						1		1
9	Abel	David							1	1
10	Aberer	Karl	1	1						2
11	Abiteboul	Serge				1			1	2
12	Acharya	Swarup							1	1
13	Adelberg	Brad						1		1
14	Adiba	Michel			1	1				2
15	Agrawal	Divy					1			1
16		Divyakant	1							1
17		Rakesh			1					1
18	Ailamaki	Natassa		1						1
19	Alagic	Suad	1	1						2
20	Albano	Antonio				1				1
21	Alonso	Gustavo	1	1			1		1	4
22		Rafael							1	1
23	Aoki	Paul	1							1
24	Apers	Peter	1		1		1	1		4
25	Arpaci-Dusseau	Remzi	1							1
26	Atkinson	Malcolm		1						1
27	Atzeni	Paolo						1		1
28	Baeza-Yates	Ricardo	1				1	1		3
29	Bailey	James	1							1
30	Bancilhon	Francois				1		1		2
31	Baralis	Elena				1			1	2
32	Barbara	Daniel						1		1
33	Barga	Roger	1							1

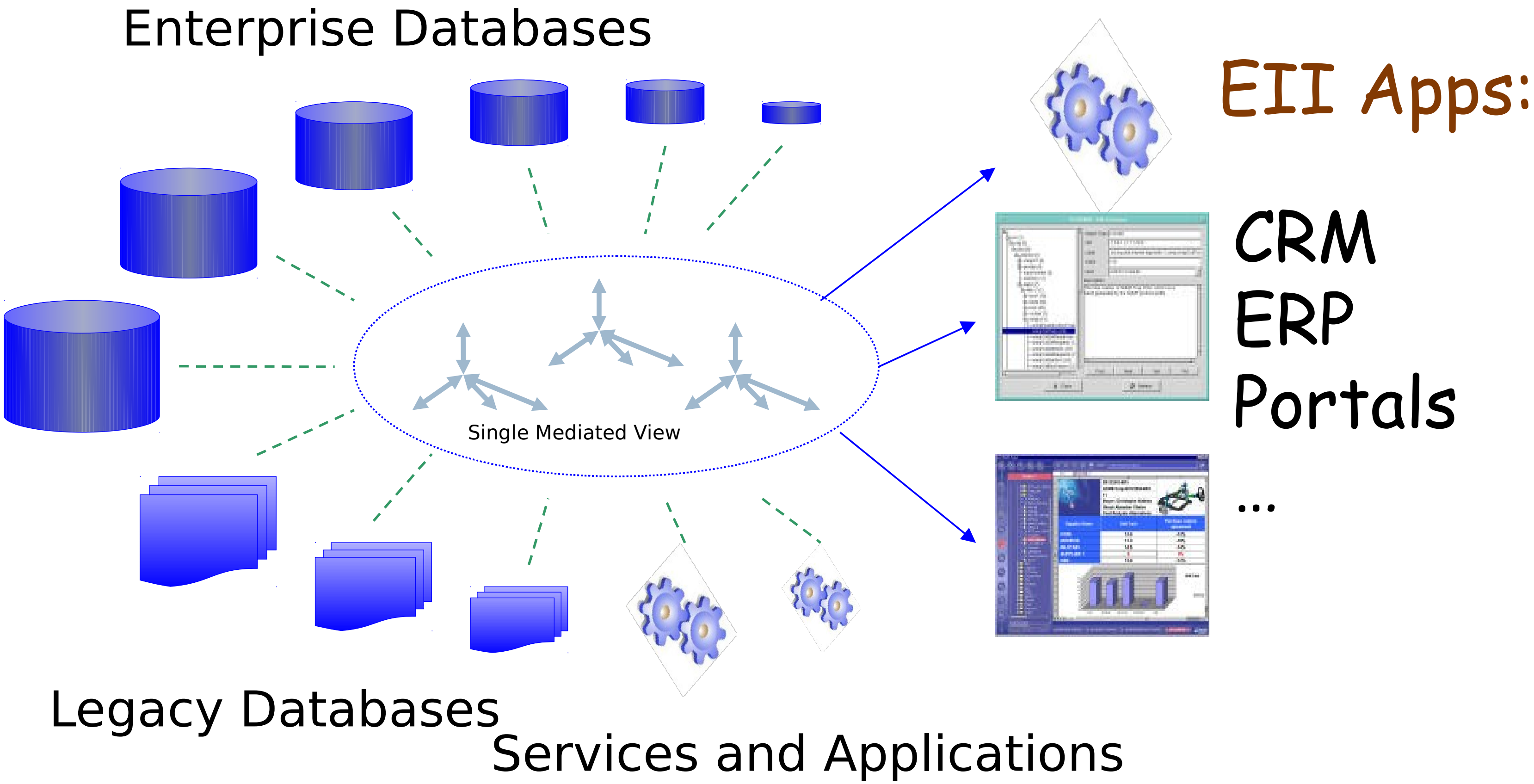
• Outline

- ✓ Introduction: data integration as a new abstraction
- ➢ Examples of data integration applications
- ▪ Schema heterogeneity
- ▪ Goal of data integration, why it’s a hard problem
- ▪ Data integration architectures

Applications of Data Integration

- Business
- Science
- Government
- The Web
- Pretty much everywhere

Application Area 1: Business



Enterprise Databases

Single Mediated View

EII Apps:

CRM

ERP

Portals

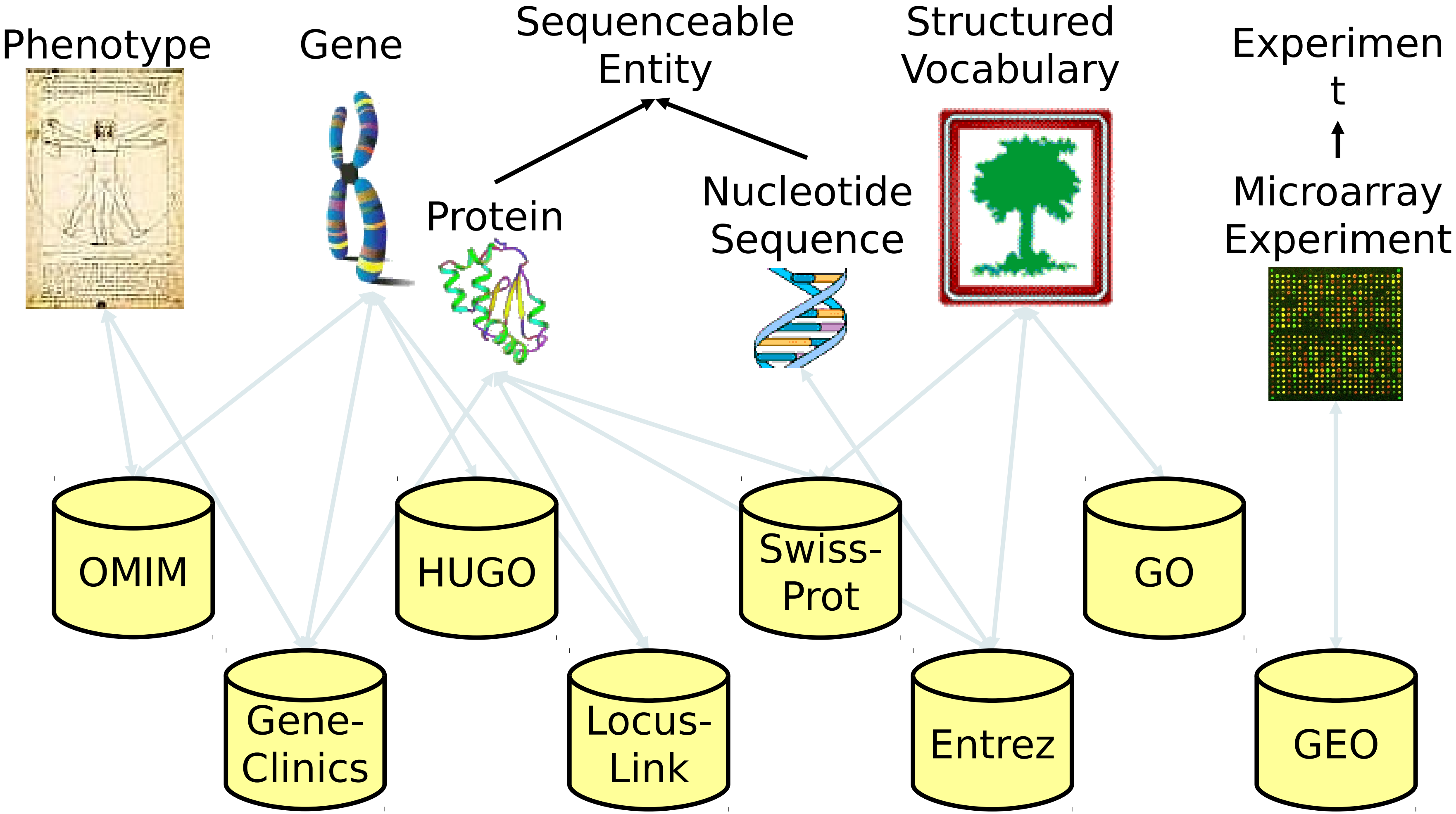
…

Legacy Databases

Services and Applications

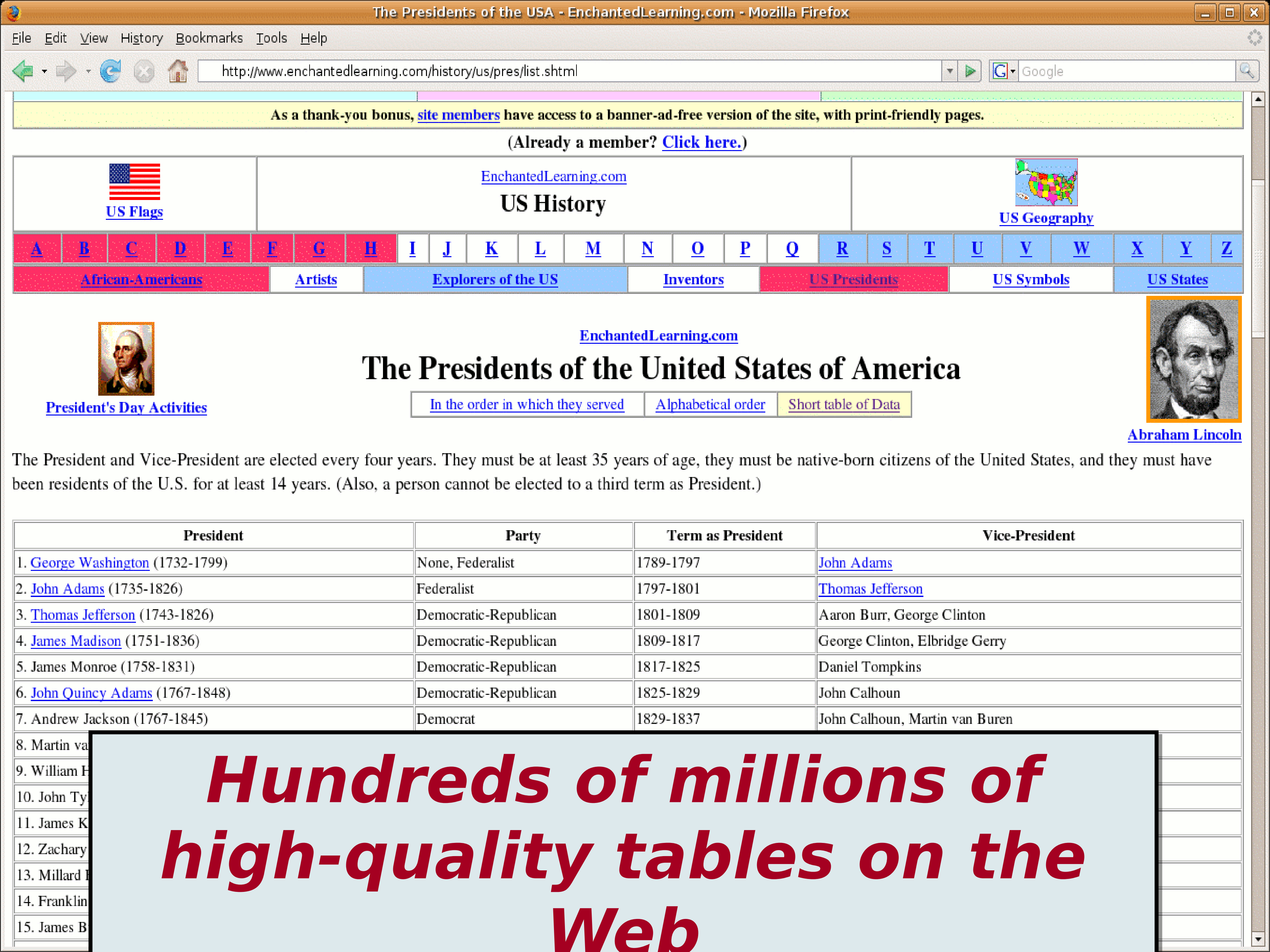
50% of all IT $$$ spent here!

Application Area 2: Science



Hundreds of biomedical data sources available; growing rapidly!

Application Area 3: The Web



The Presidents of the USA - EnchantedLearning.com - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.enchantedlearning.com/history/us/pres/list.shtml
Google

As a thank-you bonus, site members have access to a banner-ad-free version of the site, with print-friendly pages.

(Already a member? Click here.)

US Flags	EnchantedLearning.com US History	US Geography

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

African-Americans | Artists | Explorers of the US | Inventors | US Presidents | US Symbols | US States

President's Day Activities

EnchantedLearning.com

The Presidents of the United States of America

In the order in which they served | Alphabetical order | Short table of Data

Abraham Lincoln

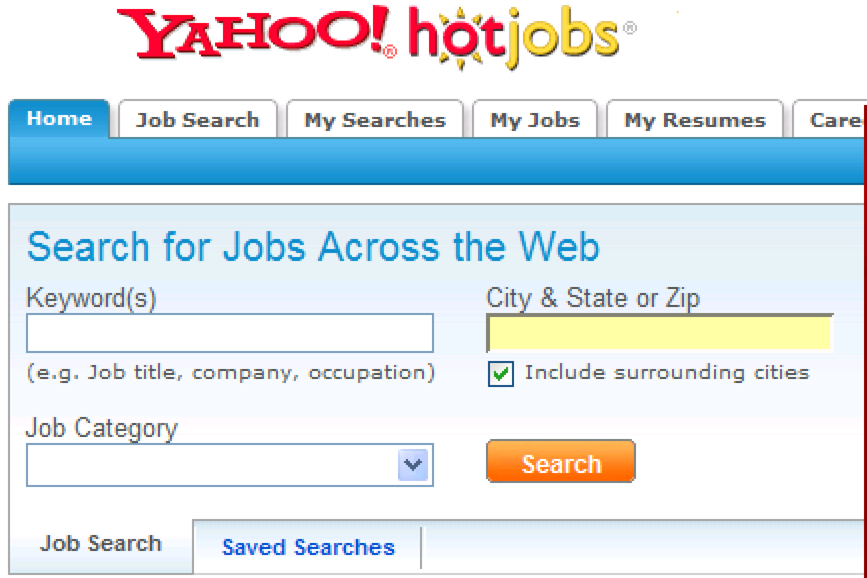
The President and Vice-President are elected every four years. They must be at least 35 years of age, they must be native-born citizens of the United States, and they must have been residents of the U.S. for at least 14 years. (Also, a person cannot be elected to a third term as President.)

President	Party	Term as President	Vice-President
1. George Washington (1732-1799)	None, Federalist	1789-1797	John Adams
2. John Adams (1735-1826)	Federalist	1797-1801	Thomas Jefferson
3. Thomas Jefferson (1743-1826)	Democratic-Republican	1801-1809	Aaron Burr, George Clinton
4. James Madison (1751-1836)	Democratic-Republican	1809-1817	George Clinton, Elbridge Gerry
5. James Monroe (1758-1831)	Democratic-Republican	1817-1825	Daniel Tompkins
6. John Quincy Adams (1767-1848)	Democratic-Republican	1825-1829	John Calhoun
7. Andrew Jackson (1767-1845)	Democrat	1829-1837	John Calhoun, Martin van Buren
8. Martin va			
9. William H			
10. John Ty			
11. James K			
12. Zachary			
13. Millard F			
14. Franklin			
15. James B			

Hundreds of millions of high-quality tables on the Web

The Deep Web

- Millions of high quality HTML forms out there
- Each form has its own special interface
    - Hard to explore data across sites.
- Goal (for some domains):
    - A single interface into a multitude of deep-web sources.



YAHOO! hotjobs

Home | Job Search | My Searches | My Jobs | My Resumes | Care

Search for Jobs Across the Web

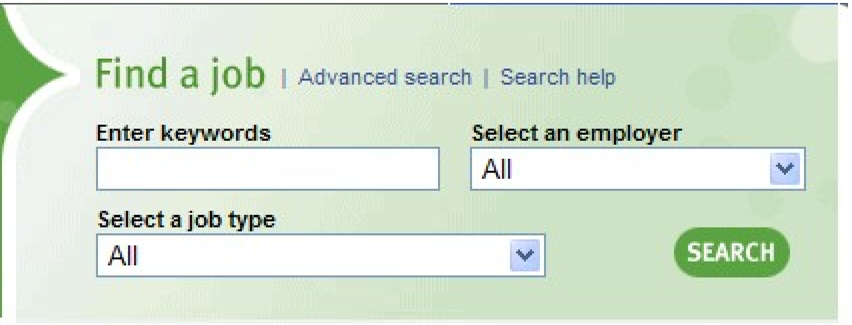
Keyword(s)

(e.g. Job title, company, occupation)

City & State or Zip

☑ Include surrounding cities

Job Category

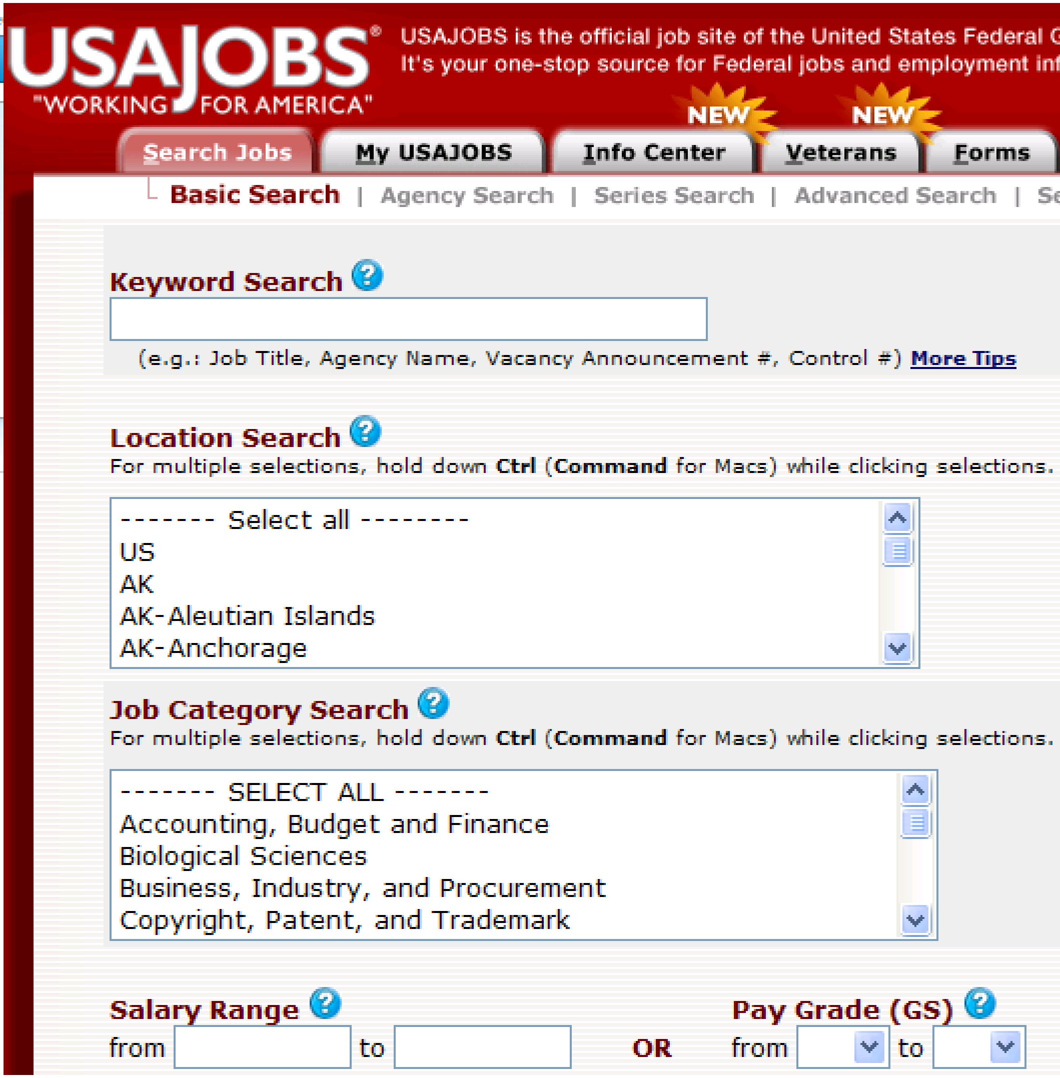


Find a job

Enter keywords

Select an employer: All

Select a job type: All



USAJOBS "WORKING FOR AMERICA"

USAJOBS is the official job site of the United States Federal G
It's your one-stop source for Federal jobs and employment inf

NEW
NEW
Search Jobs
My USAJOBS
Info Center
Veterans
Forms
Basic Search | Agency Search | Series Search | Advanced Search | Se

Keyword Search

(e.g.: Job Title, Agency Name, Vacancy Announcement #, Control #) More Tips

Location Search

For multiple selections, hold down **Ctrl** (**Command** for Macs) while clicking selections.

