MASSACHUSETTS INSTITUTE OF TECHNOLOGY
SLOAN SCHOOL OF MANAGEMENT

15.565 Integrating Information Systems:

Technology, Strategy, and Organizational Factors

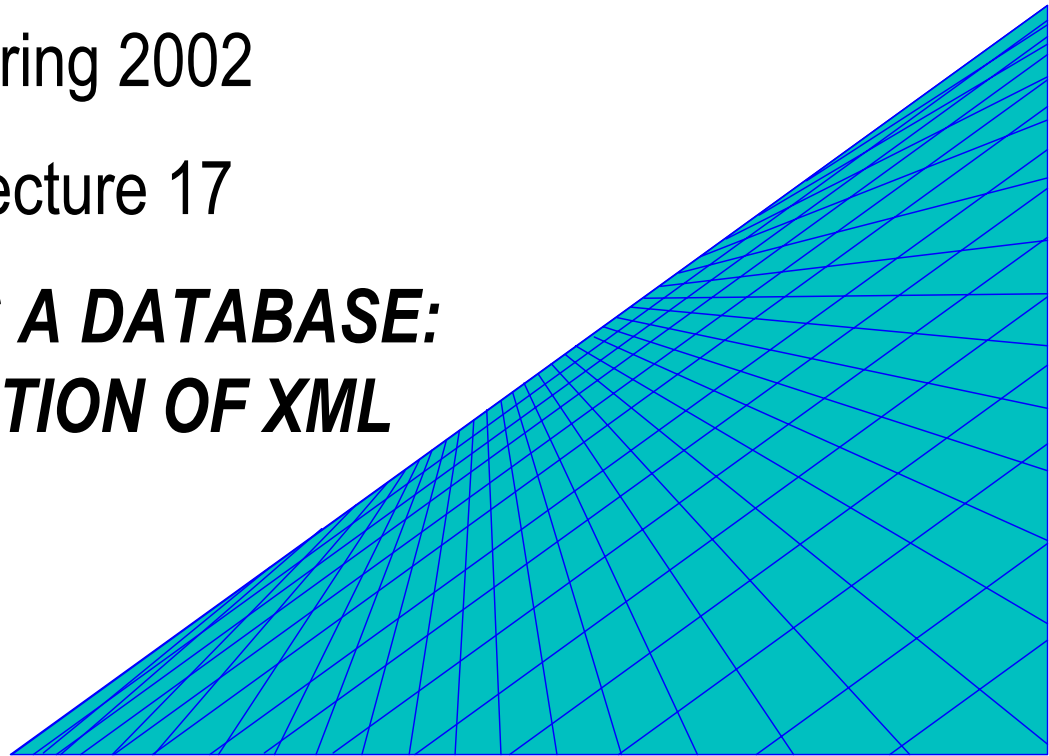
15.578 Global Information Systems:

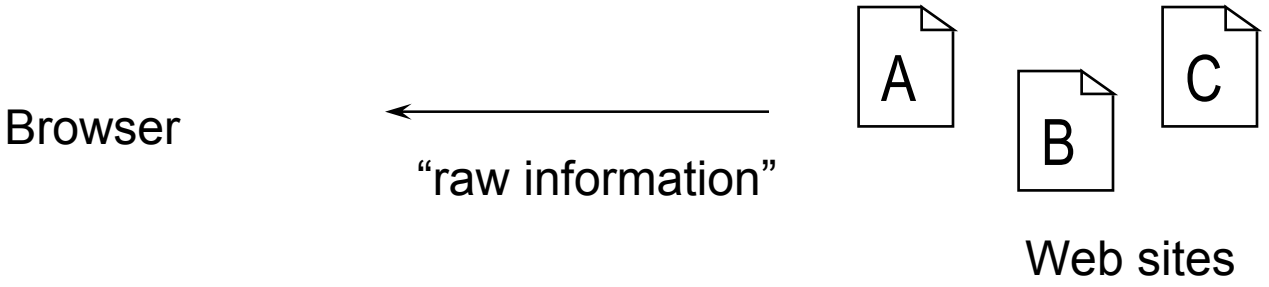
Communications & Connectivity Among Information Systems

Spring 2002

Lecture 17

WEB AS A DATABASE:
EVOLUTION OF XML





Traditional Use of the Web: For direct human usage Focus: **Entertainment**

Examples:

Extract just mortgage rate information

Compare mortgage rates offered by multiple sources

Build **cumulative** database of mortgage rates over time

Compare cumulative rates with previously stored, **alert** to new highs and lows

Programs for:

