

# CS290i - Lecture 2

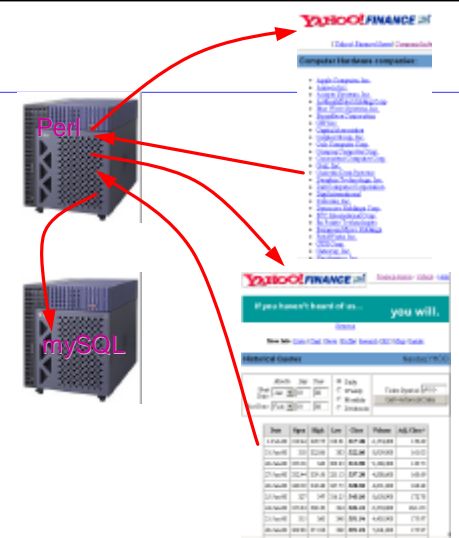
## Lift-off...

Scalable Internet Services and Systems, Spring 2001

Thorsten von Eicken  
Department of Computer Science  
University of California at Santa Barbara

## Outline

- † Crawling web sites
  - † Mechanics
  - † Robot rules
- † Perl
  - † Basics
  - † Tips & tricks
- † Yahoo finance
  - † Stock data
  - † Charts
- † SQL
  - † Basics



## Robots & Crawlers

### † Definition attempt:

- † Crawler: automated program that fetches and processes (a significant portion) of a web site's document hierarchy
- † Robot: automated program that fetches and processes specific pages of a web site

### † Uses:

- † Search engines
- † Comparison shopping
- † Performance measurements
- † Server monitoring
- † Others?

### † Project 1:

- † Fetch historical stock data from Yahoo

## Manual fetch: telnet

```
† [bugatti ~] telnet www.yahoo.com 80
Trying 204.71.200.67...
Connected to www.yahoo.akadns.net.
Escape character is '^]'.
GET / http/1.0
HTTP/1.0 200 OK
Content-Length: 18865
Content-Type: text/html
<html><head><title>Yahoo!</title><base
href=http://www.yahoo.com/><meta http-equiv="PICS-Label"
content="(PICS-1.1 "http://www.rsac.org/ratingsv01.html" 1
gen true for "http://www.yahoo.com" r (n 0 s 0 v 0 1
0))"></head><body onLoad="document.f.p.focus();">
<script><!--
...
<a href=r/ao>Advertising</a><p>Copyright &copy; 2001 Yahoo!
Inc. All rights reserved.</small><br><a href=r/pv>Privacy
Policy</a></form></center></body></html>
Connection closed by foreign host.
[bugatti ~]
```

## Manual fetch: wget

```
† [bugatti ~] wget -O- http://www.yahoo.com/
--20:14:46-- http://www.yahoo.com:80/
=> `-'
Connecting to www.yahoo.com:80... connected!
HTTP request sent, awaiting response... 200 OK
Length: 16,897 [text/html]

OK -><html><head><title>Yahoo!</title><base
href=http://www.yahoo.com/><meta httpequiv="PICS-Label"
content="(PICS-1.1 "http://www.rsac.org/ratingsv01.html" 1
gen true for "http://www.yahoo.com" r (n 0 s 0 v 0 l
0)) '></head><body onLoad="document.f.p.focus();">
<script><!--
...
<a href=r/hr>Jobs</a> -
<a href=r/ao>Advertising</a><p>Copyright &copy; 2001 Yahoo!
Inc. All rights reserved.</small><br><a href=r/pv>Privacy
Policy</a></form></center></body></html>
..... [100%]

20:14:46 (211.55 KB/s) - `-' saved [16897/16897]
[bugatti ~] wget -O- http://www.yahoo.com/ |& less
```

## Robot rules

† See <http://www.searchtools.com/robots/robots-txt.html>

† “Guidelines for Robot Writers”, M. Koster, 1993

1. Reconsider (Are you sure you really need a robot?)
2. Be Accountable
  - † Identify your Web Wanderer (HTTP User-agent field)
  - † Identify yourself (HTTP From field with your email address)
  - † Announce it to the target (tell the webmaster)
  - † Be informative (use the HTTP referer field)
  - † Be there
  - † Notify your authorities (sysadmin, network admin)
3. Test Locally
4. Walk, don't run