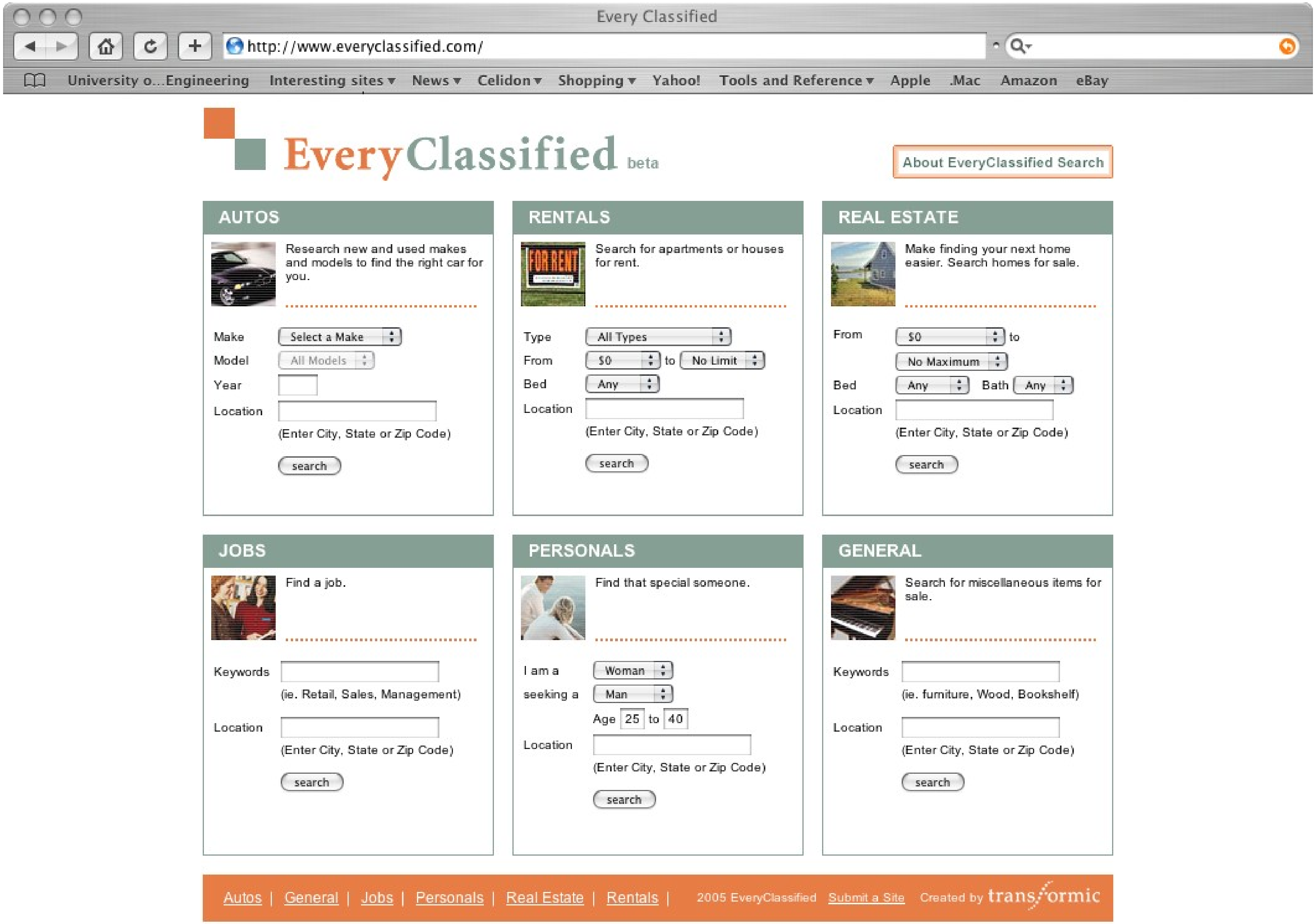
------- Select all --------
US
AK
AK-Aleutian Islands
AK-Anchorage

Job Category Search

For multiple selections, hold down **Ctrl** (**Command** for Macs) while clicking selections.

------- SELECT ALL -------
Accounting, Budget and Finance
Biological Sciences
Business, Industry, and Procurement
Copyright, Patent, and Trademark

Salary Range from to **OR** **Pay Grade (GS)** from to



Every Classified

http://www.everyclassified.com/

University o...Engineering | Interesting sites ▾ | News ▾ | Celidon ▾ | Shopping ▾ | Yahoo! | Tools and Reference ▾ | Apple | .Mac | Amazon | eBay

EveryClassified beta

About EveryClassified Search

AUTOS

Research new and used makes and models to find the right car for you.

Make	Select a Make
Model	All Models
Year	
Location	
	(Enter City, State or Zip Code)

search

RENTALS

Search for apartments or houses for rent.

Type	All Types
From	$0 to No Limit
Bed	Any
Location	
	(Enter City, State or Zip Code)

search

REAL ESTATE

Make finding your next home easier. Search homes for sale.

From	$0 to
	No Maximum
Bed	Any Bath Any
Location	
	(Enter City, State or Zip Code)

search

JOBS

Find a job.

Keywords	
	(ie. Retail, Sales, Management)
Location	
	(Enter City, State or Zip Code)

search

PERSONALS

Find that special someone.

I am a	Woman
seeking a	Man
	Age 25 to 40
Location	
	(Enter City, State or Zip Code)

search

GENERAL

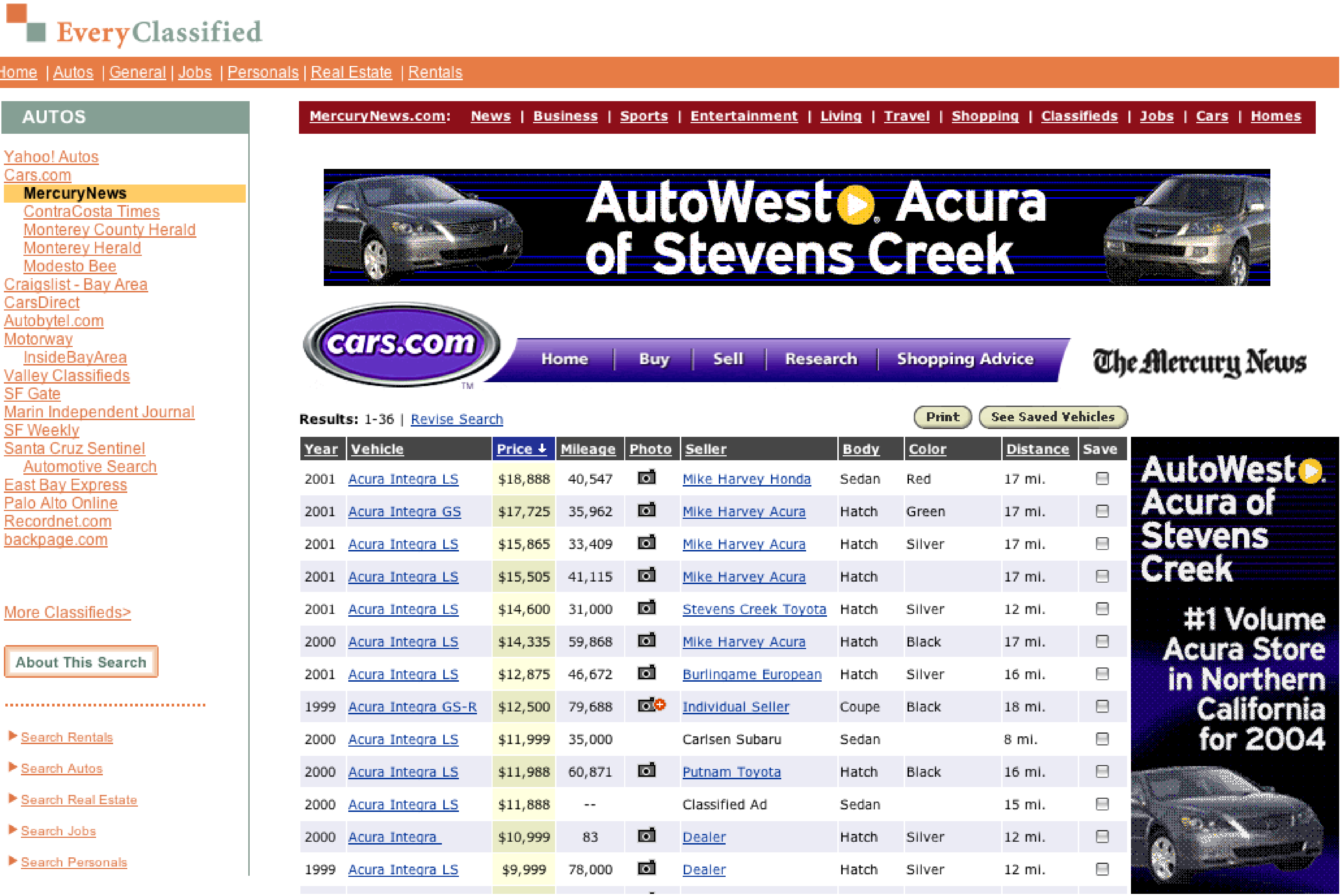
Search for miscellaneous items for sale.

Keywords	
	(ie. furniture, Wood, Bookshelf)
Location	
	(Enter City, State or Zip Code)

search

Autos | General | Jobs | Personals | Real Estate | Rentals | 2005 EveryClassified Submit a Site Created by transformic

reate a single site to search for jobs/rentals/..



Results: 1-36 | Revise Search

Year	Vehicle	Price ↓	Mileage	Photo	Seller	Body	Color	Distance	Save
2001	Acura Integra LS	$18,888	40,547	📷	Mike Harvey Honda	Sedan	Red	17 mi.	□
2001	Acura Integra GS	$17,725	35,962	📷	Mike Harvey Acura	Hatch	Green	17 mi.	□
2001	Acura Integra LS	$15,865	33,409	📷	Mike Harvey Acura	Hatch	Silver	17 mi.	□
2001	Acura Integra LS	$15,505	41,115	📷	Mike Harvey Acura	Hatch		17 mi.	□
2001	Acura Integra LS	$14,600	31,000	📷	Stevens Creek Toyota	Hatch	Silver	12 mi.	□
2000	Acura Integra LS	$14,335	59,868	📷	Mike Harvey Acura	Hatch	Black	17 mi.	□
2001	Acura Integra LS	$12,875	46,672	📷	Burlingame European	Hatch	Silver	16 mi.	□
1999	Acura Integra GS-R	$12,500	79,688	📷	Individual Seller	Coupe	Black	18 mi.	□
2000	Acura Integra LS	$11,999	35,000		Carlsen Subaru	Sedan		8 mi.	□
2000	Acura Integra LS	$11,988	60,871	📷	Putnam Toyota	Hatch	Black	16 mi.	□
2000	Acura Integra LS	$11,888	--		Classified Ad	Sedan		15 mi.	□
2000	Acura Integra	$10,999	83	📷	Dealer	Hatch	Silver	12 mi.	□
1999	Acura Integra LS	$9,999	78,000	📷	Dealer	Hatch	Silver	12 mi.	□

ly traverse between the site by clicking its na

Outline

- ✓ Introduction: data integration as a new abstraction
- ✓ Examples of data integration applications
- ➢ Schema heterogeneity
- ▪ Goal of data integration, why it’s a hard problem
- ▪ Data integration architectures

Enterprise Data Integration:
FullServe Corporation

Employees

- FullTimeEmp
- Hire
- TempEmployees

Training

- Courses
- Enrollments

Sales

- Products
- Sales

Resumes

- Interview
- CV

Services

- Services
- Customers
- Contracts

HelpLine

- Calls

EuroCard Corporation

Employees

Employees
Hire

Resumes

Interview

Credit Cards

Customer
CustDetail

HelpLine

Calls

Examples of Heterogeneity

FullServe

FullTimeEmp

ssn, empId, firstName middleName, lastName

Hire

empId, hireDate, recruiter

TempEmployees

ssn, hireStart, hireEnd

EuroCard

Employees

ID, firstNameMiddleInitial, lastName

Hire

ID, hireDate, recruiter

Find all employees (making over $100K)

Customer Call Center

Agents should have a full view of customer when they call in.

Sales

Products
Sales

Services

Services
Customers
Contracts

Credit Cards

Customer
CustDetail

Other Reasons to Integrate Data

- Create a (useful) web site for tracking services
- Collaborate with third parties
  - E.g., create branded services
- Comply with government regulations
  - Find “risky” employees
- Business intelligence
  - What’s really wrong with our products?

Outline

- ✓ Introduction: data integration as a new abstraction
- ✓ Examples of data integration applications
- ✓ Schema heterogeneity
- ➢ Goal of data integration, why it’s a hard problem
- ▪ Data integration architectures

Goal of Data Integration

- Uniform query access to a set of data sources
- Handle:
  - Scale of sources: from tens to millions
  - Heterogeneity
  - Autonomy
  - Semi-structure

Why is it Hard?

- Systems-level reasons:
  - Managing different platforms
  - SQL across multiple systems is not so simple
  - Distributed query processing
- Logical reasons:
  - Schema (and data) heterogeneity
- ‘Social’ reasons:
  - Locating and capturing relevant data in the enterprise.
  - Convincing people to share (data fiefdoms)
    - Security, privacy and performance implications.

Setting Expectations

Data integration is **AI-Complete**.

- Completely automated solutions unlikely.

Goal 1:

- Reduce the effort needed to set up an integration application.

Goal 2:

- Enable the system to perform gracefully with uncertainty (e.g., on the web)

Data Integration Smorgasbord

Something for everyone:

- **Theory** of modeling data sources
- **Systems** aspects of data integration
- **Architectural** issues: e.g., P2P data sharing
- **AI** @ work: automated schema matching
- **Web**: latest on data integration & web
- **Commercial** products: BEA, IBM
- **Semantic Web**: what does it have to offer?
- New trends in DBMS: **uncertainty, dataspaces**

Outline

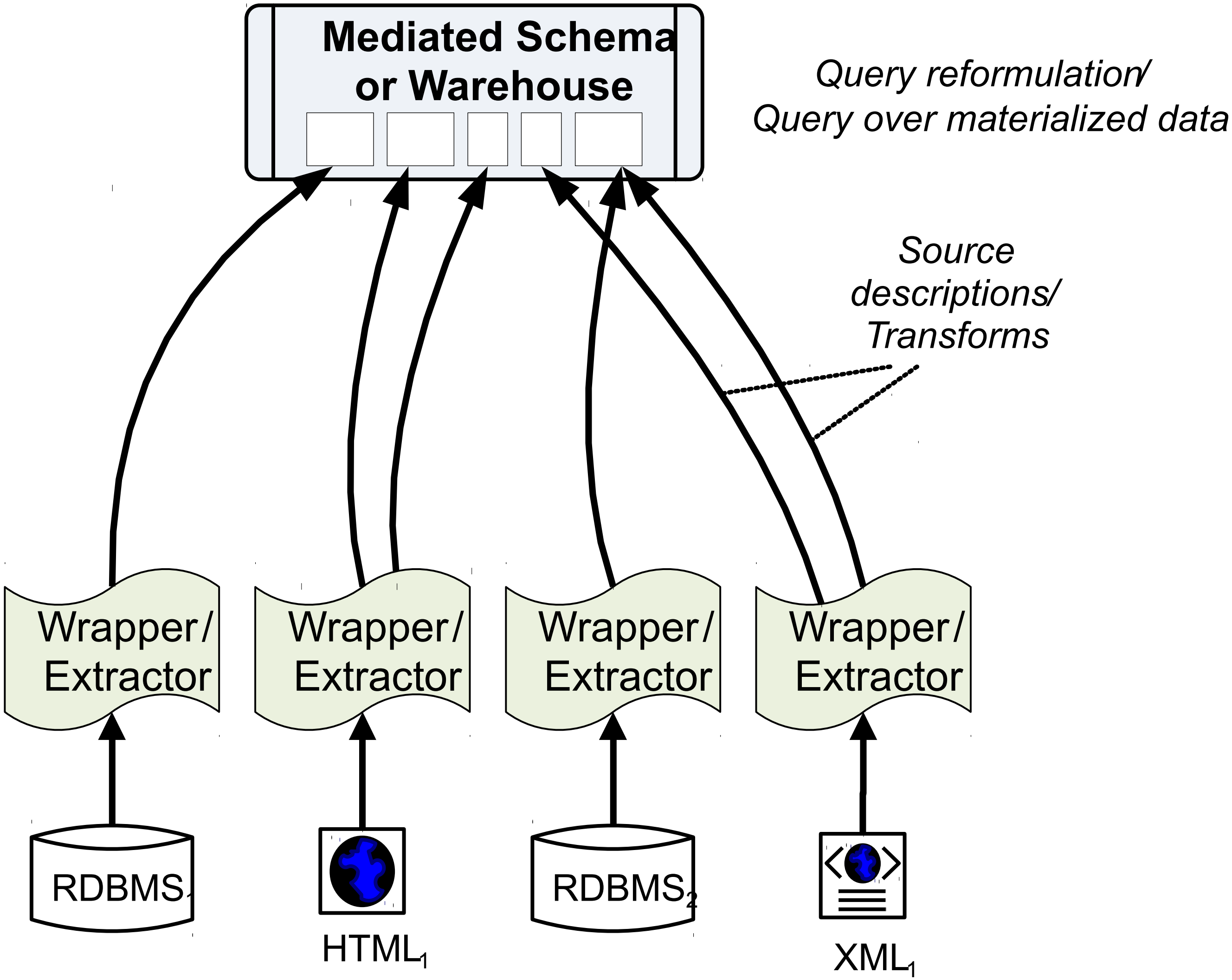
- ✓ Introduction: data integration as a new abstraction
- ✓ Examples of data integration applications
- ✓ Schema heterogeneity
- ✓ Goal of data integration, why it’s a hard problem
- ➢ Data integration architectures

Virtual, Warehousing and in Between

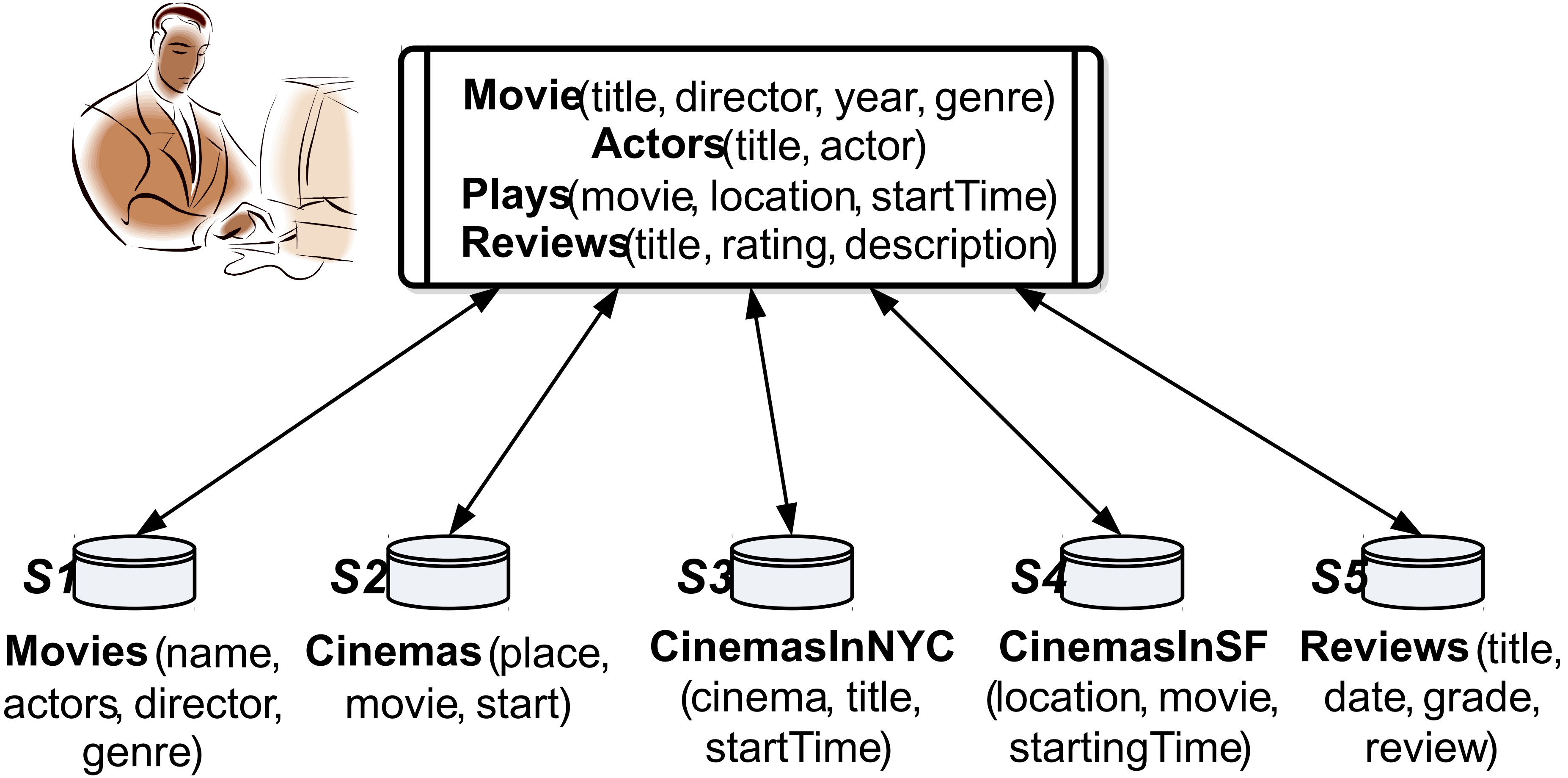
- Data warehousing: integrate by bringing the data into a single physical warehouse
- Virtual data integration: leave the data at the sources and access it at query time.