Web Wrapper and Context Mediator Technology

Enhancing

Analyzing

Consolidating

Processing

Repository

Exception reports

A

B

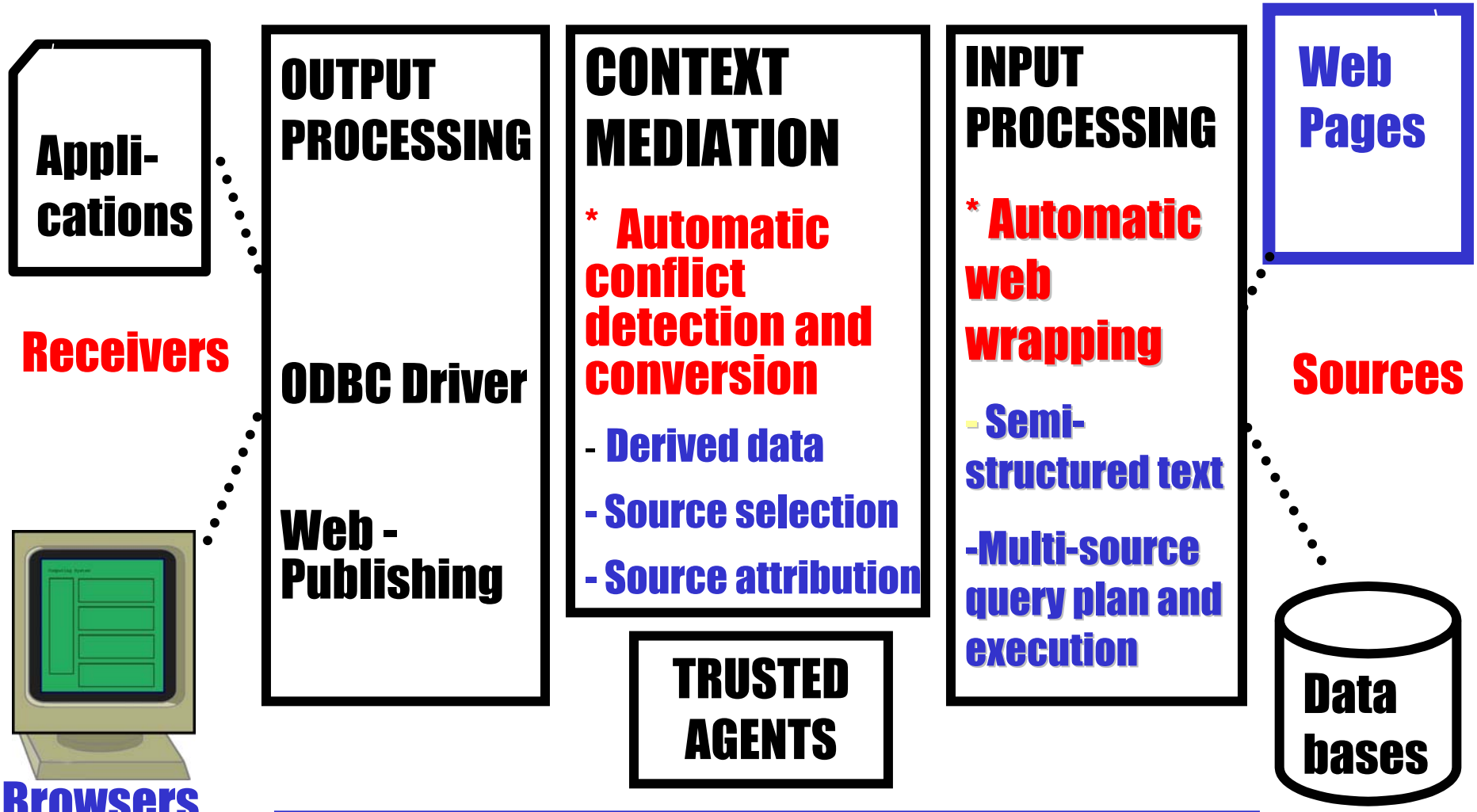
C

Web sites

Internal data bases

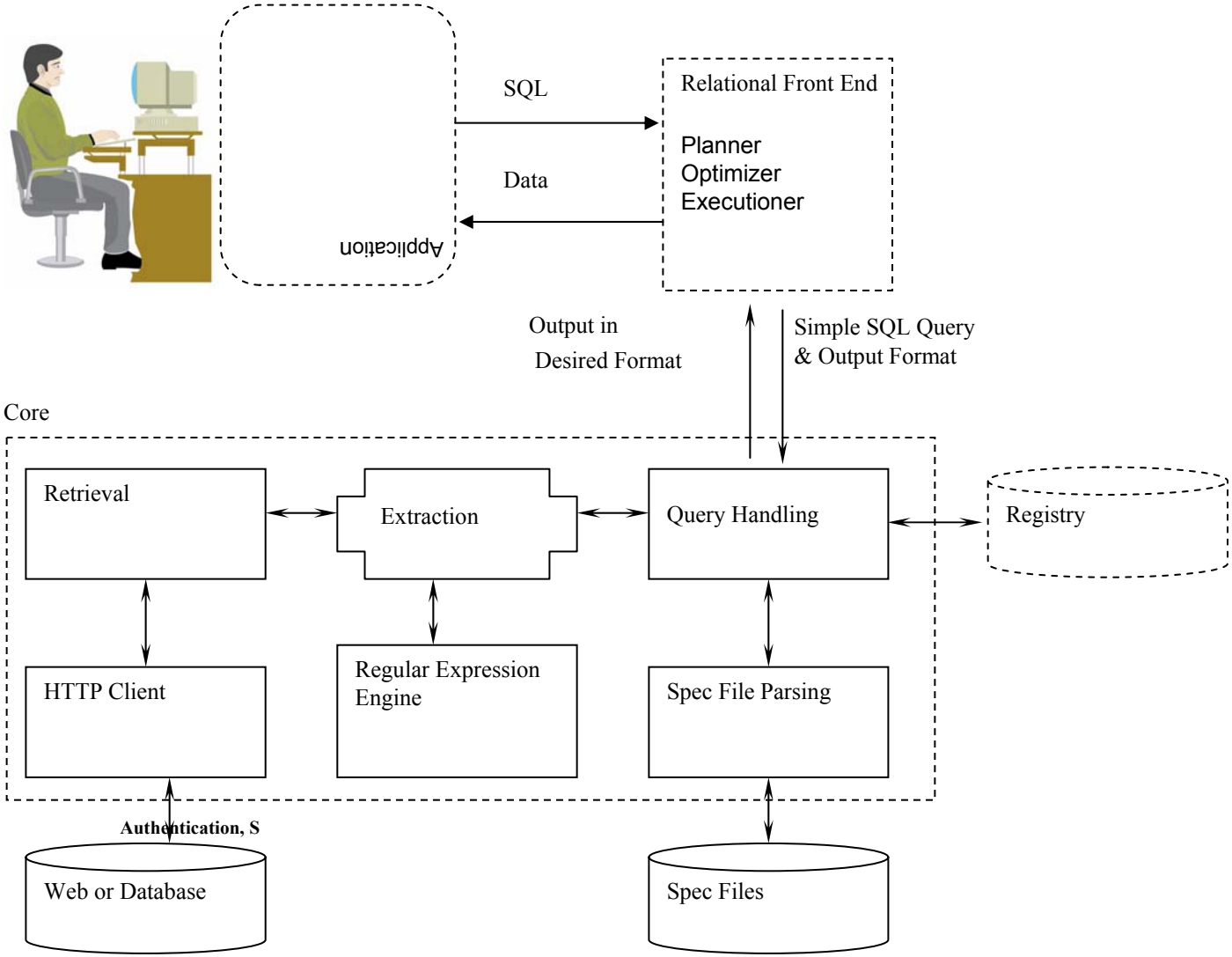
New Use of the Web: Program intermediaries Focus: **Productivity**

MIT Sloan COntext INterchange (COIN) Project



Example Semi-structured Web data: Intel SEC filing

Cameleon Architecture



CIA Fact Book Example

Singapore

Introduction

[Top of Page]

Background: Founded as a British trading colony in 1819, Singapore joined Malaysia in 1963, but withdrew two years later and became independent. It subsequently became one of the world's most prosperous countries, with strong international trading links (its port is one of the world's busiest) and with per capita GDP above that of the leading nations of Western Europe.

Geography

[Top of Page]

Location: Southeastern Asia, islands between Malaysia and Indonesia

Geographic coordinates: 1 22 N, 103 48 E

Map references: Southeast Asia

Area:
total: 647.5 sq km
land: 637.5 sq km
water: 10 sq km

Area - comparative: slightly more than 3.5 times the size of Washington, DC

Land boundaries: 0 km

Coastline: 193 km

Maritime claims:
exclusive fishing zone: within and beyond territorial sea, as defined in treaties and practice
territorial sea: 3 nm

Climate: tropical; hot, humid, rainy; no pronounced rainy or dry seasons; thunderstorms occur on 40% of all days (67% of days in April)

Terrain: lowland; gently undulating central plateau contains water catchment area and nature preserve

Elevation extremes:
lowest point: Singapore Strait 0 m
highest point: Bukit Timah 166 m

http://www.odci.gov/cia/publications/factbook/geos/sn.html

CAMELEON **QUERY:**

Select capital, location, coordinates, totalarea, climate, population, GDP from cia where Country="Singapore"

CAMELEON **RESULTS:**

Record 1

CAPITAL

Singapore

LOCATION

Southeastern Asia, islands between Malaysia and Indonesia

COORDINATES

1 22 N, 103 48 E

TOTALAREA

647.5 sq km

CLIMATE

tropical; hot, humid, rainy; no pronounced rainy or dry seasons; thunderstorms occur on 40% of all days (67% of days in April)

POPULATION

4,151,264 (July 2000 est.)