## Robot rules (cont.)

5. Use If-modified-since or HEAD where possible
6. Ask for what you want (HTTP Accept field)
7. Ask only for what you want (check suffixes of links)
8. Check the results
9. Don't Loop or Repeat (keep list of visited URLs, handle DNS aliases)
10. Run at opportune times (e.g. their night)
11. Don't run it often
12. Don't try queries
13. Log
14. Be interactive (start/pause/checkpoint/stop robot)

## Robots.txt

† A Standard for Robot Exclusion

- † Text file, URI: `/robots.txt`
- † Lines of the form:  
`<field>:<optionalspace><value><optionalspace>`
- † Comments start with #
- † Disallow record: URI prefix that should not be fetched, empty=>allow all
- † User-Agent field: following Disallow records apply to specific robot
- † Example:  
# robots.txt for http://www.example.com/  
User-agent: \*  
Disallow: /cyberworld/map/  
User-agent: cybermapper  
Disallow:

## Robot/crawler limitations

- † Unlinked pages (plain text links)
- † Authentication, cookies
- † Parameters/dynamic pages
- † Browser/host type
- † Document types
- † Others?
  - ‡ Q: how about your favorite search engine?

9

## Perl

- † Interpreted high-level programming language
- † Developed by Larry Wall
- † According to Larry, he included in Perl all the cool features found in other languages and left out those features that weren't so cool

### † The name:

- † Practical Extraction and Report Language
- † Pathologically Eclectic Rubbish Lister

### † On [bugatti.cs.ucsb.edu](http://bugatti.cs.ucsb.edu)

- † Use `/usr/local/bin/perl` !!!  
(or else you won't have the modules you need)

10

## Perl motto

- † The Perl motto is "There's more than one way to do it."  
Divining how many more is left as an exercise...
- † Perl's learning curve is therefore shallow (easy to learn) and long (there's a whole lot you can do if you really want)
- † The first of the four Perl Paradoxes:  
Perl programs are easy to write but not always easy to read
- † For example, the following lines are equivalent!
  - ‡ `if ($x == 0) {$y = 10;} else {$y = 20;}`
  - ‡ `$y = $x==0 ? 10 : 20;`
  - ‡ `$y = 20; $y = 10 if $x==0;`
  - ‡ `unless ($x == 0) {$y=20} else {$y=10}`
  - ‡ `if ($x) {$y=20} else {$y=10}`
  - ‡ `$y = (10,20)[$x != 0];`

11

## Perl documentation

### † Man

- † `Man perl` => list of man pages
- † `Perldoc` => same as man, without other stuff

### † Web pages

- † [www.perl.com](http://www.perl.com) (O'Reilly)
- † [www.cpan.org](http://www.cpan.org) (Comprehensive Perl Archive Network)

### † Modules

- † `HTTP::Request`
- † `HTTP::Response`
- † `LWP::UserAgent`
- † And more...

12

## Example: extract HREFs

### † Extract all HREFs out of www.yahoo.com

```
† [bugatti ~] cat test.pl
#!/usr/local/bin/perl -w
use strict;
require LWP::UserAgent;
require HTTP::Request;

my $ua = LWP::UserAgent->new;
my $request = HTTP::Request->new(GET =>
'http://www.yahoo.com');
my $response = $ua->request($request);
print "Response code: ", $response->code, "\n";
my @response = split /\r?\n/ , $response->as_string;
for(my $i=0; $i<5; $i++) {
    print $response[$i], "\n";
}
print "---\n";
my @hrefs = $response->as_string =~ /href="?(^">+);/g;
foreach my $h (@hrefs) {
    print "$h\n";
}
[bugatti ~]
```

13

## Extract HREFs (cont.)

```
† [bugatti ~] ./test.pl |& head -20
Response code: 200
HTTP/1.0 200 OK
Content-Base: http://www.yahoo.com/
Content-Length: 16897
Content-Type: text/html
Client-Date: Thu, 05 Apr 2001 06:33:05 GMT
----
http://www.yahoo.com/
r/al
r/pl
r/ml
r/wl
r/il
r/hw
/homet/?http://taxes.yahoo.com
http://messenger.yahoo.com
/homet/?http://baseball.fantasysports.yahoo.com/baseball/
r/so
/homet/?http://travel.yahoo.com
r/a2
[bugatti ~]
```