- Some differences, but semantic heterogeneity arises in both cases.
- Numerous intermediate architectures.
- The course illustrates data integration technology mostly through the virtual architecture.

Virtual Data Integration Architecture



Example



Movie(title, director, year, genre)
Actors(title, actor)
Plays(movie, location, startTime)
Reviews(title, rating, description)

S1 **Movies** (name, actors, director, genre)

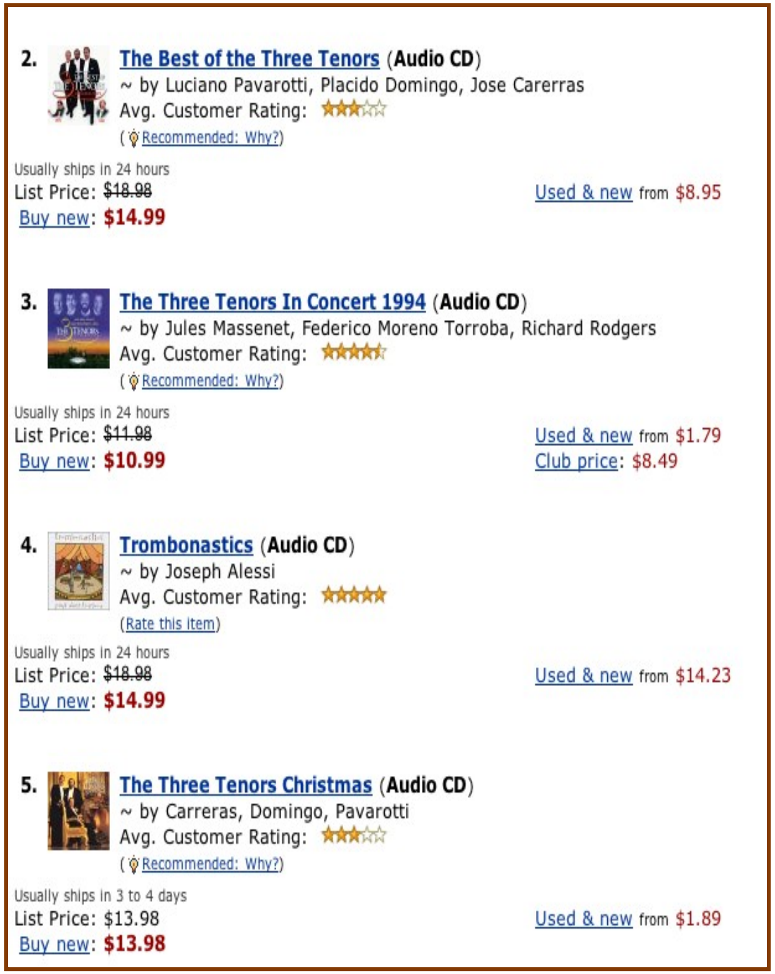
S2 **Cinemas** (place, movie, start)

S3 **CinemasInNYC** (cinema, title, startTime)

S4 **CinemasInSF** (location, movie, startingTime)

S5 **Reviews** (title, date, grade, review)

Wrappers



2. The Best of the Three Tenors (Audio CD)
~ by Luciano Pavarotti, Placido Domingo, Jose Carerras
Avg. Customer Rating: ★★★☆☆
(Recommended: Why?)

Usually ships in 24 hours
List Price: ~~$18.98~~
Buy new: $14.99
Used & new from $8.95

3. The Three Tenors In Concert 1994 (Audio CD)
~ by Jules Massenet, Federico Moreno Torroba, Richard Rodgers
Avg. Customer Rating: ★★★★☆
(Recommended: Why?)

Usually ships in 24 hours
List Price: ~~$11.98~~
Buy new: $10.99
Used & new from $1.79
Club price: $8.49

4. Trombonastics (Audio CD)
~ by Joseph Alessi
Avg. Customer Rating: ★★★★★
(Rate this item)

Usually ships in 24 hours
List Price: ~~$18.98~~
Buy new: $14.99
Used & new from $14.23

5. The Three Tenors Christmas (Audio CD)
~ by Carreras, Domingo, Pavarotti
Avg. Customer Rating: ★★★☆☆
(Recommended: Why?)

Usually ships in 3 to 4 days
List Price: $13.98
Buy new: $13.98
Used & new from $1.89

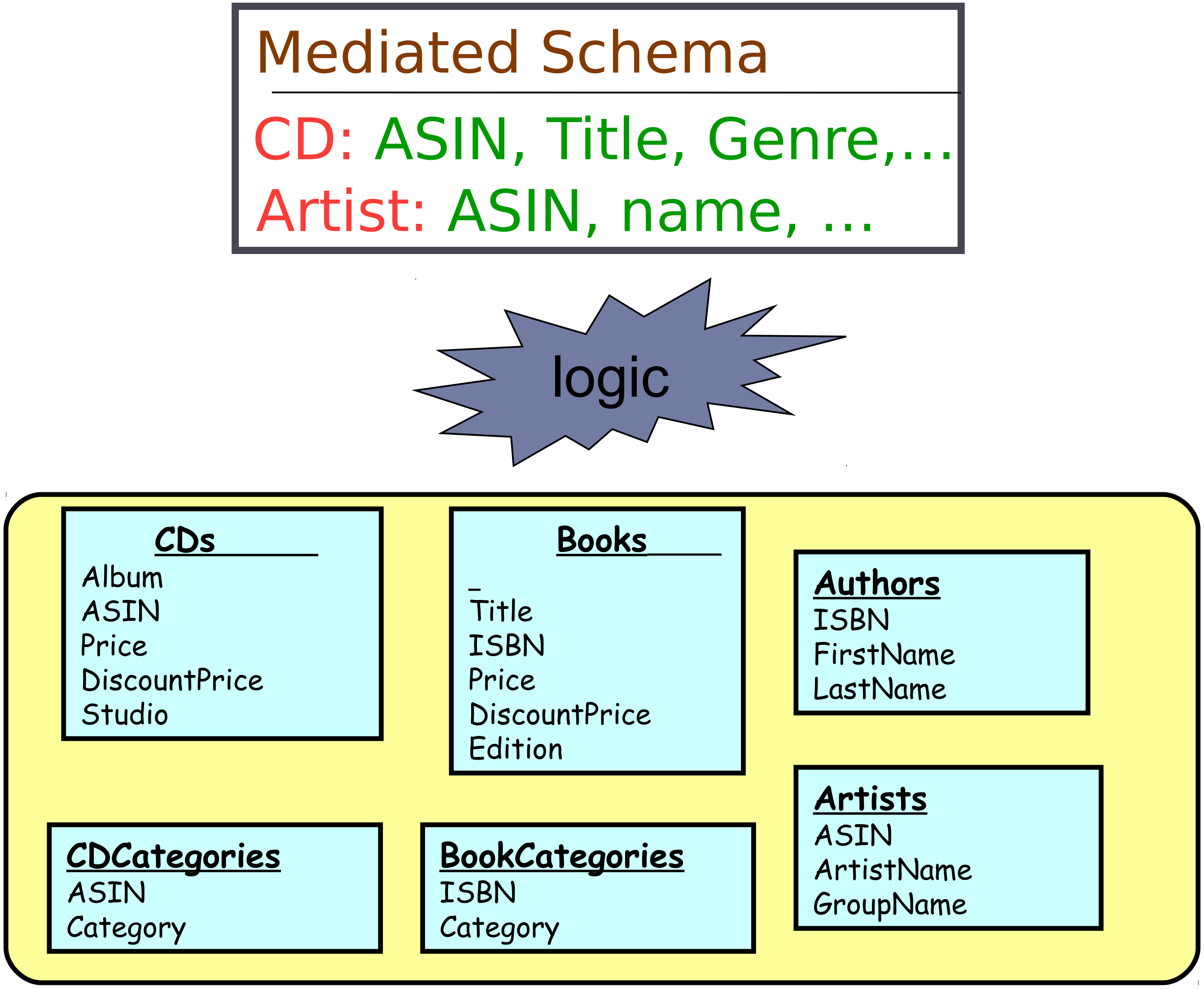
→

```
<cd>    <title> The best of … </title>
              <artist> Abiteboul </artist>
              <artist> Pavarotti  </artist>
              <artist> Domingo  </artist>
              <price> 19.95      </price>
     </cd>
       …
```

Send queries to data sources and transform answers into tuples (or other internal data model). (Chapter 9)

Mediation Languages

Describe relationships between mediated schema and data sources (Chapter 3).



Mediated Schema

CD: ASIN, Title, Genre,...

Artist: ASIN, name, ...

logic

CDs
Album
ASIN
Price
DiscountPrice
Studio

Books
Title
ISBN
Price
DiscountPrice
Edition

Authors
ISBN
FirstName
LastName

CDCategories
ASIN
Category

BookCategories
ISBN
Category

Artists
ASIN
ArtistName
GroupName

Woody Allen Comedies in NY

Mediated schema:

Movie: Title, director, year, genre
Actors: title, actor
Plays: movie, location, startTime
Reviews: title, rating, description

```
select title, startTime
from Movie, Plays
where Movie.title=Plays.movie AND
      location=“New York”  AND
      director=“Woody Allen”
```

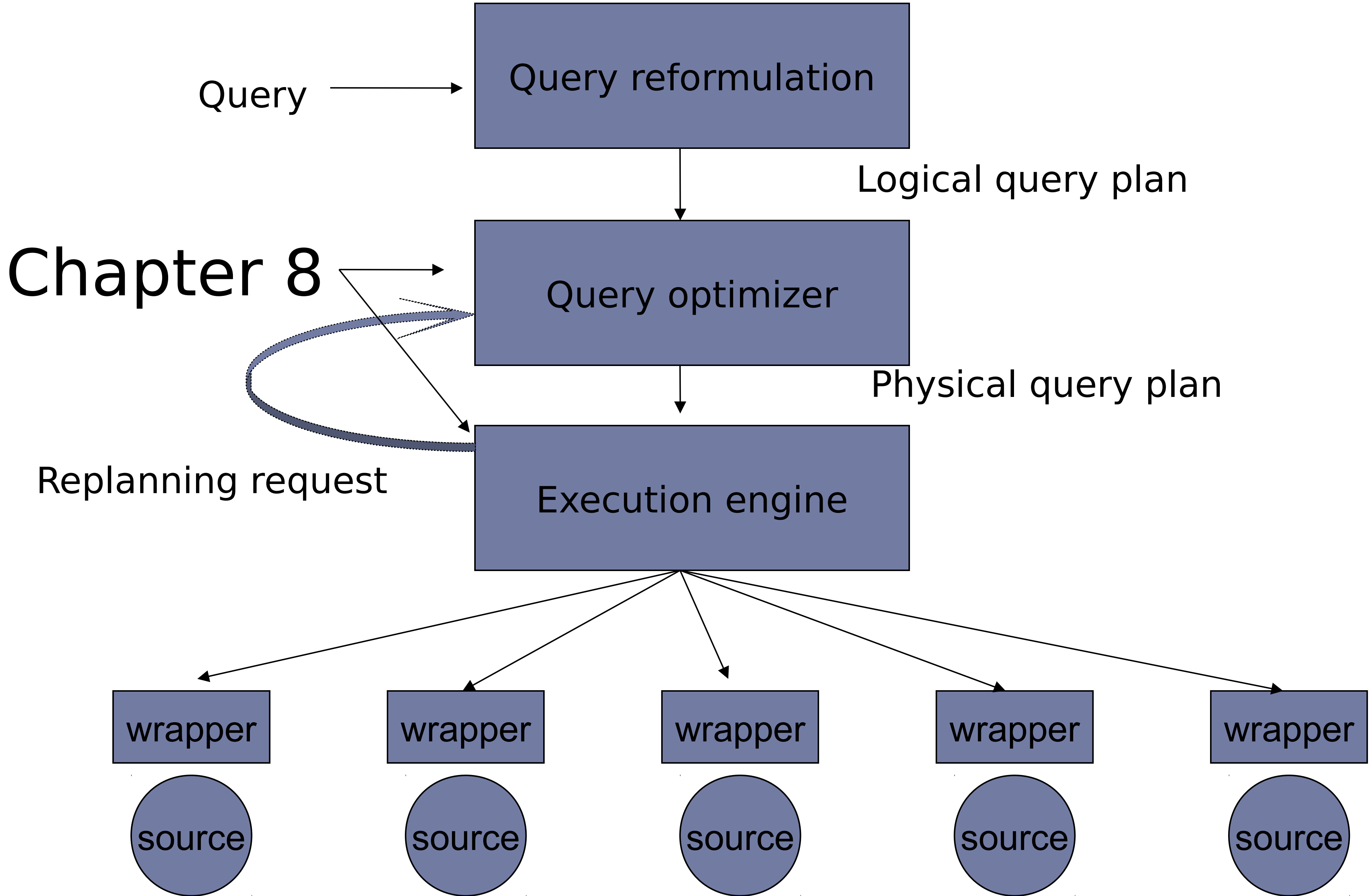
Movie: Title, director, year, genre
Actors: title, actor
Plays: movie, location, startTime
Reviews: title, rating, description

```
select title, startTime
from Movie, Plays
where Movie.title=Plays.movie AND
        location="New York"  AND
        director="Woody Allen"
```

Sources S1 and S3 are relevant, sources S4 and S5 are irrelevant, and source S2 is relevant but possibly redundant.

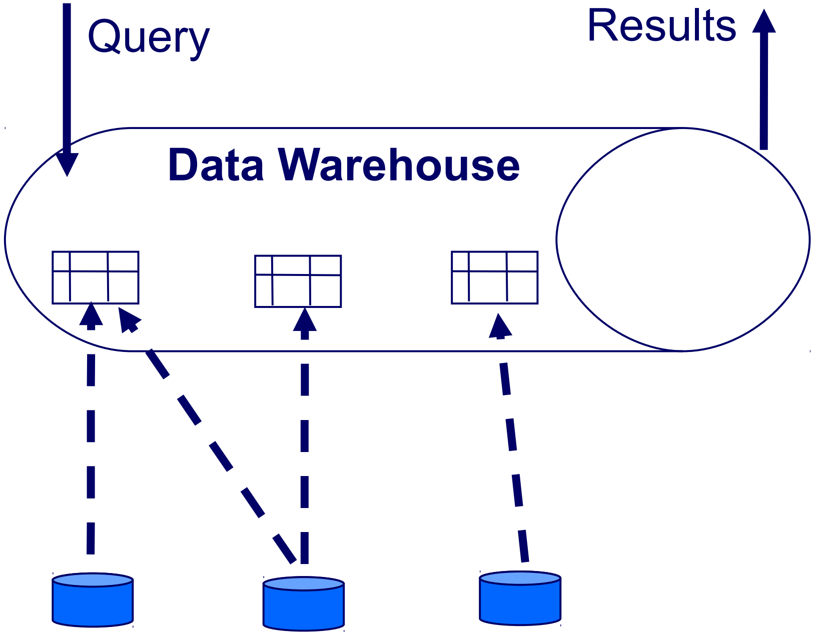
S1	S2	S3	S4	S5
Movies: name, actors, director, genre	Cinemas: place, movie, start	Cinemas in NYC: cinema, title, startTime	Cinemas in SF: location, movie, startingTime	Reviews: title, date grade, review

Query Processing



Data Warehouses – Offline Replication

- Determine physical schema
- Define a database with this schema
- Define procedural *mappings* in an “ETL tool” to import the data and clean it.
- Periodically copy all of the data from the data sources
  - Note that the sources and the warehouse are basically independent at this point



Pros and Cons of Data Warehouses

- ✘ Need to spend time to design the physical database layout, as well as logical
  - ✘ This actually takes a lot of effort!
- ✘ Data is generally not up-to-date (lazy or offline refresh)

- ✓ Queries over the warehouse don’t disrupt the data sources
- ✓ Can run very heavy-duty computations, including data mining and cleaning

Summary: Data Integration

- Data integration: abstract away the fact that data comes from multiple sources in varying schemata.
- Problem occurs everywhere: it’s key to business, science, Web and government.
- Goal: reduce the effort involved in integrating.
- Regardless of the architecture, heterogeneity is a key issue.
- Architectures range from warehousing to virtual integration.

Summary: Data Management Matters

- Assumed and used in most sciences (and sociology, linguistics, etc etc)
- Core to many businesses (finance, healthcare, policy, advertising, city management, etc)
- Cool mix of theory and systems
- Something useful and interesting for everyone (even if data management isn't your main focus).

Next Class

Database Query Compilation & Optimization Principles, Indices

Course Staff and Information

- Instructor:
  - Jerome Simeon, IBM Research, T.J. Watson Lab
- Reach me at js1491@nyu.edu
- In our wiki you will find:

Tentative schedule

News and announcements

Reading list

Assignments

http://www.vistrails.org/index.php/Course:_Advanced_Databases

Check it often!!!

What we will cover

- Query compilation (beyond DB 101): Architecture, indices, query rewritings, new data models, distribution
- Data Integration: Architecture**s**, Schema and data mappings, addressing heterogeneity, wrappers/mediators, query decomposition
- Tentative schedule in:

http://www.vistrails.org/index.php/Course:_Advanced_Databases

Pre-Requisites

- A course in database systems, covering application programming in SQL (and other database-related languages such as OQL or XQuery)
- A course on algorithms and data structures
- Good programming skills
  - *Useful but not required: course on compilers (e.g., for programming languages)*

Readings

- Scientific papers, specific per class
- Textbooks for further study:

"Database Management Systems", by Ramakrishnan and Gehrke, McGraw-Hill, 2002. This book is for background, but we will go beyond its content.

"Query Compilers" by Guido Moerkoette at:
http://pi3.informatik.uni-mannheim.de/~moer/querycompiler.pdf
A bible of query compilation techniques, notably covering techniques that go beyond relational stores.