GDP

$98 billion (1999 est.)

Spec file for CIA Fact Book (partial)

```
#Relation=cia
#Source=http://www.odci.gov/cia/publications/factbook/country.html
#Attribute=Link#String
  #Begin=Top\s*of\s*Page
  #Pattern=<LI><FONT SIZE=-1><a href="([^"]*)">#Country#</a></font>
  #End=</[Bb][oO][dD][yY]>
#Source=http://www.odci.gov/cia/publications/factbook/#Link#
#Attribute=Telephone#String
  #Begin=Telephones:
  #Pattern=</b>\s*([\0-\377]*?)\s*<p>
  #End=Telephone system:
#Attribute=Background#String
  #Begin=Background:
  #Pattern=</b>\s*([\0-\377]*?)\s*<
  #End=Location:
#Attribute=Location#String
  #Begin=Location:
  #Pattern=</b>\s*([\0-\377]*?)\s*<p>
  #End=Geographic\s*coordinates:
```

. . .

Regular Expressions Used in Spec Files

* Match 0 or more times (greedy).

- ***?** Match 0 or more times (non-greedy).
- **+** Match 1 or more times (greedy).
- **?** Match 0 or 1 time (greedy).

Greedy quantifiers such as * matches as much as possible, whereas non-greedy quantifiers stop at the minimum match. Example:

 hello lovely   world

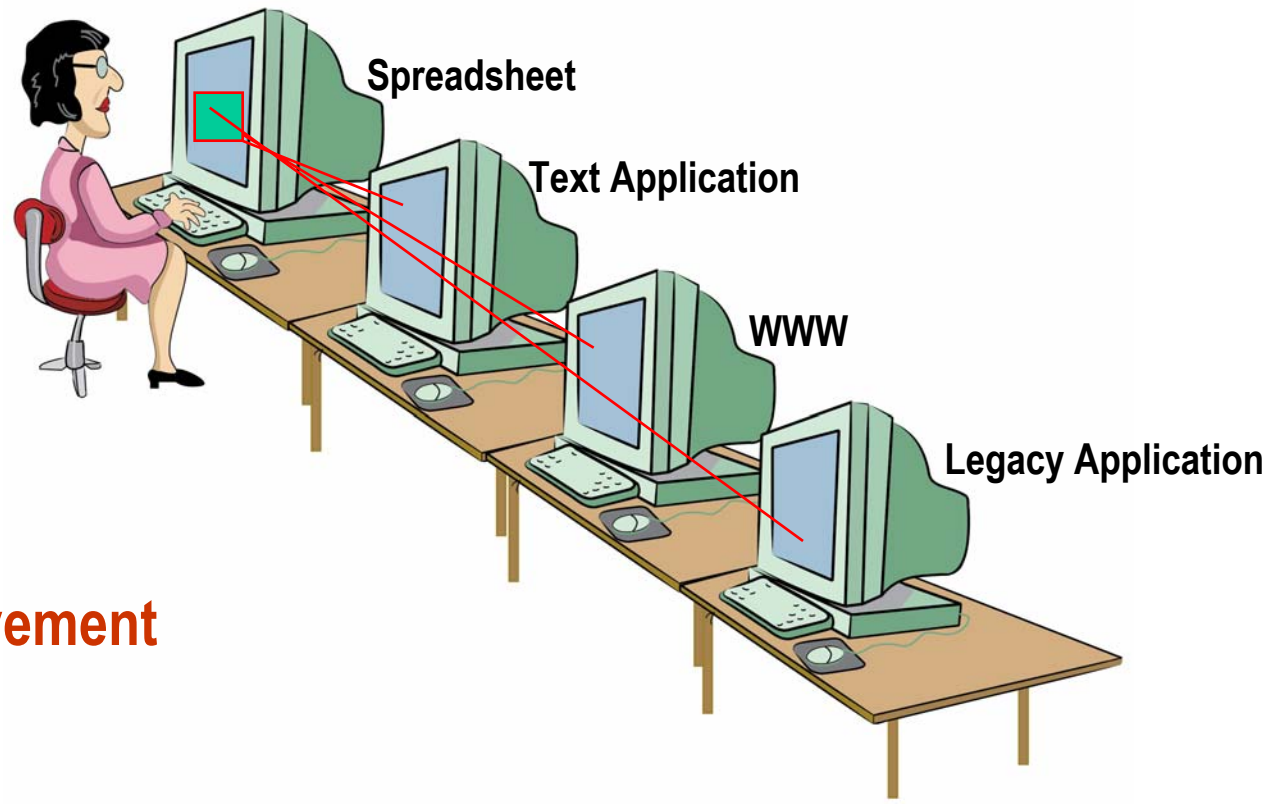
(.*) would match 'hello lovely   world' whereas