14

## Perl Modules

### † Find them on CPAN

- † Installation:
  - † wget www.cpan.org/.../foo.tar.gz
  - † gunzip <foo.tar.gz | tar xfv -; cd foo
  - † perl Makefile.PL PREFIX=/home/tve/perl
  - † make;make test; make install
  - † use lib "/home/tve/perl/lib/site\_perl";
- † Database module
  - † use DBI;
- † Request object manipulation, perform remote access
  - † use HTTP::Request;
  - † se HTTP::Request::Common;
  - † use LWP::UserAgent;
- † URI manipulation and HTML conversions
  - † use URI::URL;
  - † use HTML::Entities;

15

## Yahoo finance

### † Project 1

- † Fetch list of computer hardware companies
  - † <http://biz.yahoo.com/p/tech-o-cmptrs.html>
- † Fetch information about each company
- † Fetch 1 year of historical stock data for each company
  - † Day-by-day data is only available for one year
- † For each stock/day/company:
  - † Open price
  - † High price
  - † Low price
  - † Close price
  - † Any splits



16

# Companies

YAHOO! FINANCE

What are you waiting for  
Go online with YAHOO! Tax Center

Computer Hardware companies

- Apple Computer, Inc.
- Ariston, Inc.
- Avaya Systems, Inc.
- Avaya Data Networking
- Blue Win Systems, Inc.
- Broadcom Corporation
- CDI, Inc.
- Digital Equipment
- Desk Computer Corp.
- Empire Computer Corp.
- Computer Concepts Corp.
- Genie, Inc.
- Computer Data Systems
- Dynalco Technology, Inc.
- Dell Computer Corporation
- Dig International
- Edutronics, Inc.
- Dynalco Technology Corp.

Statistics at a Glance - Microsoft (MSFT)

| Price and Volume | Price | Volume | 52-Week High | 52-Week Low | 52-Week Range |
|------------------|-------|--------|--------------|-------------|---------------|
| MSFT             | 31.37 | 11,352 | 40.00        | 20.00       | 20.00 - 40.00 |

Stock Performance

Shareholder Data

| Market Capitalization | Shares Outstanding | Book Value |
|-----------------------|--------------------|------------|
| \$11.1B               | 353.8M             | \$31.48    |

# Historical quotes

Historical Quotes

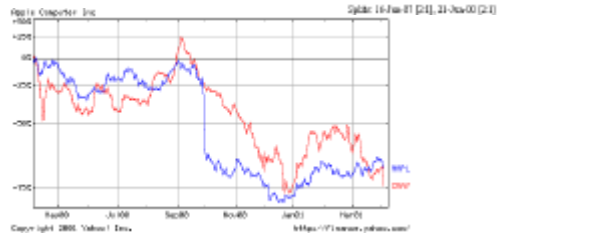
Start Date: Jan 01 00  
End Date: Jul 01 00