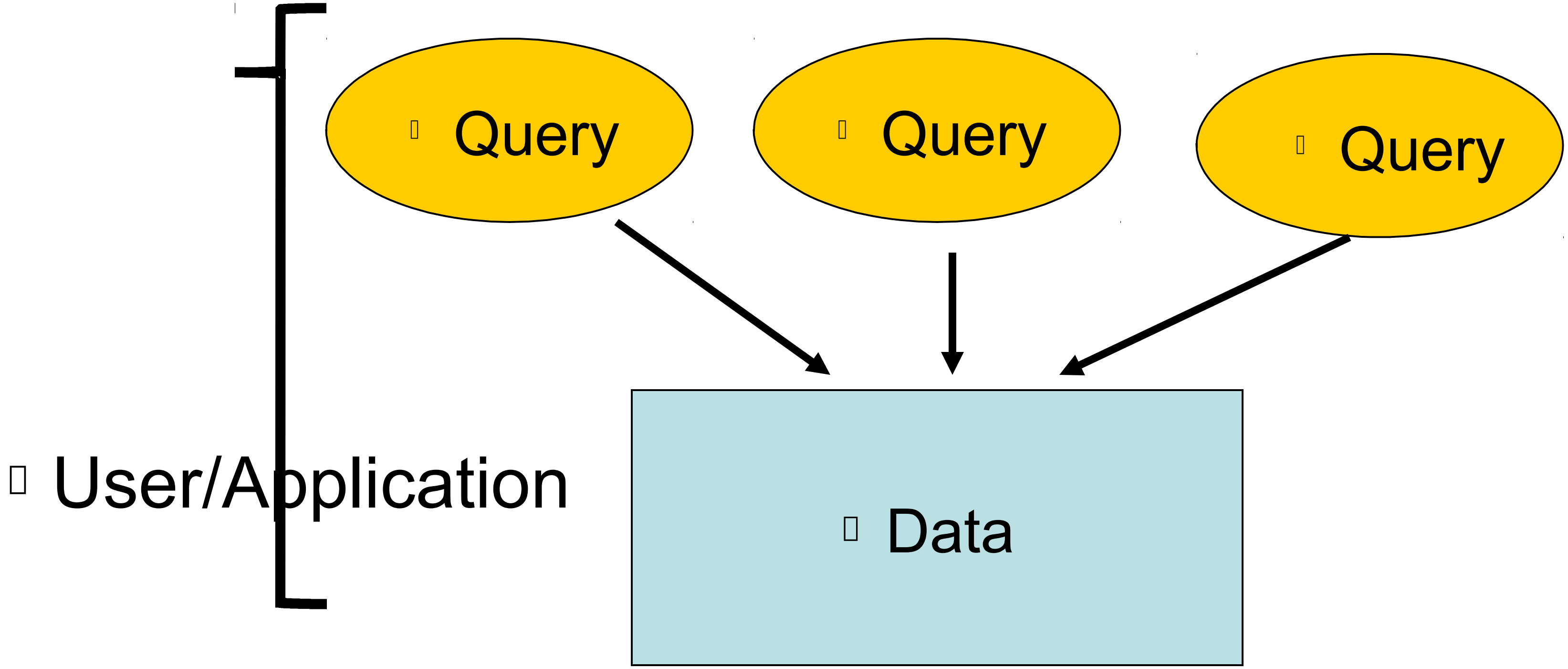
"Principles of Data Integration" by Anhai Doan, Alon Halevy, Zachary Ives. Morgan Kaufman, 2012.
A recent book on data integration that covers most technical aspects important to this area.

What you will do

- Reading assignments and review (33.3%) *done in pairs*
  - *You will need TIME and WORK*
- Quizzes (33.3%): you will use Gradiance
  - Register at http://www.newgradiance.com/services
  - Use token ***(tbd)***
- Final exam (33.3%)

Overview:
Query Compilation –
Data Integration

Data Management



- DataBase Management System (DBMS)

Example: At a Company

Query 1: Is there an employee named “Nemo”?

Query 2: What is “Nemo’s” salary?

Query 3: How many departments are there in the company?

Query 4: What is the name of “Nemo’s” department?

Query 5: How many employees are there in the
□ “Accounts” department?

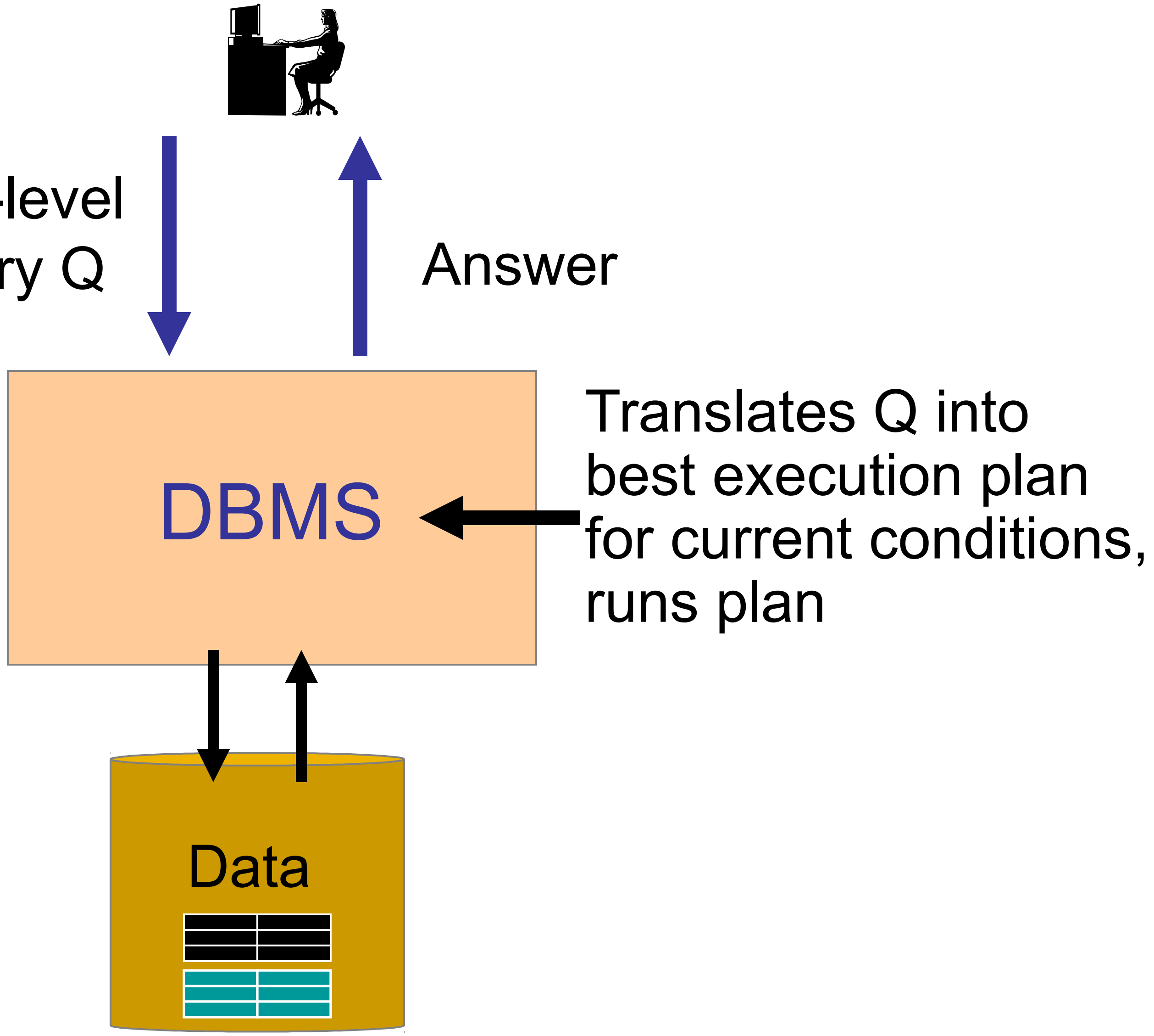
Employee

ID	Name	DeptID	Salary	…
10	Nemo	12	**120K**	…
20	Dory	156	**79K**	…
40	Gill	89	**76K**	…
52	Ray	34	**85K**	…
…	…	…	…	…

Department

ID	Name	…
12	IT	…
34	Accounts	…
89	HR	…
156	Marketing	…
…	…	…

DataBase Management System (DBMS)



High-level Query Q

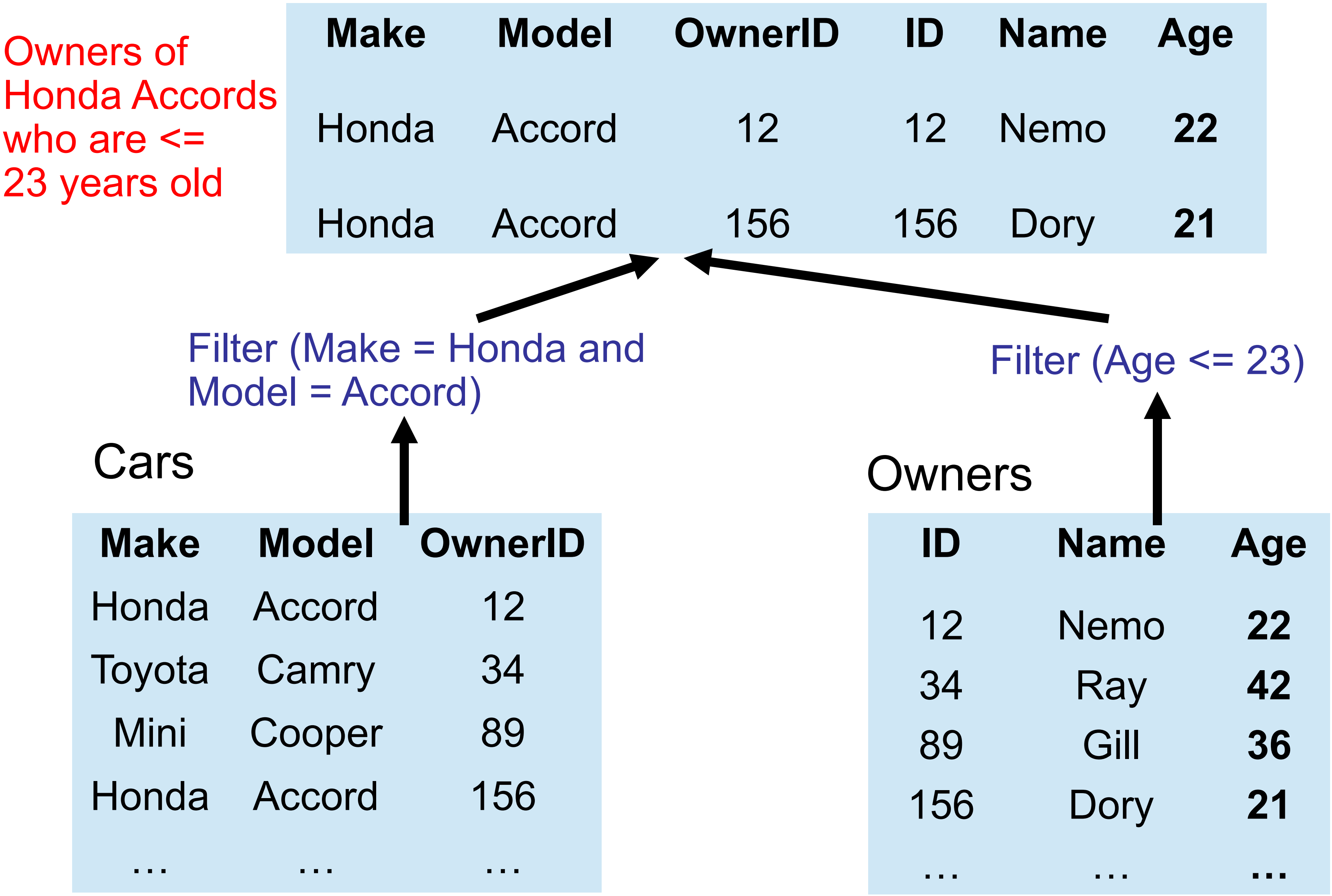
Answer

DBMS

Translates Q into best execution plan for current conditions, runs plan

Data

Example: Store that Sells Cars



Owners of Honda Accords who are <= 23 years old

Make	Model	OwnerID	ID	Name	Age
Honda	Accord	12	12	Nemo	**22**
Honda	Accord	156	156	Dory	**21**

Filter (Make = Honda and Model = Accord)

Filter (Age <= 23)

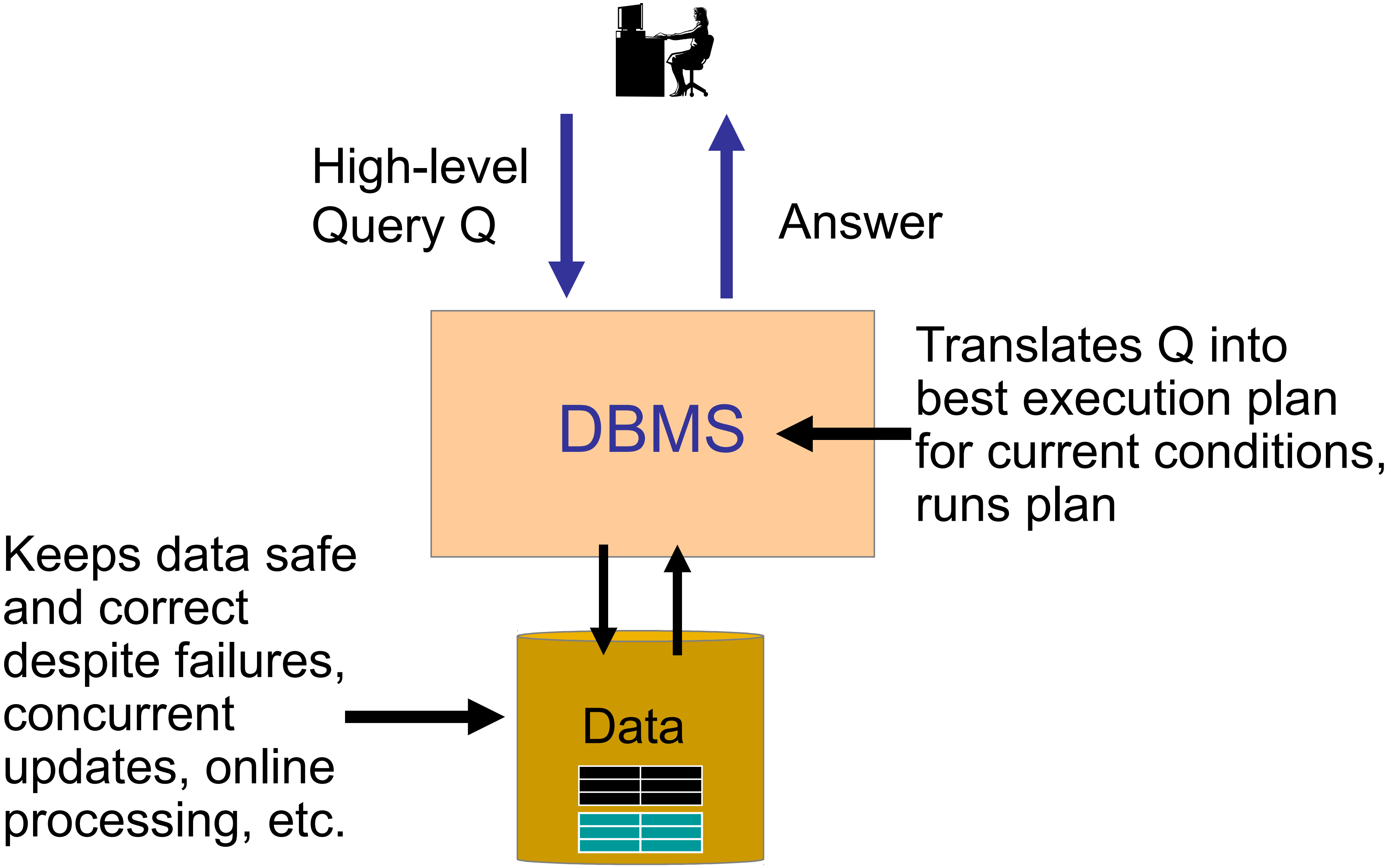
Cars

Make	Model	OwnerID
Honda	Accord	12
Toyota	Camry	34
Mini	Cooper	89
Honda	Accord	156
…	…	…

Owners

ID	Name	Age
12	Nemo	**22**
34	Ray	**42**
89	Gill	**36**
156	Dory	**21**
…	…	**…**

DataBase Management System (DBMS)



Translates Q into best execution plan for current conditions, runs plan

Keeps data safe and correct despite failures, concurrent updates, online processing, etc.

DBMS is multi-user

Example

```
Get account balance from database;
If balance > amount of withdrawal then
    balance = balance - amount of withdrawal;
    dispense cash;
    store new balance into database;
```

- Homer at ATM1 withdraws $100
- Marge at ATM2 withdraws $50
- Initial balance = $400, final balance = ?
    - Should be $250 no matter who goes first

Final balance = $250

Homer withdraws $100:

```
read balance; $400
if balance > amount then
    balance = balance - amount; $300
    write balance; $300
```

Marge withdraws $50:

```
read balance; $300
if balance > amount then
    balance = balance - amount; $250
    write balance; $250
```

Final balance = $300

```
Homer withdraws $100:                Marge withdraws $50:

  read balance; $400
                                     read balance; $400
                                     If balance > amount then
                                      balance = balance - amount; $350
                                      write balance; $350


  if balance > amount then
    balance = balance - amount; $300
    write balance; $300
```

Final balance = $350

```
Homer withdraws $100:                 Marge withdraws $50:

  read balance; $400
                                      read balance; $400

  if balance > amount then
   balance = balance - amount; $300
   write balance; $300

                                      if balance > amount then
                                       balance = balance - amount; $350
                                       write balance; $350
```

Concurrency control in DBMS

- Similar to concurrent programming problems
  - But data is not all in main-memory
- Appears similar to file system concurrent access?
  - Approach taken by MySQL initially; now MySQL offers better alternatives
- But want to control at much finer granularity
  - Or else one withdrawal would lock up all accounts!