(.*?) would match 'hello' and 'world'

- **.** matches everything except \n
- **[\0-\377]** matches everything
- **^** matches the beginning of a string or line
- **[^ a character]** matches everything except the specified character.
  For instance [^<] matches anything but <
- **$** matches the end of a string or line
- **\s** matches a whitespace character
- **\S** matches a non-whitespace character
- **\d** matches a digit
- Expressions within parentheses are saved.

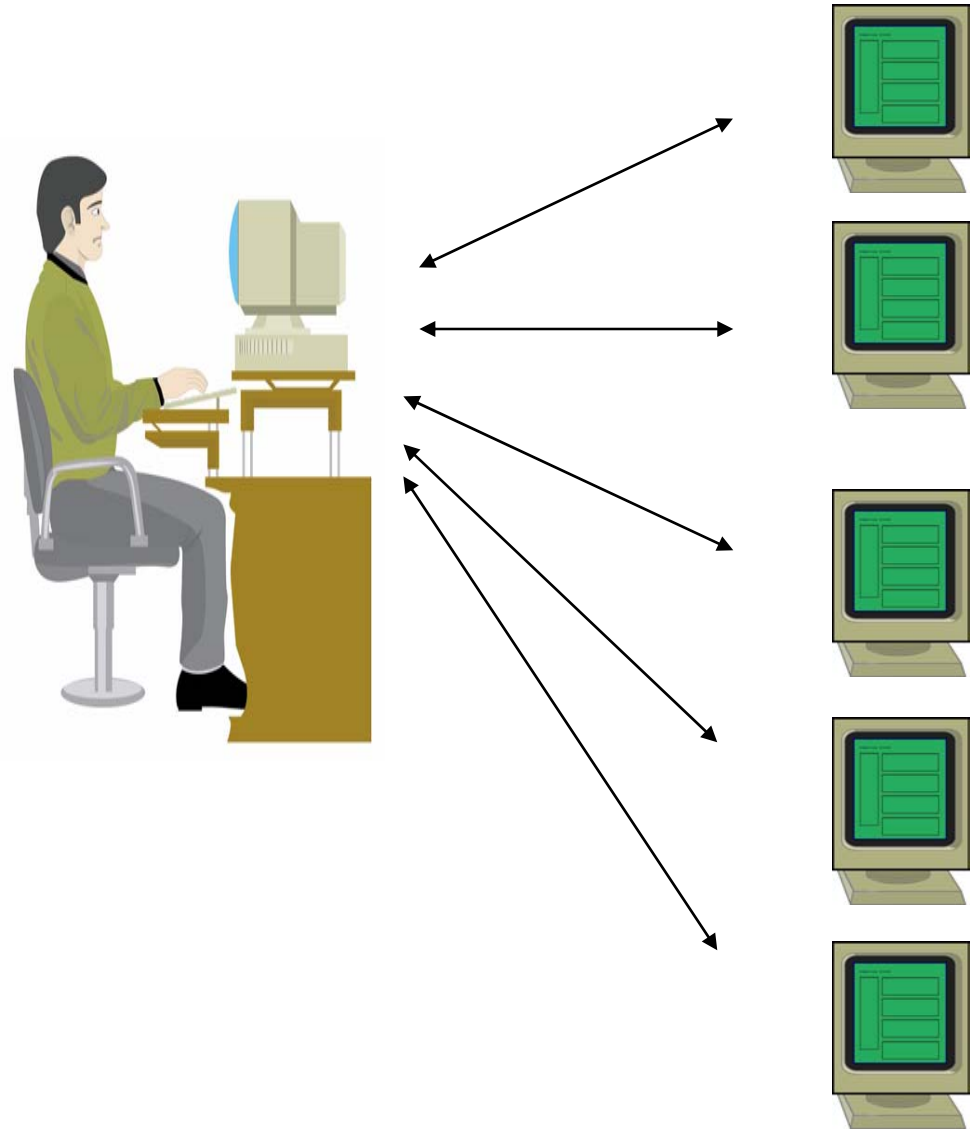
Sample Application

Research Analyst
or
Trader



Manual Data Movement

Providing Integrated Data andAnalysis



Equity prices - TIBCO
Real-time feed

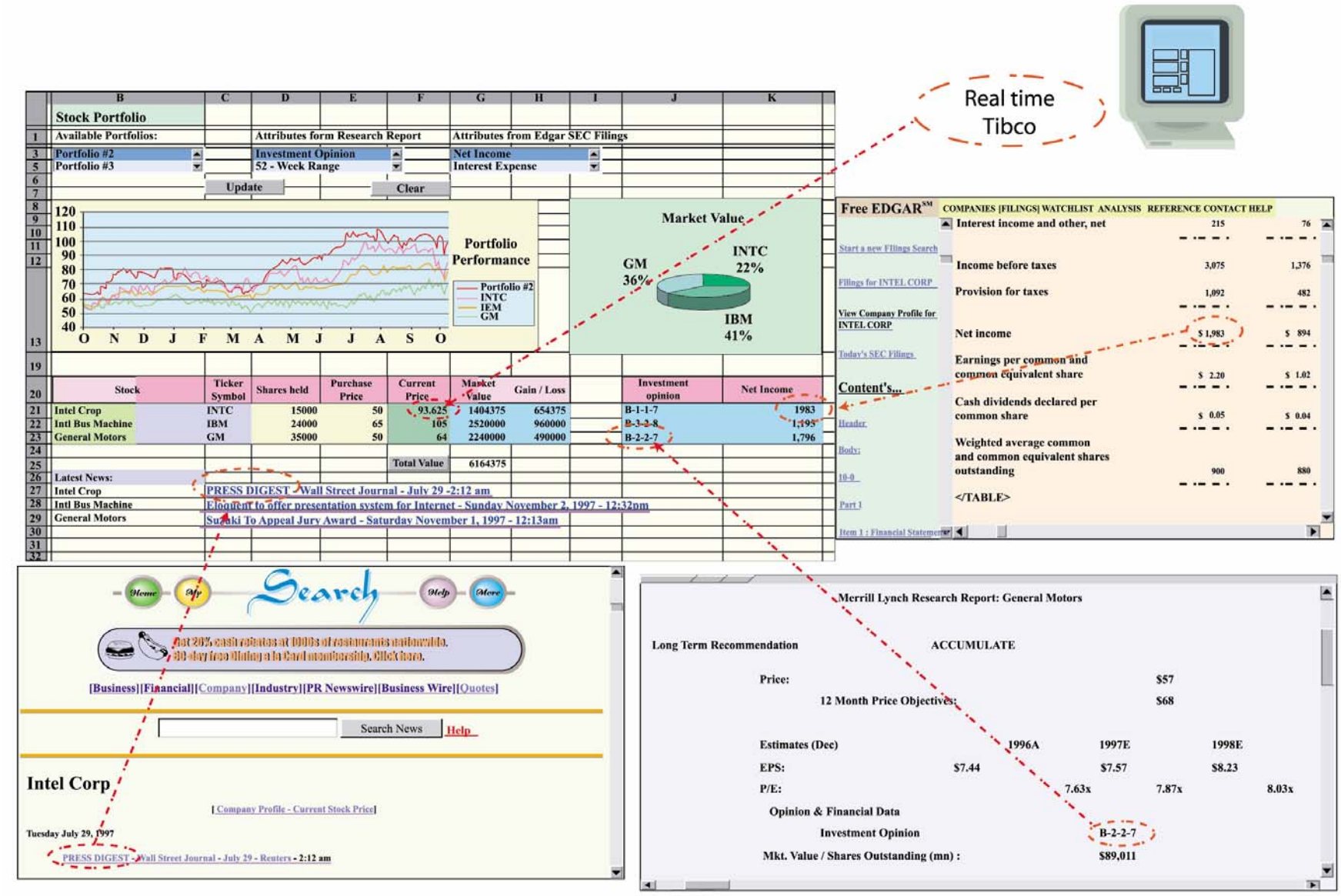
SEC Filings - EDGAR
Web based - Internet

News - Reuters, Newswire and Business wire
Web based - Internet

Merrill Research Reports
Text based - Merrill Intranet

Market updates - Merrill Lynch home page
Web based - Internet

Spreadsheet Interface



XML – The Silver Bullet ?

- XML is (according to press reports …)
  - "HTML on steroids" ?
  - "a Rosetta Stone" ?
  - "a universal way to translate data" ?
  - "a miraculous way to" … information integration ?
  - "a silver bullet" ?

XML What is it?

- XML - EXtensible Markup Language
- Meta language for defining a markup language
- Based on SGML - Standard Generalized Markup Language
- Data model for syntax for structuring data
- Can define tags at will
- Can nest document structures to arbitrary levels of complexity
- Can use Document Type Definition (DTD)
- Many other members of “family”:
  - XSL, XSLT, XLL, XML-Query, etc.