Frequency: Daily  
Time Period: 1000 Days

| Date      | Open                                 | High     | Low     | Close    | Volume    | Adj. Close |
|-----------|--------------------------------------|----------|---------|----------|-----------|------------|
| 30-Jun-00 | 32.8125                              | 34.9375  | 31.6875 | 32.3750  | 5,369,000 | 32.3750    |
| 29-Jun-00 | 32.8125                              | 33.8375  | 31.6875 | 31.2500  | 2,636,000 | 31.2500    |
| 28-Jun-00 | 33.3125                              | 33.3125  | 31.3800 | 31.7500  | 3,114,000 | 31.7500    |
| 27-Jun-00 | 33.7812                              | 33.7800  | 31.6000 | 31.7500  | 3,053,000 | 31.7500    |
| 26-Jun-00 | 33.5800                              | 33.7125  | 33.1250 | 33.2250  | 3,308,000 | 33.2250    |
| 23-Jun-00 | 33.7812                              | 34.6250  | 33.1250 | 31.6875  | 2,668,000 | 31.6875    |
| 22-Jun-00 | 33.7125                              | 37.6250  | 33.5000 | 33.7500  | 3,232,000 | 33.7500    |
| 21-Jun-00 | 33.5800                              | 36.9375  | 33.3125 | 33.6250  | 3,139,000 | 33.6250    |
| 21-Jun-00 | 1:1 Stock Split (before market open) |          |         |          |           |            |
| 20-Jun-00 | 98.5800                              | 100.8375 | 98.3125 | 101.2500 | 4,473,000 | 38.6250    |
| 19-Jun-00 | 99.3625                              | 99.3125  | 89.1250 | 89.6250  | 3,171,000 | 48.5125    |
| 16-Jun-00 | 99.5800                              | 99.3125  | 89.6250 | 91.8750  | 2,713,000 | 45.5938    |
| 15-Jun-00 | 98.2500                              | 99.3125  | 89      | 92.3750  | 2,219,000 | 46.1875    |
| 14-Jun-00 | 94.6875                              | 95.2500  | 98.1250 | 96.4375  | 2,477,000 | 43.2188    |

# Project 2

- † Company info
- † Stock price charts
- † Stock comparison charts



# SQL information

- † SQL: Structured Query Language
  - † pronounced "ess-cue-el" or "sequel"
- † Tutorials
  - † Philip Greenspun's "SQL for Web Nerds" (<http://www.arsdigita.com/books/sql/>)
  - † Mike Chapple's Introduction to SQL (<http://databases.about.com/compute/databases/library/weekly/aa020401a.htm>)
  - † James Hoffman's Introduction to SQL (<http://w3.one.net/~jhoffman/sqltut.htm>)
- † MySQL: an open source SQL database
  - † [www.mysql.com](http://www.mysql.com)
  - † An Introduction to MySQL ([http://www.mysql.com/articles/mysql\\_intro.html](http://www.mysql.com/articles/mysql_intro.html))

## SQL example

```
† [bugatti ~] mysql -u root -p tve
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 8 to server version: 3.22.26a

Type 'help' for help

mysql> create table companies (ticker varchar(8) primary
key, name varchar(8));
Query OK, 0 rows affected (0.07 sec)

mysql> insert into companies (ticker, name) values
('AAPL', 'Apple Computers Inc');
Query OK, 1 row affected (0.01 sec)

mysql> insert into companies (ticker, name) values
('CRAY', 'Cray Inc');
Query OK, 1 row affected (0.00 sec)
```

21

## SQL example (cont.)

```
† mysql> select * from companies;
+-----+-----+
| ticker | name   |
+-----+-----+
| AAPL   | Apple Co |
| CRAY   | Cray Inc |
+-----+-----+
2 rows in set (0.00 sec)

mysql> select name from companies where ticker='AAPL';
+-----+
| name   |
+-----+
| Apple Co |
+-----+
1 row in set (0.00 sec)

mysql> drop table companies;
Query OK, 0 rows affected (0.06 sec)

mysql> show tables;
Empty set (0.00 sec)

mysql>
```

22

## Perl DBI example

```
† Connect to the database
my $dsn = "DBI:mysql:$dbName";
$scrawler->{dbh} = DBI->connect($dsn, $user, $password)
    or return undef; # Failure while connecting to db

† Fill the books table
my $command = "INSERT INTO books (isbn,title,listprice) " .
    "VALUES (?, ?, ?)";
$dbh->do($command, undef,
    $self->{isbn},
    HTML::Entities::decode($self->{title} ),
    $self->{listPrice})
    or die $dbh->errstr;

† Disconnect from the mysql server
$self->{dbh}->disconnect if defined $dbh;
```

23

## Speed of fiber

- † Refractive index = speed of light in space / speed of light in medium
- † Refractive index of glass in fiber = 1.5 (approx)
- † Light in fiber = 2/3 Speed of light in space
- † Europe->US (LON-NYC) = 3500mi = 5600km
- † Time:  $5600 * 1.5 / 300000 = 28\text{ms}$  -> 56ms r/t
- † Measurements: LON~20ms, NYC~100ms

24