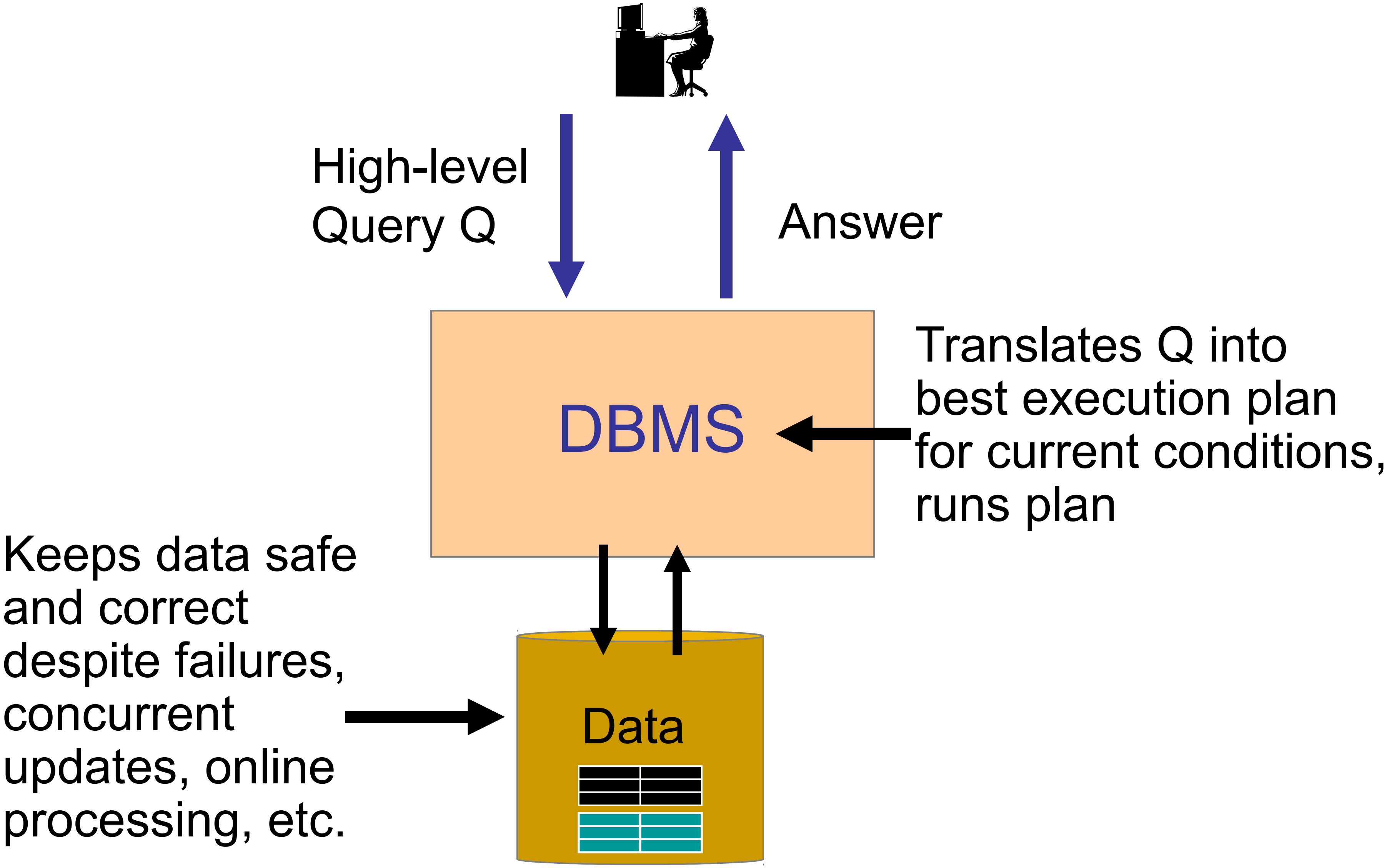
Recovery in DBMS

Example: balance transfer

decrement the balance of account X by $100;

- increment the balance of account Y by $100;
- Scenario 1: Power goes out after the first instruction
- Scenario 2: DBMS buffers and updates data in memory (for efficiency); before they are written back to disk, power goes out
- Log updates; undo/redo during recovery

DataBase Management System (DBMS)



High-level
Query Q

Answer

DBMS

Translates Q into best execution plan for current conditions, runs plan

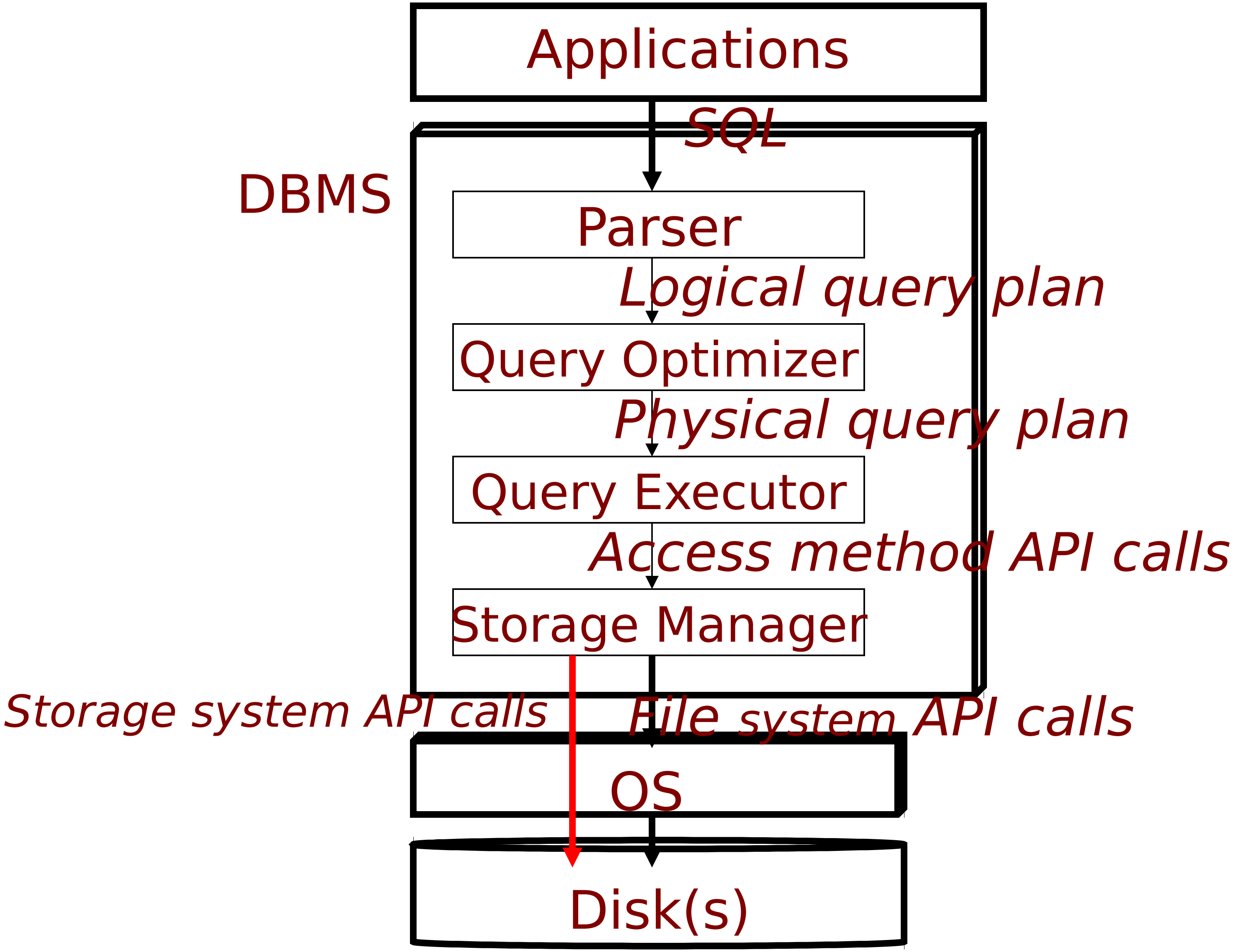
Keeps data safe and correct despite failures, concurrent updates, online processing, etc.

Data

Summary of modern DBMS features

- Persistent storage of data
- Logical data model; declarative queries and updates ! physical data independence
- Multi-user concurrent access
- Safety from system failures
- Performance, performance, performance
  - Massive amounts of data (terabytes ~ petabytes)
  - High throughput (thousands ~ millions transactions per minute)
  - High availability (¸ 99.999% uptime)

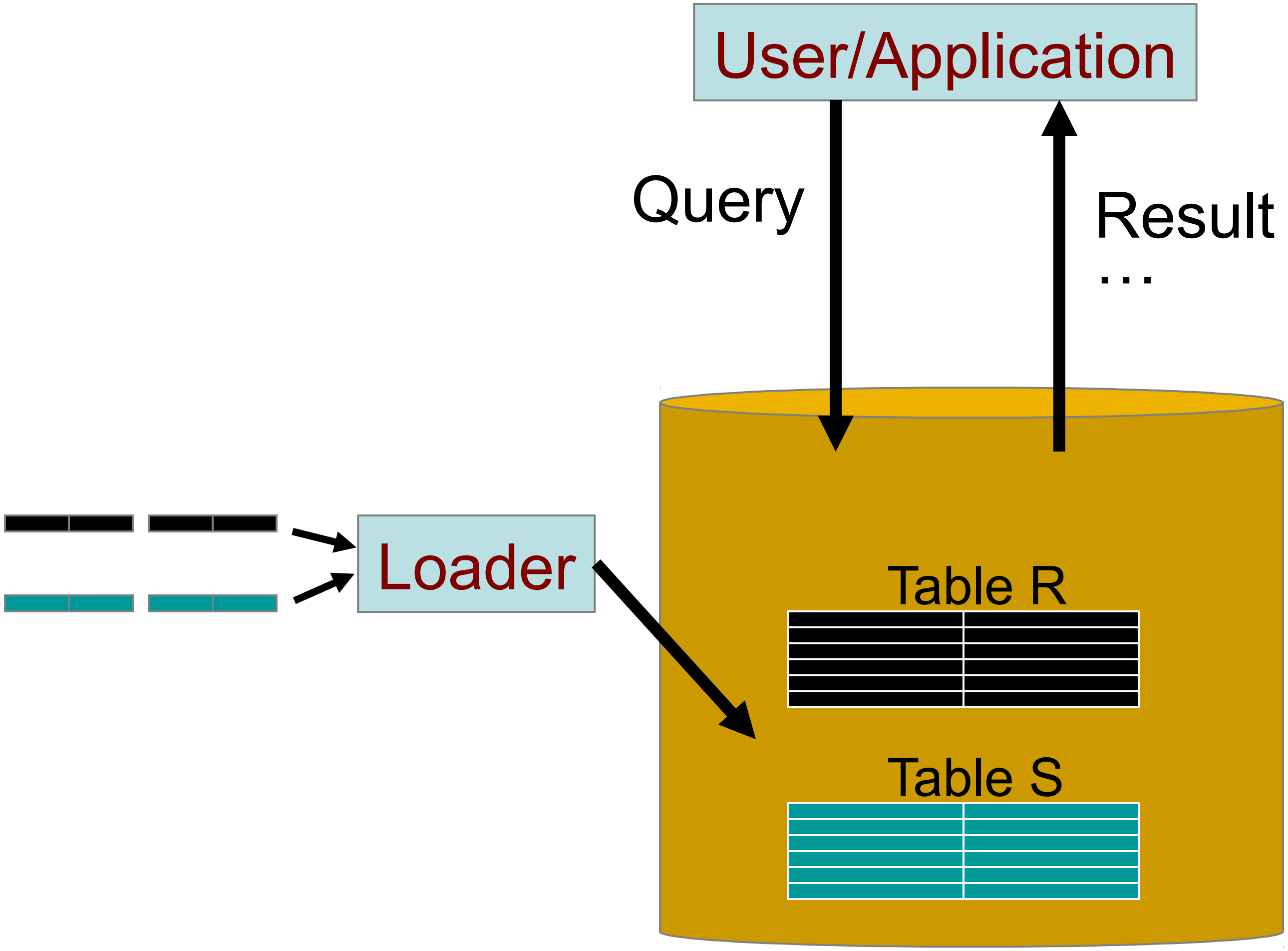
Modern DBMS Architecture



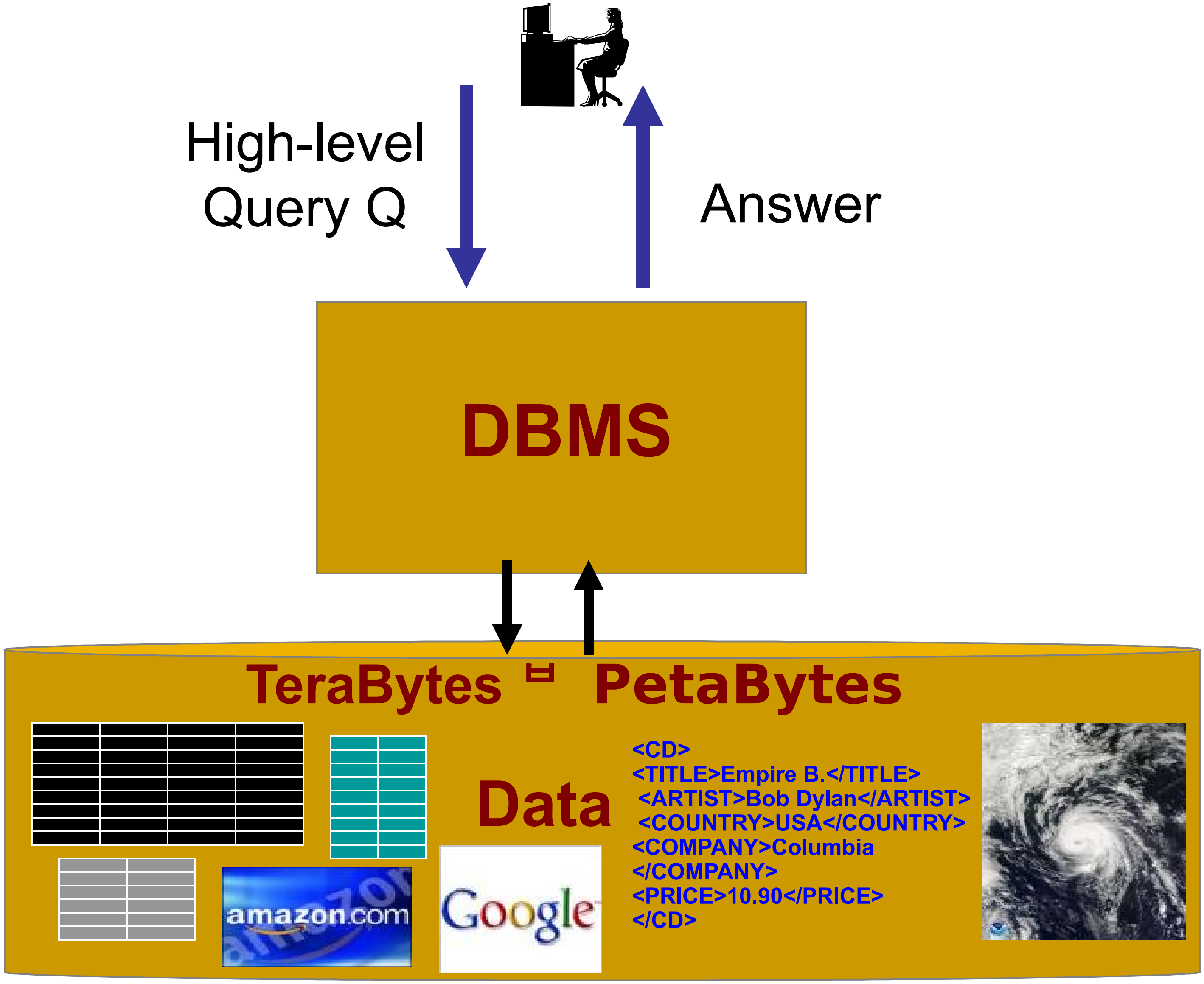
Course Outline

- 50% is about modern DBMS technology
  - Query execution, query optimization, transactions, recovery, etc.
  - Textbook material is starting point, but we will go beyond
- 50% is about one important class of “what is happening today in data management”: Data Integration
  - Structured vs unstructured data
  - Data is not locally stored
  - Data is in many places
  - Data is heterogeneous
  - Data is inconsistent

Using a Traditional DBMS



New Challenges in DBMSs



```
<CD>
<TITLE>Empire B.</TITLE>
 <ARTIST>Bob Dylan</ARTIST>
 <COUNTRY>USA</COUNTRY>
<COMPANY>Columbia
</COMPANY>
<PRICE>10.90</PRICE>
</CD>
```

Query Processing

Declarative Query → Query Plan

- NOTE: You will need to be familiar with SQL. We will
- Look at other query languages in class. A SQL refresher is available on the Wiki.

- Focus: Relational System and approach (i.e., data is organized as tables, or relations) is still central for query optimization. We will look at extensions, in particular nested relational algebra, and limited schema which are common in modern scenarios.

Data Integration: Overview

- ➢Introduction: data integration as a new abstraction
- Examples of data integration applications
- Schema heterogeneity
- Goal of data integration, why it’s a hard problem
- Data integration architectures

Data Integration

- Databases are great: they let us manage huge amounts of data
  - Assuming you’ve put it all into your schema.
- In reality, data sets are often created independently
  - Only to discover later that they need to combine their data!
  - At that point, they’re using different systems, different schemata and have limited interfaces to their data.
- The goal of data integration: tie together different sources, controlled by many people, under a common schema.

DBMS: it’s all about abstraction

- *Logical* vs. *Physical*; *What* vs. *How.*

Students:

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

Takes:

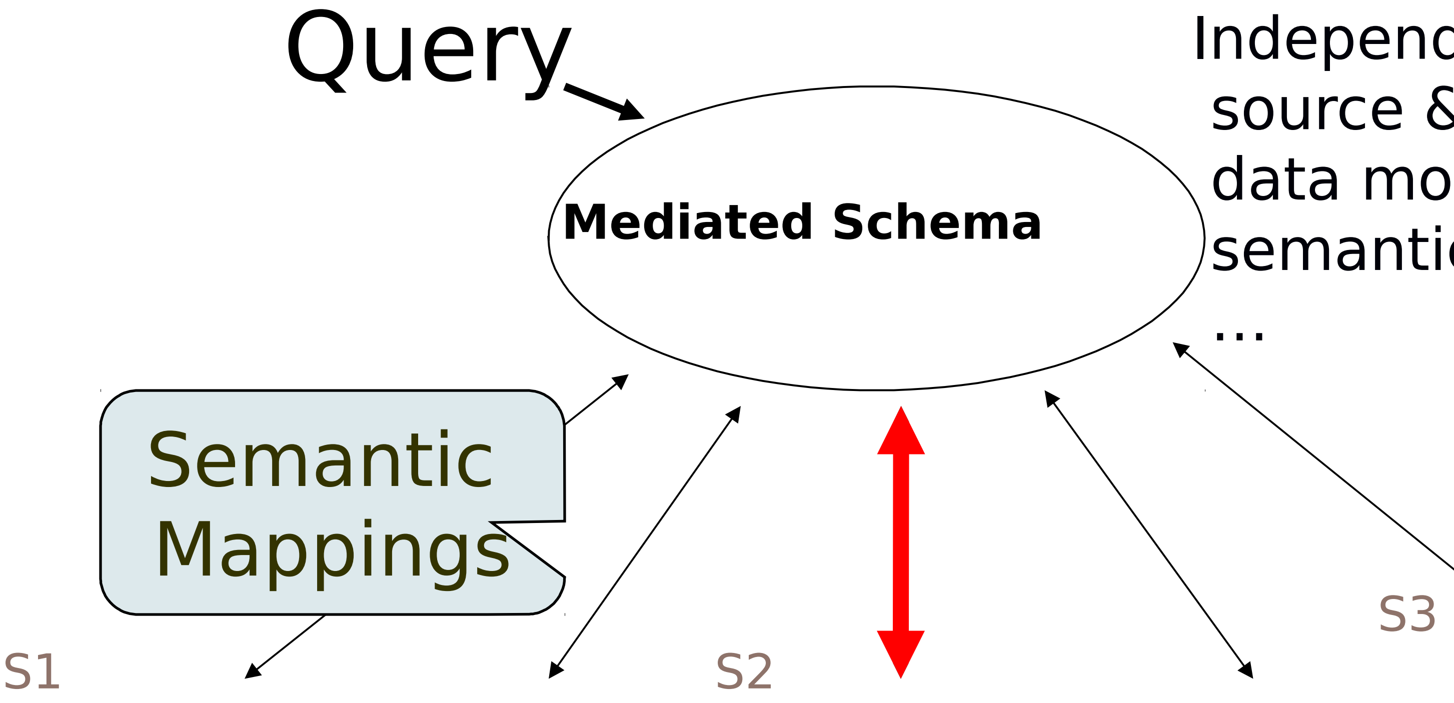
SSN	CID
123-45-6789	CSE444
123-45-6789	CSE444
234-56-7890	CSE142
	…

Courses:

CID	Name	Quarter
CSE444	Databases	fall
CSE541	Operating systems	winter

```
SELECT  C.name
FROM Students S, Takes T, Courses C
WHERE S.name=“Mary” and
        S.ssn = T.ssn and T.cid = C.cid
```

Data Integration:
A Higher-level Abstraction



Query

Mediated Schema

Independence of:
source & location
data model, syntax
semantic variations
…

Semantic Mappings

S1

S2

S3

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

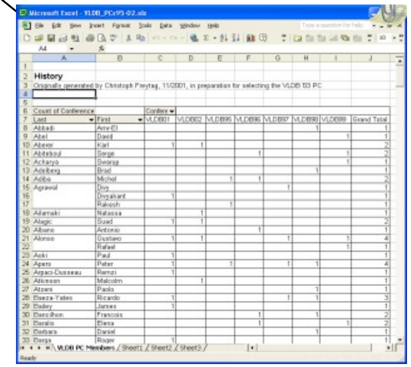
SSN	CID
123-45-6789	CSE444
123-45-6789	CSE444
234-56-7890	CSE142
	…

CID	Name	Quarter
CSE444	Databases	fall
CSE541	Operating systems	winter

. .

```
• <cd>   <title> The best of … </title>
   –            <artist> Carreras  </artist>
   –            <artist> Pavarotti  </artist>
   –            <artist> Domingo  </artist>
   –            <price> 19.95        </price>
                                 –   </cd>
```

. . .