XML Does help create structured Web pages

Feature	HTML	XML
Extensibility	Fixed set of tags	Extensible set of tags
Tag purpose	Presentation	Content
Views	Single	Multiple (XSL)
Orientation	Documents	Documents + semi-structured data
Search	Keyword	Keywork plus Field-sensitive queries

Example: HTML Compared with XML

HTML *

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 3.2 Final//EN">
<html>
<head> . . .
<BODY topmargin=18 leftmargin=6 bgcolor="#ffffff" link="#0000ee" VLINK="#551A8B" ALINK="#ff0000">
<pre><font size=2>
                        Regular   Our
                        Price     Price
      Palm Pilot V      329.00    236.00    In stock
</font></pre>
<table cellpadding=0 cellspacing=0 border=0>
<tr><td align=left valign=middle width=455 nowrap height=20>
<tr><td align=left valign=top nowrap width=455>
<font size=1 face="helvetica,arial"> . . .
</BODY>
</HTML>
```

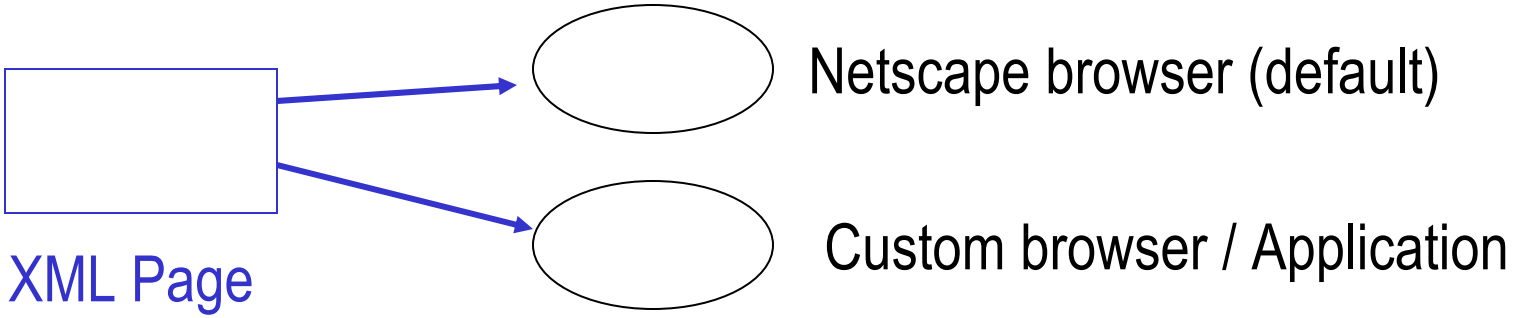
* The HTML is normally much messier, with lots more formatting details and <tr> and <td> tags for defining tables and tab positions.

XML