Microsoft Excel - VLDB_PCs95-02.xls

History

Originally generated by Christoph Freytag, 11/2001, in preparation for selecting the VLDB 03 PC

Count of Conference / Confere

Last	First	VLDB01	VLDB02	VLDB95	VLDB96	VLDB97	VLDB98	VLDB99	Grand Total
Abbadi	Amr El						1		1
Abel	David							1	1
Aberer	Karl	1	1						2
Abiteboul	Serge				1			1	2
Acharya	Swarup							1	1
Adelberg	Brad						1		1
Adiba	Michel			1	1				2
Agrawal	Divy					1			1
	Divyakant	1							1
	Rakesh			1					1
Ailamaki	Natassa		1						1
Alagic	Suad	1	1						2
Albano	Antonio				1				1
Alonso	Gustavo	1	1			1		1	4
	Rafael							1	1
Aoki	Paul	1							1
Apers	Peter	1		1		1	1		4
Arpaci-Dusseau	Remzi	1							1
Atkinson	Malcolm		1						1
Atzeni	Paolo						1		1
Baeza-Yates	Ricardo	1				1	1		3
Bailey	James	1							1
Bancilhon	Francois				1		1		2
Baralis	Elena				1			1	2
Barbara	Daniel						1		1
Barga	Roger	1							1

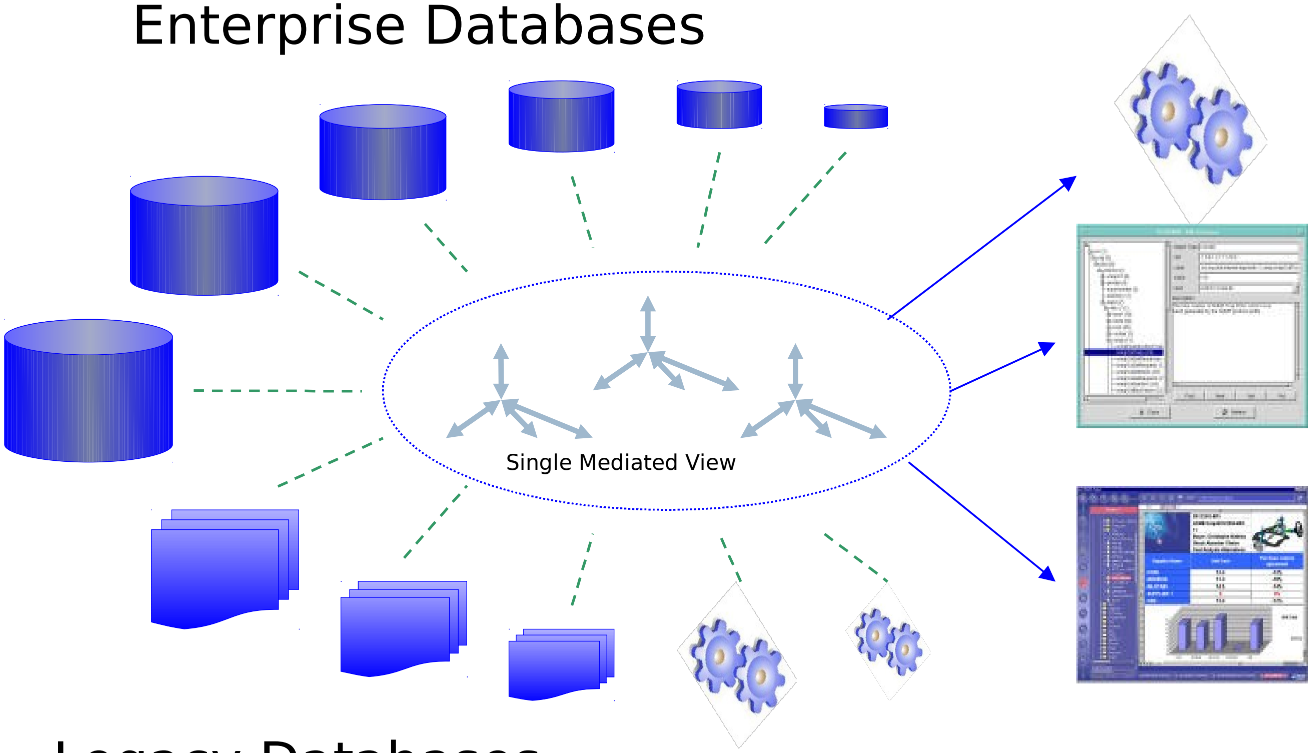
Outline

- ✓ Introduction: data integration as a new abstraction
- ➢ Examples of data integration applications
- ▪ Schema heterogeneity
- ▪ Goal of data integration, why it’s a hard problem
- ▪ Data integration architectures

Applications of Data Integration

- Business
- Science
- Government
- The Web
- Pretty much everywhere

Application Area 1: Business



Enterprise Databases

EII Apps:

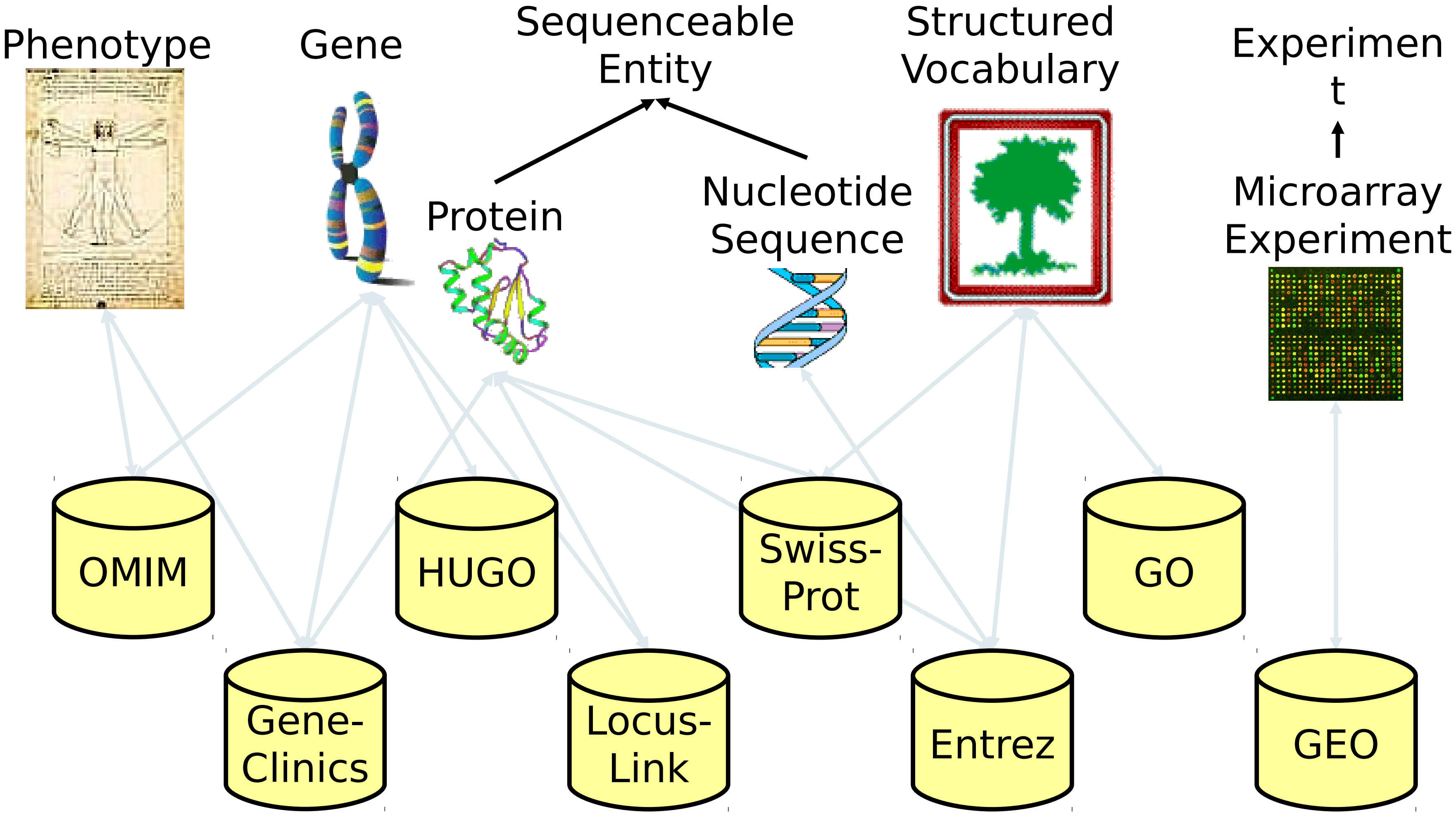
CRM
ERP
Portals
…

Legacy Databases

Services and Applications

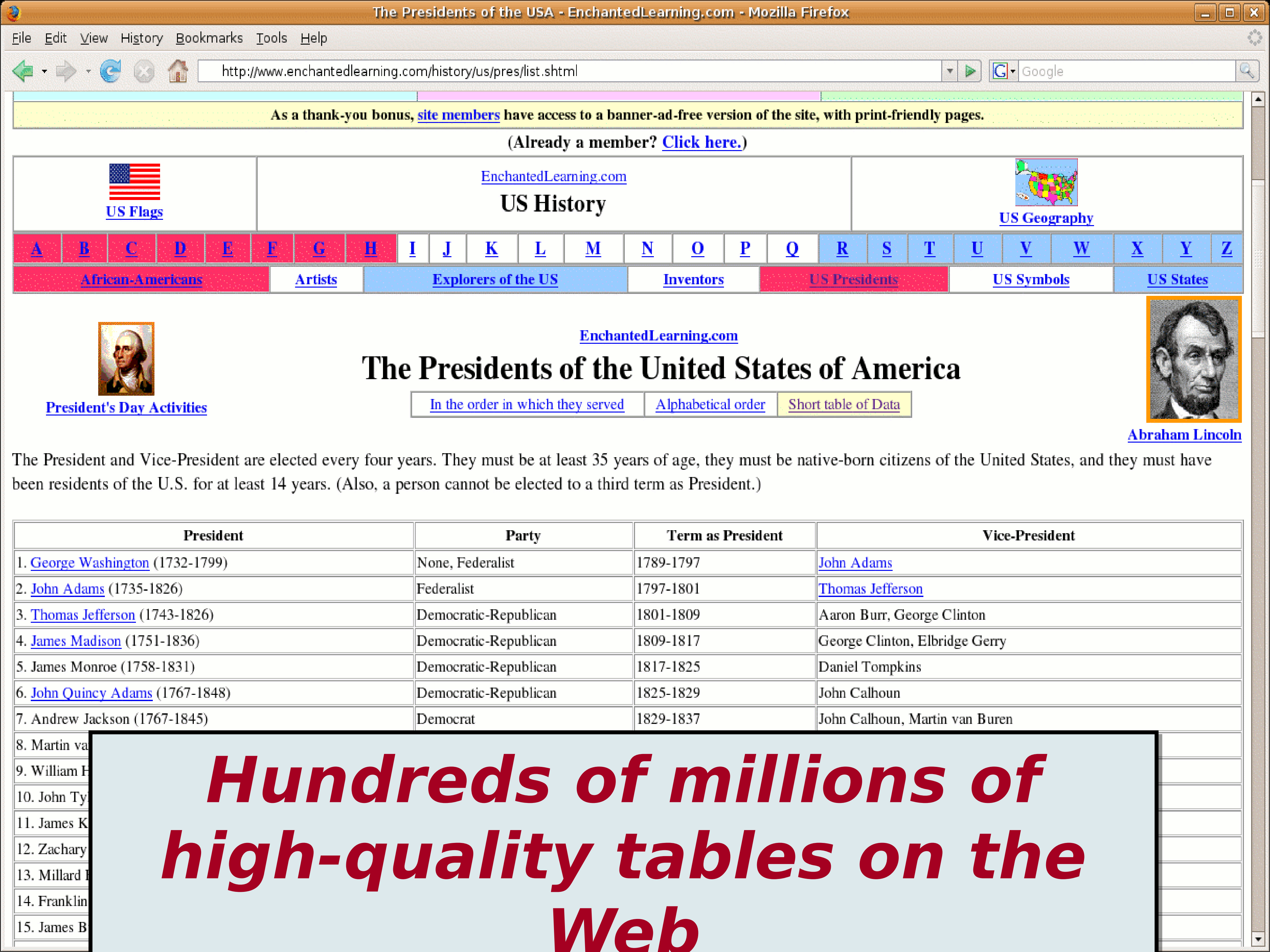
50% of all IT $$$ spent here!

Application Area 2: Science



Hundreds of biomedical data sources available; growing rapidly!

Application Area 3: The Web


The Presidents of the USA - EnchantedLearning.com - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.enchantedlearning.com/history/us/pres/list.shtml
Google

As a thank-you bonus, site members have access to a banner-ad-free version of the site, with print-friendly pages.

(Already a member? Click here.)

US Flags	EnchantedLearning.com US History	US Geography

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

African-Americans | Artists | Explorers of the US | Inventors | US Presidents | US Symbols | US States

President's Day Activities

EnchantedLearning.com

The Presidents of the United States of America

In the order in which they served | Alphabetical order | Short table of Data

Abraham Lincoln

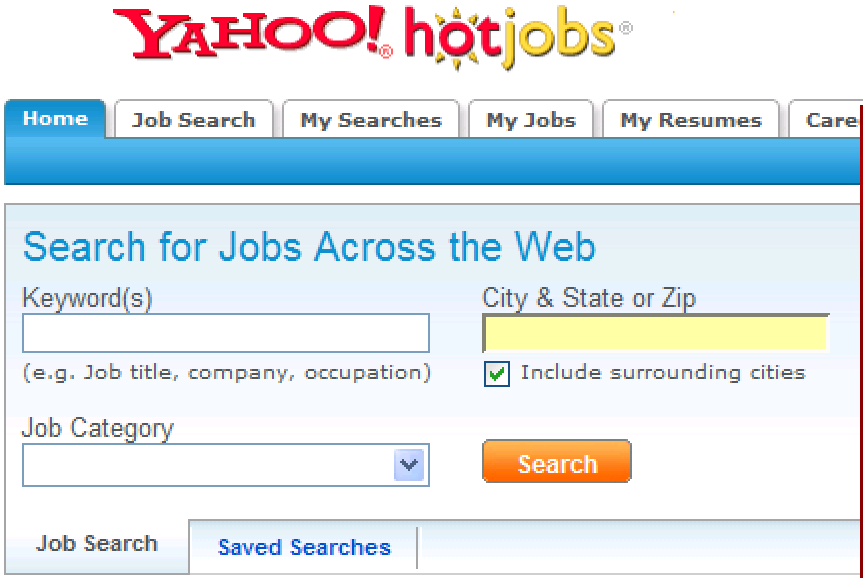
The President and Vice-President are elected every four years. They must be at least 35 years of age, they must be native-born citizens of the United States, and they must have been residents of the U.S. for at least 14 years. (Also, a person cannot be elected to a third term as President.)

President	Party	Term as President	Vice-President
1. George Washington (1732-1799)	None, Federalist	1789-1797	John Adams
2. John Adams (1735-1826)	Federalist	1797-1801	Thomas Jefferson
3. Thomas Jefferson (1743-1826)	Democratic-Republican	1801-1809	Aaron Burr, George Clinton
4. James Madison (1751-1836)	Democratic-Republican	1809-1817	George Clinton, Elbridge Gerry
5. James Monroe (1758-1831)	Democratic-Republican	1817-1825	Daniel Tompkins
6. John Quincy Adams (1767-1848)	Democratic-Republican	1825-1829	John Calhoun
7. Andrew Jackson (1767-1845)	Democrat	1829-1837	John Calhoun, Martin van Buren
8. Martin va			
9. William H			
10. John Ty			
11. James K			
12. Zachary			
13. Millard F			
14. Franklin			
15. James B			

Hundreds of millions of high-quality tables on the Web

The Deep Web

- Millions of high quality HTML forms out there
- Each form has its own special interface
  - Hard to explore data across sites.
- Goal (for some domains):
  - A single interface into a multitude of deep-web sources.



YAHOO! hotjobs

Home | Job Search | My Searches | My Jobs | My Resumes | Care

Search for Jobs Across the Web

Keyword(s)

(e.g. Job title, company, occupation)

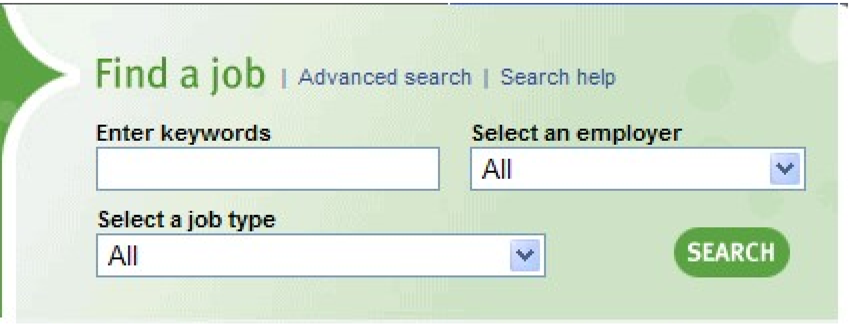
City & State or Zip

☑ Include surrounding cities

Job Category

Search

Job Search | Saved Searches



Find a job

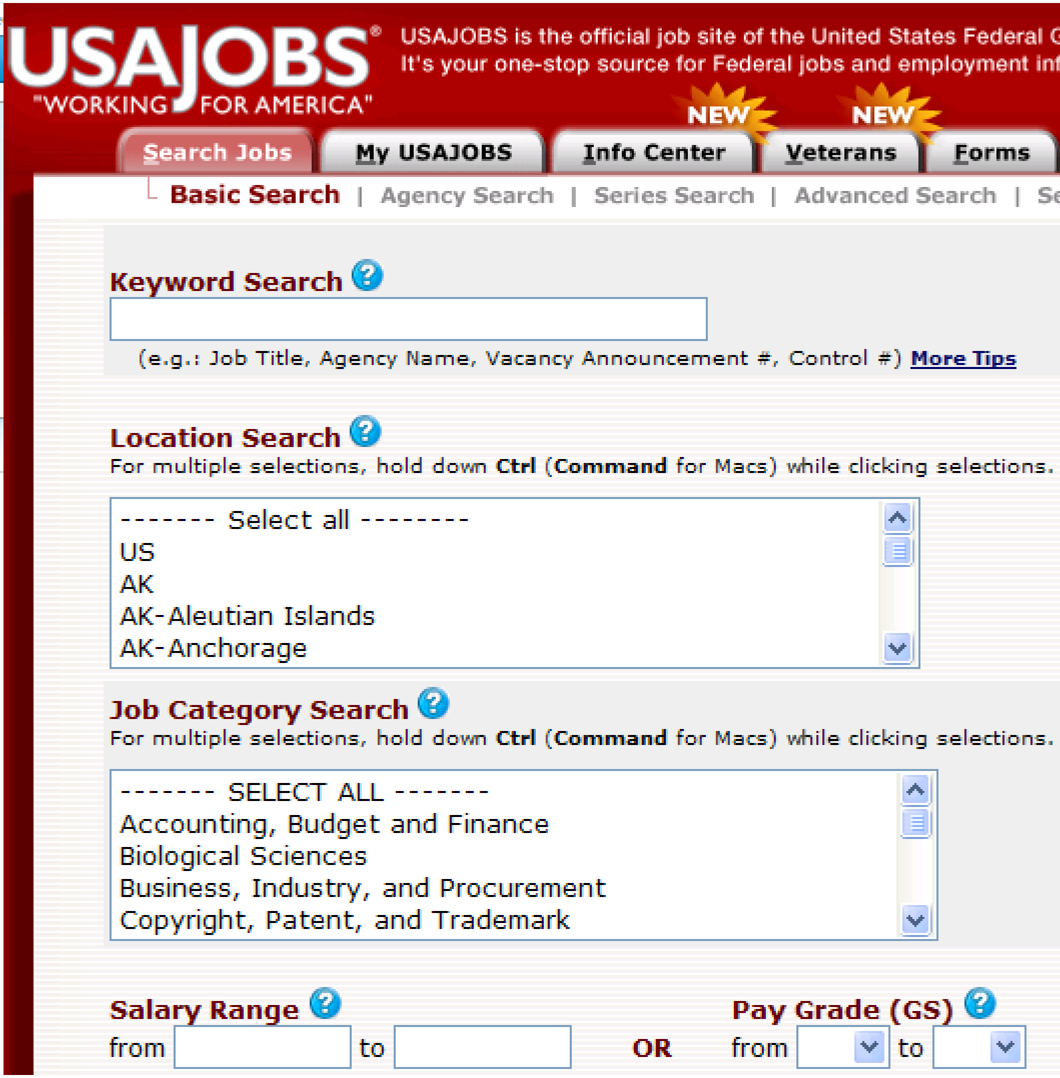
| Advanced search | Search help

Enter keywords

Select an employer: All

Select a job type: All

SEARCH



USAJOBS "WORKING FOR AMERICA"

USAJOBS is the official job site of the United States Federal G

It's your one-stop source for Federal jobs and employment inf

NEW NEW

Search Jobs | My USAJOBS | Info Center | Veterans | Forms

Basic Search | Agency Search | Series Search | Advanced Search | Se

Keyword Search

(e.g.: Job Title, Agency Name, Vacancy Announcement #, Control #) More Tips

Location Search

For multiple selections, hold down **Ctrl** (**Command** for Macs) while clicking selections.