```
<XML>
<Product info>
    <Product> Palm Pilot V <\Product>
    <Regular price> 329.00 <\Regular price>
    <Our price> 236.00 <\Our price>
    <InStock> yes <\InStock>
<\Product info>
<\XML>
```

XML Why do we need it?

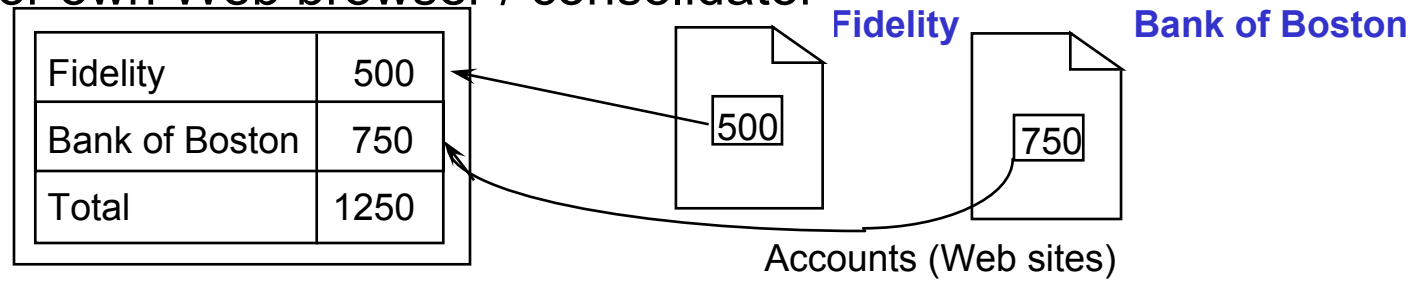
- W3C wanted to get out of the tag creation business
- Separate data from presentation
  - Use of a style sheet instead of “hardcoded” HTML formatting
  - Flexibility / scalability / extensibility



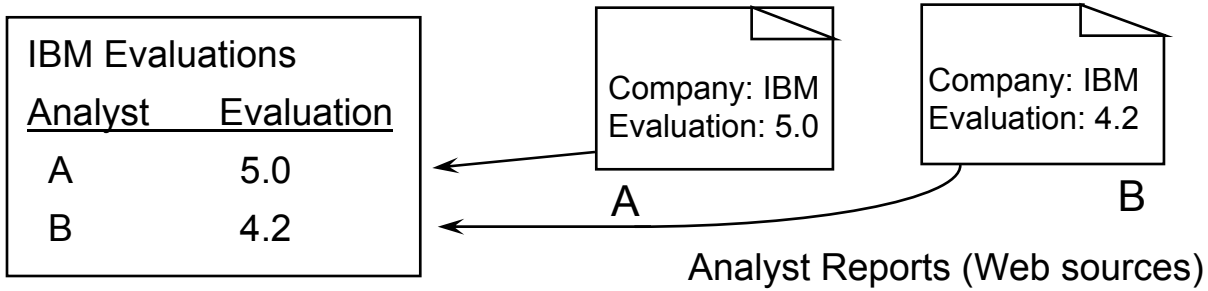
- Also important for Wireless Applications (WML/ XHTML)
- Human readability
- Computer processable
- Information interchange

Sample Applications of Semi-Structured Web Data

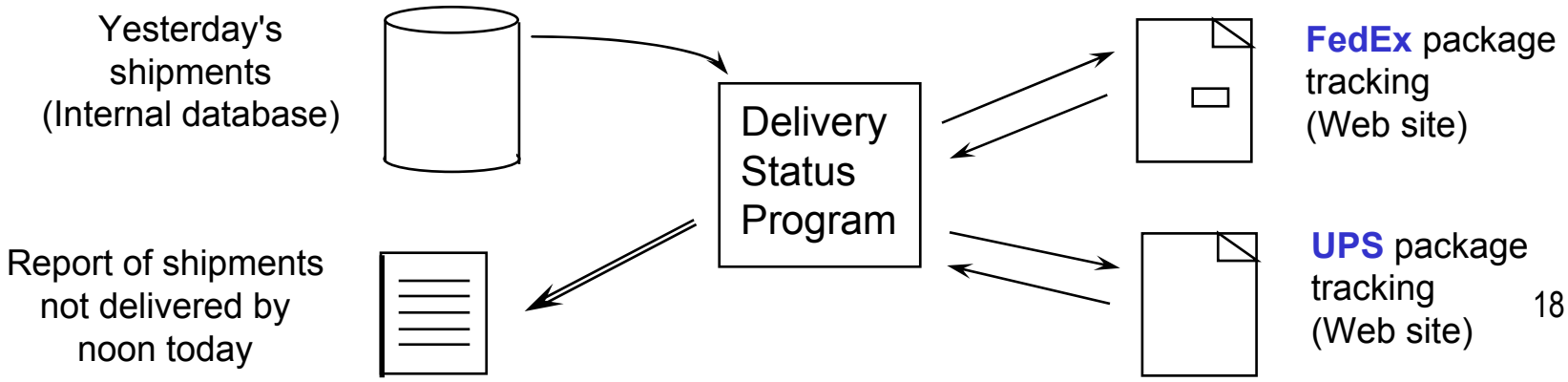
• Automate Extraction of data from specific Web sites into user tool, like Excel, or own Web browser / consolidator



• Automatically Select and Consolidate information across Web sites



• Integrate Internet / Intranet / Client Server networks for internal operations



XML . . . Multiple Standards

- What is so great about XML standards – is
  that there are so many . . .
- Should tag for catalog be called “price” or “cost” ?

- “The director of electronic trading at Credit Suisse First Boston and chairman of financial services XML working group is struggling with more than a dozen XML protocols … for financial trading applications.” (*ComputerWorld*, 9 July 2001)

XML – The Silver Bullet ?

- XML is not quite:
  - "a Rosetta Stone"
  - "a universal way to translate data"
  - "a miraculous way to" … information integration
  - "a silver bullet"

- It is a helpful tool toward information integration . . .
- Some background sources: w3c.org/XML and XML.org
- But much more is needed for information integration
  - Context Interchange and Semantic Web research are promising areas . . .

Summary

- **Tim Berners-Lee**, W3C Director:

  - "The Web is quickly becoming the world's fastest growing repository of data"

- In past: Primarily processed by humans.
- In future, must be processible by programs (agents for humans)
- Tools, such as MIT's Automatic Web Wrapper and W3C's XML, are providing these capabilities.

Work reported herein has been supported, in part, by the USA Advanced Research Projects Agency, Banco Santander Central Hispano, Citibank, Fleet Bank, First Logic, Merrill Lynch, PricewaterhouseCoopers, MIT Total Data Quality Management Program, MIT Center for eBusiness, Suruga Bank, and USAF/Rome Laboratory.