------- Select all --------
US
AK
AK-Aleutian Islands
AK-Anchorage

Job Category Search

For multiple selections, hold down **Ctrl** (**Command** for Macs) while clicking selections.

------- SELECT ALL -------
Accounting, Budget and Finance
Biological Sciences
Business, Industry, and Procurement
Copyright, Patent, and Trademark

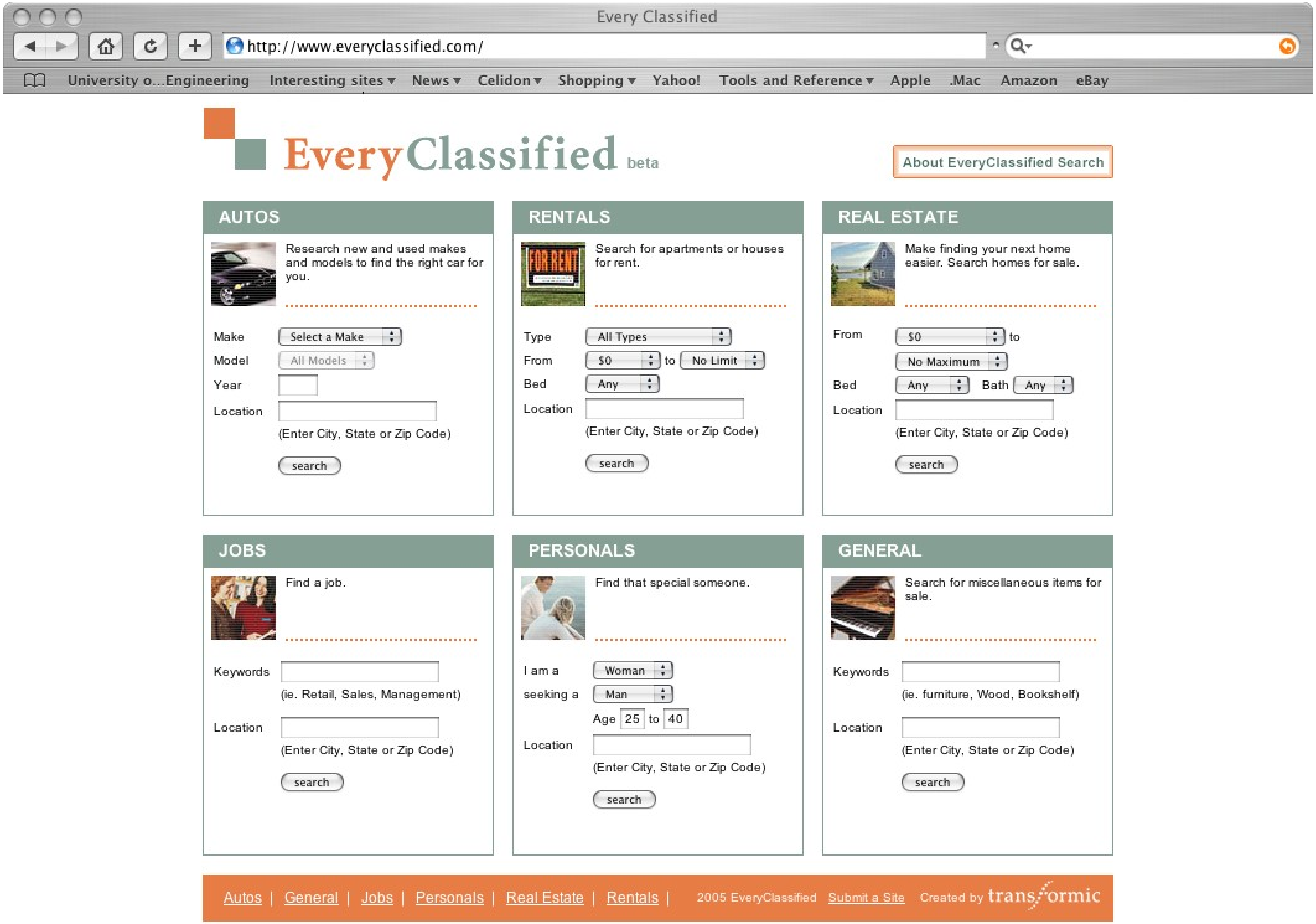
Salary Range

from ___ to ___

OR

Pay Grade (GS)

from ___ to ___



EveryClassified beta

About EveryClassified Search

AUTOS

Research new and used makes and models to find the right car for you.

Make: Select a Make
Model: All Models
Year:
Location:
(Enter City, State or Zip Code)

search

RENTALS

Search for apartments or houses for rent.

Type: All Types
From: $0 to No Limit
Bed: Any
Location:
(Enter City, State or Zip Code)

search

REAL ESTATE

Make finding your next home easier. Search homes for sale.

From: $0 to No Maximum
Bed: Any Bath: Any
Location:
(Enter City, State or Zip Code)

search

JOBS

Find a job.

Keywords:
(ie. Retail, Sales, Management)
Location:
(Enter City, State or Zip Code)

search

PERSONALS

Find that special someone.

I am a: Woman
seeking a: Man
Age 25 to 40
Location:
(Enter City, State or Zip Code)

search

GENERAL

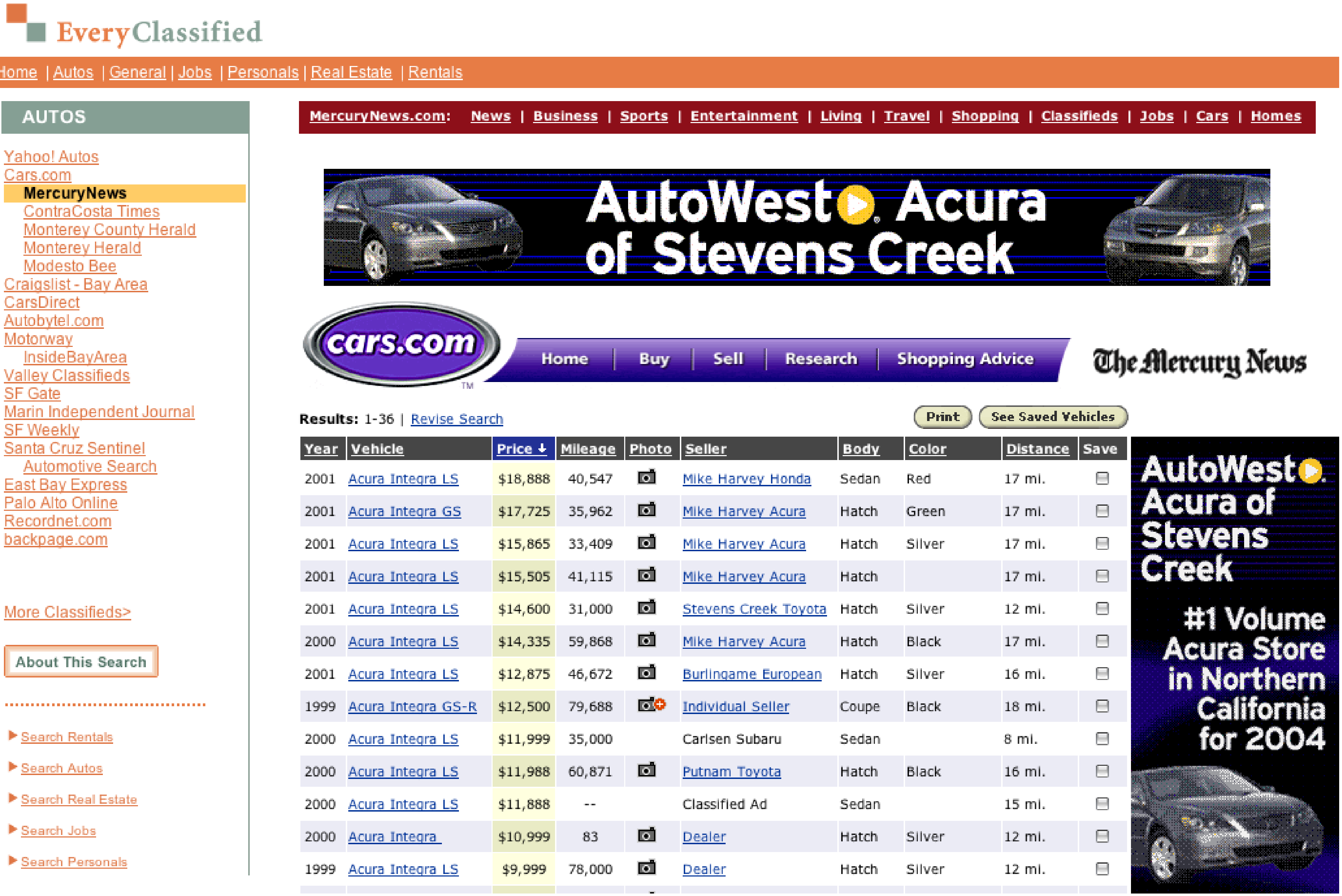
Search for miscellaneous items for sale.

Keywords:
(ie. furniture, Wood, Bookshelf)
Location:
(Enter City, State or Zip Code)

search

Autos | General | Jobs | Personals | Real Estate | Rentals | 2005 EveryClassified Submit a Site Created by transformic

reate a single site to search for jobs/rentals/..



EveryClassified

Home | Autos | General | Jobs | Personals | Real Estate | Rentals

AUTOS

Yahoo! Autos
Cars.com
MercuryNews
ContraCosta Times
Monterey County Herald
Monterey Herald
Modesto Bee
Craigslist - Bay Area
CarsDirect
Autobytel.com
Motorway
InsideBayArea
Valley Classifieds
SF Gate
Marin Independent Journal
SF Weekly
Santa Cruz Sentinel
Automotive Search
East Bay Express
Palo Alto Online
Recordnet.com
backpage.com

More Classifieds>

About This Search

- Search Rentals
- Search Autos
- Search Real Estate
- Search Jobs
- Search Personals

MercuryNews.com: News | Business | Sports | Entertainment | Living | Travel | Shopping | Classifieds | Jobs | Cars | Homes

cars.com Home | Buy | Sell | Research | Shopping Advice — The Mercury News

Results: 1-36 | Revise Search

Print | See Saved Vehicles

Year	Vehicle	Price ↓	Mileage	Photo	Seller	Body	Color	Distance	Save
2001	Acura Integra LS	$18,888	40,547	📷	Mike Harvey Honda	Sedan	Red	17 mi.	□
2001	Acura Integra GS	$17,725	35,962	📷	Mike Harvey Acura	Hatch	Green	17 mi.	□
2001	Acura Integra LS	$15,865	33,409	📷	Mike Harvey Acura	Hatch	Silver	17 mi.	□
2001	Acura Integra LS	$15,505	41,115	📷	Mike Harvey Acura	Hatch		17 mi.	□
2001	Acura Integra LS	$14,600	31,000	📷	Stevens Creek Toyota	Hatch	Silver	12 mi.	□
2000	Acura Integra LS	$14,335	59,868	📷	Mike Harvey Acura	Hatch	Black	17 mi.	□
2001	Acura Integra LS	$12,875	46,672	📷	Burlingame European	Hatch	Silver	16 mi.	□
1999	Acura Integra GS-R	$12,500	79,688	📷	Individual Seller	Coupe	Black	18 mi.	□
2000	Acura Integra LS	$11,999	35,000		Carlsen Subaru	Sedan		8 mi.	□
2000	Acura Integra LS	$11,988	60,871	📷	Putnam Toyota	Hatch	Black	16 mi.	□
2000	Acura Integra LS	$11,888	--		Classified Ad	Sedan		15 mi.	□
2000	Acura Integra	$10,999	83	📷	Dealer	Hatch	Silver	12 mi.	□
1999	Acura Integra LS	$9,999	78,000	📷	Dealer	Hatch	Silver	12 mi.	□

ly traverse between the site by clicking its na

Outline

- ✓ Introduction: data integration as a new abstraction
- ✓ Examples of data integration applications
- ➢ Schema heterogeneity
- ▪ Goal of data integration, why it’s a hard problem
- ▪ Data integration architectures

Enterprise Data Integration:

FullServe Corporation

Employees

- FullTimeEmp
- Hire
- TempEmployees

Training

- Courses
- Enrollments

Sales

- Products
- Sales

Resumes

- Interview
- CV

Services

- Services
- Customers
- Contracts

HelpLine

- Calls

EuroCard Corporation

Employees

Employees
Hire

Resumes

Interview

Credit Cards

Customer
CustDetail

HelpLine

Calls

Examples of Heterogeneity

FullServe

FullTimeEmp

ssn, empId, firstName
middleName, lastName

Hire

empId, hireDate, recruiter

TempEmployees

ssn, hireStart, hireEnd

EuroCard

Employees

ID, firstNameMiddleInitial,
lastName

Hire

ID, hireDate, recruiter

Find all employees (making over $100K)

Customer Call Center

Agents should have a full view of customer when they call in.

Sales

Products
Sales

Credit Cards

Customer
CustDetail

Services

Services
Customers
Contracts

Other Reasons to Integrate Data

- Create a (useful) web site for tracking services
- Collaborate with third parties
  - E.g., create branded services
- Comply with government regulations
  - Find “risky” employees
- Business intelligence
  - What’s really wrong with our products?

Outline

- ✓ Introduction: data integration as a new abstraction
- ✓ Examples of data integration applications
- ✓ Schema heterogeneity
- ➢ Goal of data integration, why it’s a hard problem
- ▪ Data integration architectures

Goal of Data Integration

- Uniform query access to a set of data sources
- Handle:
  - Scale of sources: from tens to millions
  - Heterogeneity
  - Autonomy
  - Semi-structure

Why is it Hard?

- Systems-level reasons:
  - Managing different platforms
  - SQL across multiple systems is not so simple
  - Distributed query processing
- Logical reasons:
  - Schema (and data) heterogeneity
- ‘Social’ reasons:
  - Locating and capturing relevant data in the enterprise.
  - Convincing people to share (data fiefdoms)
    - Security, privacy and performance implications.

Setting Expectations

Data integration is **AI-Complete**.

- Completely automated solutions unlikely.

Goal 1:

- Reduce the effort needed to set up an integration application.

Goal 2:

- Enable the system to perform gracefully with uncertainty (e.g., on the web)

Data Integration Smorgasbord

Something for everyone:

- **Theory** of modeling data sources
- **Systems** aspects of data integration
- **Architectural** issues: e.g., P2P data sharing
- **AI** @ work: automated schema matching
- **Web**: latest on data integration & web
- **Commercial** products: BEA, IBM
- **Semantic Web**: what does it have to offer?
- New trends in DBMS: **uncertainty, dataspaces**

Outline

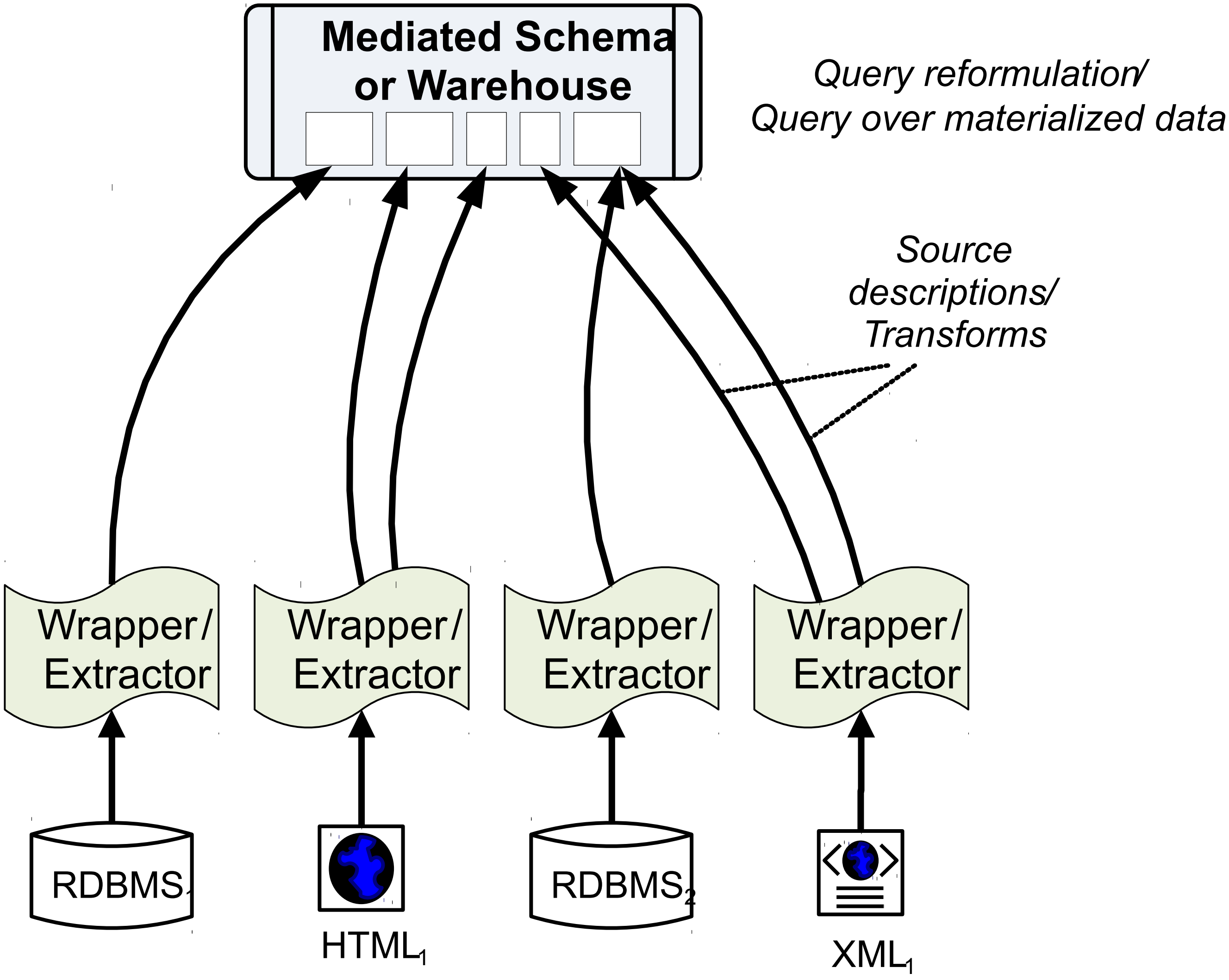
- ✓ Introduction: data integration as a new abstraction
- ✓ Examples of data integration applications
- ✓ Schema heterogeneity
- ✓ Goal of data integration, why it’s a hard problem
- ➢ Data integration architectures

Virtual, Warehousing and in Between

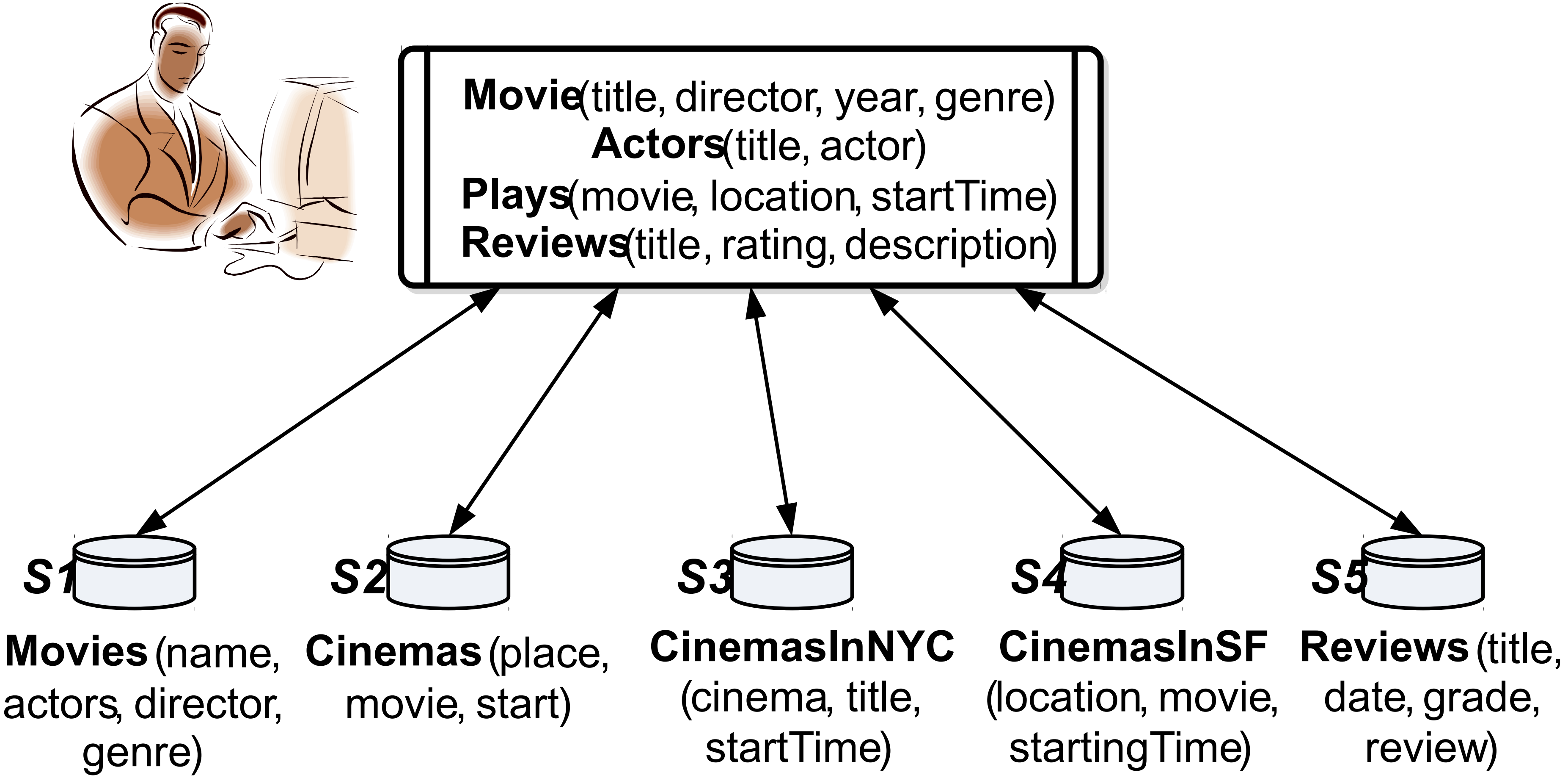
- Data warehousing: integrate by bringing the data into a single physical warehouse
- Virtual data integration: leave the data at the sources and access it at query time.

- Some differences, but semantic heterogeneity arises in both cases.
- Numerous intermediate architectures.
- The course illustrates data integration technology mostly through the virtual architecture.

Virtual Data Integration Architecture



Example



Movie(title, director, year, genre)
Actors(title, actor)
Plays(movie, location, startTime)
Reviews(title, rating, description)

S1 **Movies** (name, actors, director, genre)

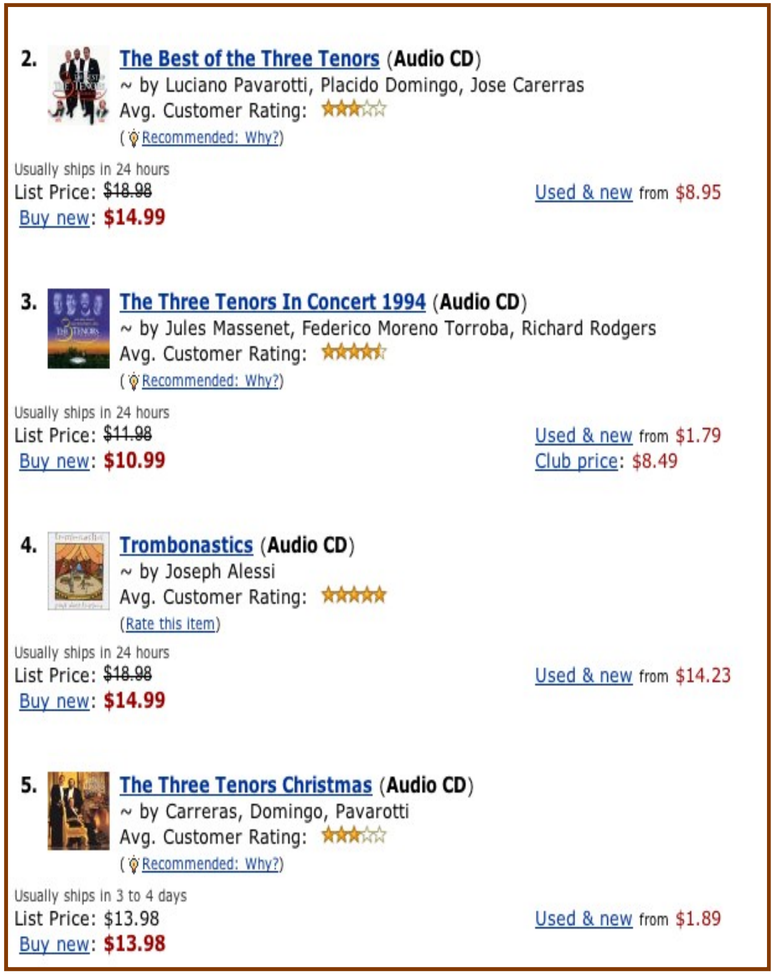
S2 **Cinemas** (place, movie, start)

S3 **CinemasInNYC** (cinema, title, startTime)

S4 **CinemasInSF** (location, movie, startingTime)

S5 **Reviews** (title, date, grade, review)

Wrappers



2. **The Best of the Three Tenors** (**Audio CD**)
~ by Luciano Pavarotti, Placido Domingo, Jose Carerras
Avg. Customer Rating: ★★★☆☆
(Recommended: Why?)

Usually ships in 24 hours
List Price: ~~$18.98~~
Buy new: **$14.99**
Used & new from $8.95

3. **The Three Tenors In Concert 1994** (**Audio CD**)
~ by Jules Massenet, Federico Moreno Torroba, Richard Rodgers
Avg. Customer Rating: ★★★★½
(Recommended: Why?)

Usually ships in 24 hours
List Price: ~~$11.98~~
Buy new: **$10.99**
Used & new from $1.79
Club price: $8.49

4. **Trombonastics** (**Audio CD**)
~ by Joseph Alessi
Avg. Customer Rating: ★★★★★
(Rate this item)

Usually ships in 24 hours
List Price: ~~$18.98~~
Buy new: **$14.99**
Used & new from $14.23

5. **The Three Tenors Christmas** (**Audio CD**)
~ by Carreras, Domingo, Pavarotti
Avg. Customer Rating: ★★★☆☆
(Recommended: Why?)

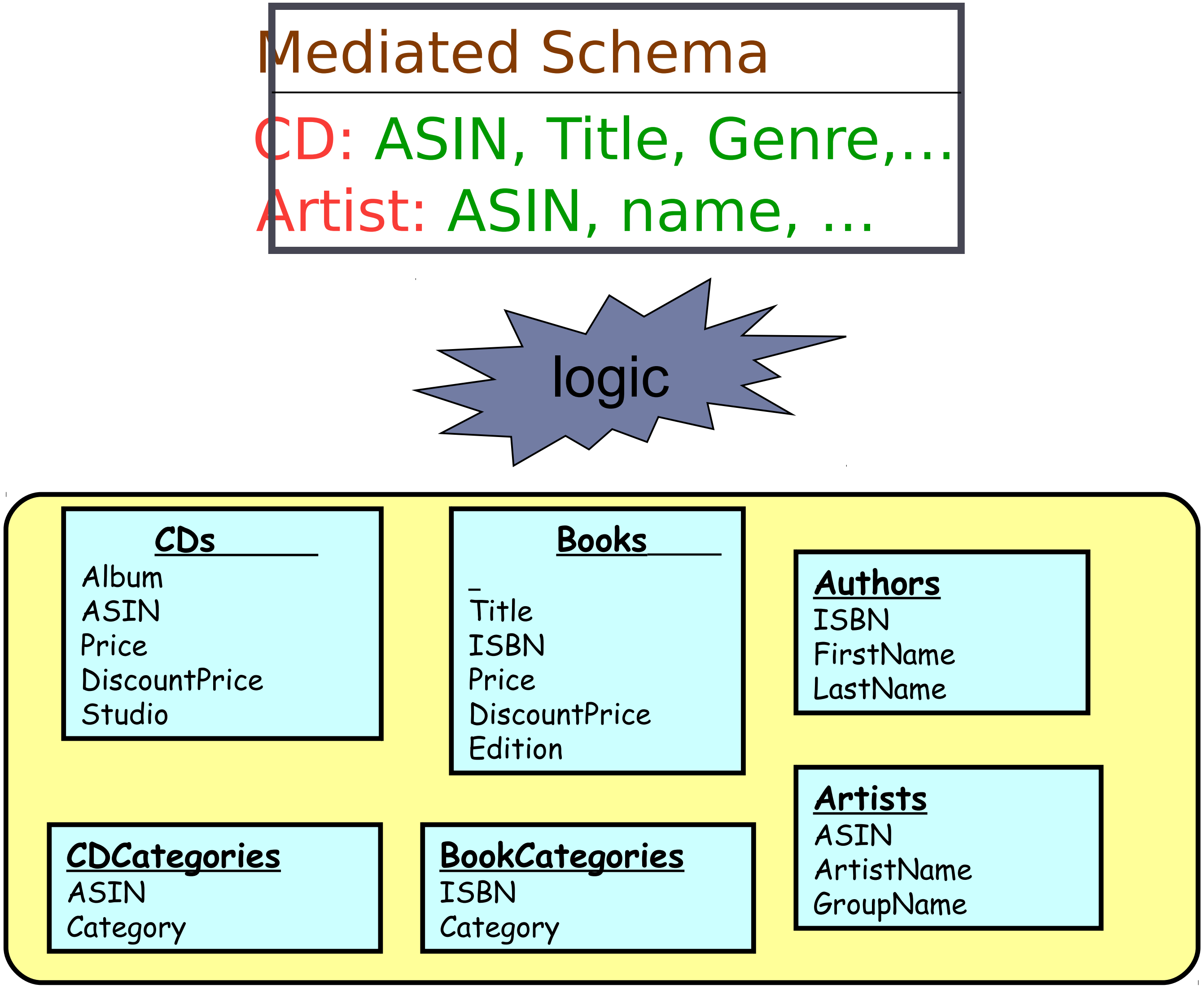
Usually ships in 3 to 4 days
List Price: $13.98
Buy new: **$13.98**
Used & new from $1.89

```
<cd>    <title> The best of … </title>
              <artist> Abiteboul </artist>
              <artist> Pavarotti  </artist>
              <artist> Domingo  </artist>
              <price> 19.95      </price>
     </cd>
      …
```

Send queries to data sources and transform answers into tuples (or other internal data model). (Chapter 9)

Mediation Languages

Describe relationships between mediated schema and data sources (Chapter 3).



Mediated Schema

CD: ASIN, Title, Genre,…

Artist: ASIN, name, …

logic

CDs
Album
ASIN
Price
DiscountPrice
Studio

Books
Title
ISBN
Price
DiscountPrice
Edition

Authors
ISBN
FirstName
LastName

CDCategories
ASIN
Category

BookCategories
ISBN
Category

Artists
ASIN
ArtistName
GroupName

Woody Allen Comedies in NY

Mediated schema:

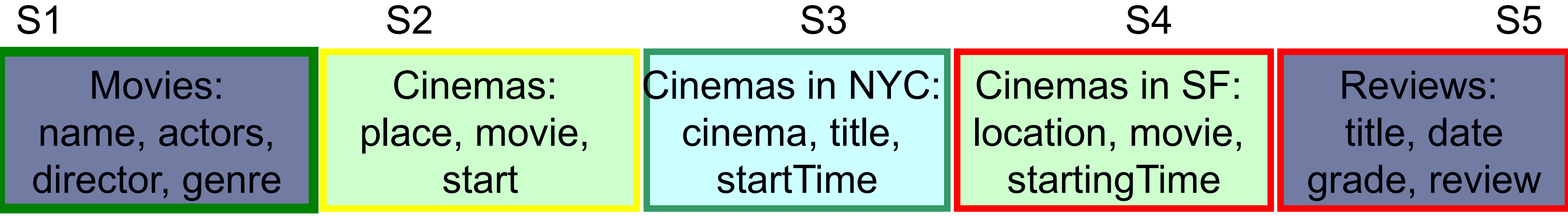
Movie: Title, director, year, genre
Actors: title, actor
Plays: movie, location, startTime
Reviews: title, rating, description

```
select title, startTime
from Movie, Plays
where Movie.title=Plays.movie AND
      location="New York"  AND
      director="Woody Allen"
```

Movie: Title, director, year, genre
Actors: title, actor
Plays: movie, location, startTime
Reviews: title, rating, description

```
select title, startTime
from Movie, Plays
where Movie.title=Plays.movie AND
        location="New York"  AND
        director="Woody Allen"
```

Sources S1 and S3 are relevant, sources S4 and S5 are irrelevant, and source S2 is relevant but possibly redundant.



S1
Movies:
name, actors,
director, genre

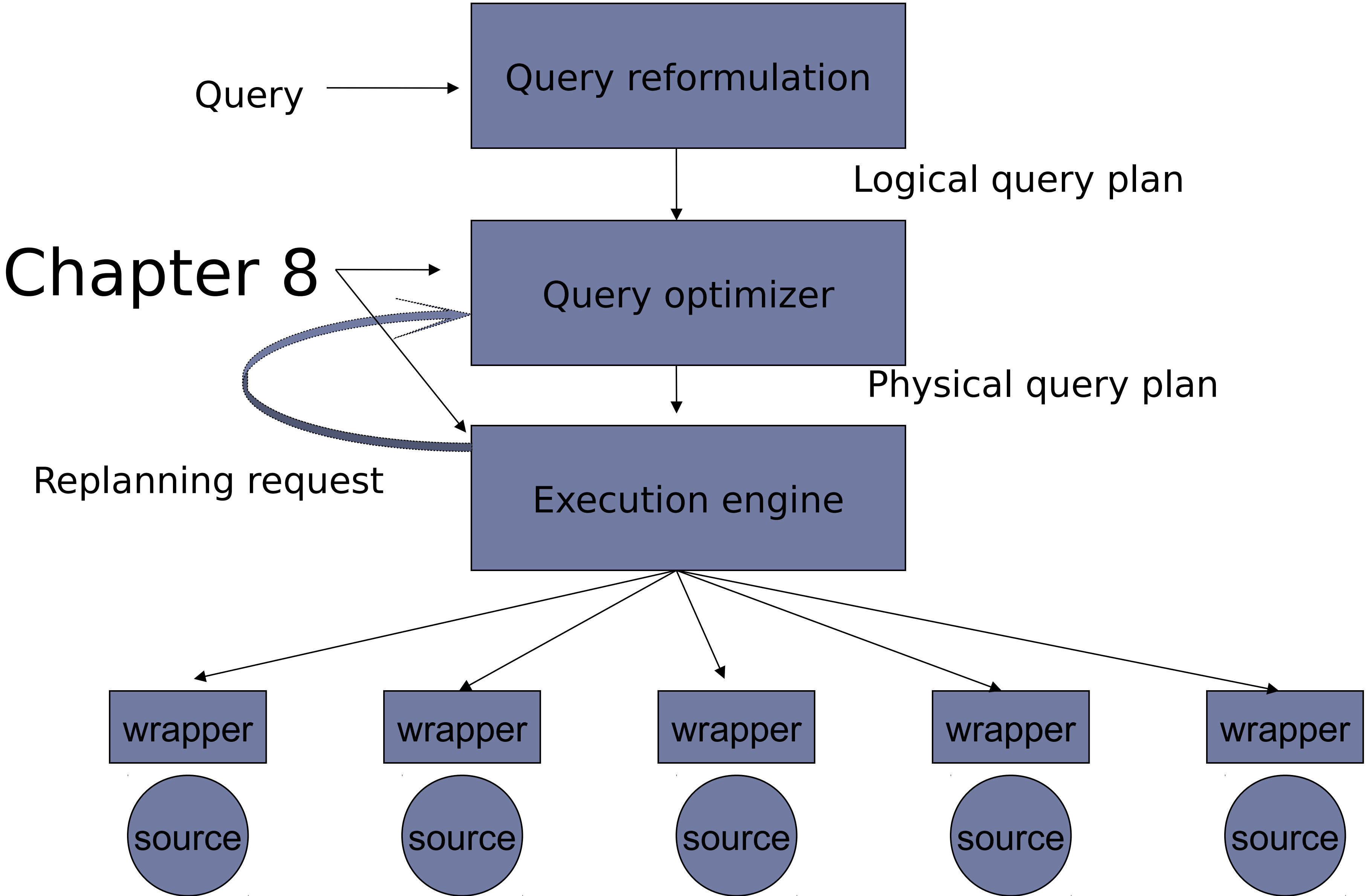
S2
Cinemas:
place, movie,
start

S3
Cinemas in NYC:
cinema, title,
startTime

S4
Cinemas in SF:
location, movie,
startingTime

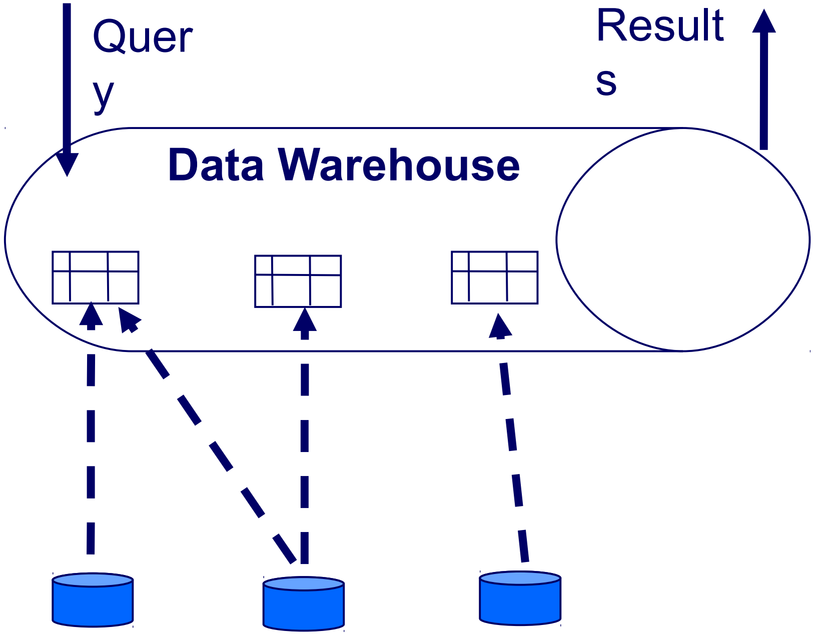
S5
Reviews:
title, date
grade, review

Query Processing



Data Warehouses – Offline Replication

- Determine physical schema
- Define a database with this schema
- Define procedural *mappings* in an “ETL tool” to import the data and clean it.
- Periodically copy all of the data from the data sources
  - Note that the sources and the



Pros and Cons of Data Warehouses

- ✘ Need to spend time to design the physical database layout, as well as logical
  - ✘ This actually takes a lot of effort!
- ✘ Data is generally not up-to-date (lazy or offline refresh)

- ✓ Queries over the warehouse don’t disrupt the data sources
- ✓ Can run very heavy-duty computations, including data mining and cleaning

Summary of Chapter 1

- Data integration: abstract away the fact that data comes from multiple sources in varying schemata.
- Problem occurs everywhere: it’s key to business, science, Web and government.
- Goal: reduce the effort involved in integrating.
- Regardless of the architecture, heterogeneity is a key issue.
- Architectures range from warehousing to virtual integration.

Summary: Data Management is Important

- Assumed and used in most sciences and engineering today
- Core need in industry
- Cool mix of theory and systems
- Chances are you will find something interesting even if you primary interest is elsewhere

Next Class

Database Query Compilation & Optimization Principles, Indices