Seek and Ye shall Find

The continuum of computer “intelligence”

COS 116: 2/22/2007

Adam Finkelstein

Recap: Binary Representation

Powers of 2

2^0	2^1	2^2	2^3	2^4	2^5	2^6	2^7	2^8	2^9	2^{10}
1	2	4	8	16	32	64	128	256	512	1024

$2^{10} = 1024 \approx 10^3$

Fact: Every integer can be *uniquely* represented as a sum of powers of 2.

Ex: $25 = 16 + 8 + 1$

$= 1 \times 2^4 + 1 \times 2^3 + 0 \times 2^2 + 0 \times 2^1 + 1 \times 2^0$

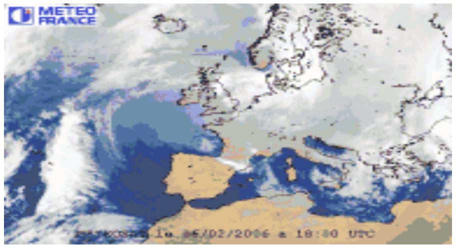
$[25]_2 = 11001$

Misconceptions about Computers

Just a calculator
on steroids

→

Weather Forecast



Just maintains large
amount of data

→

Airline Reservation System

Just does what the
programmer tells it

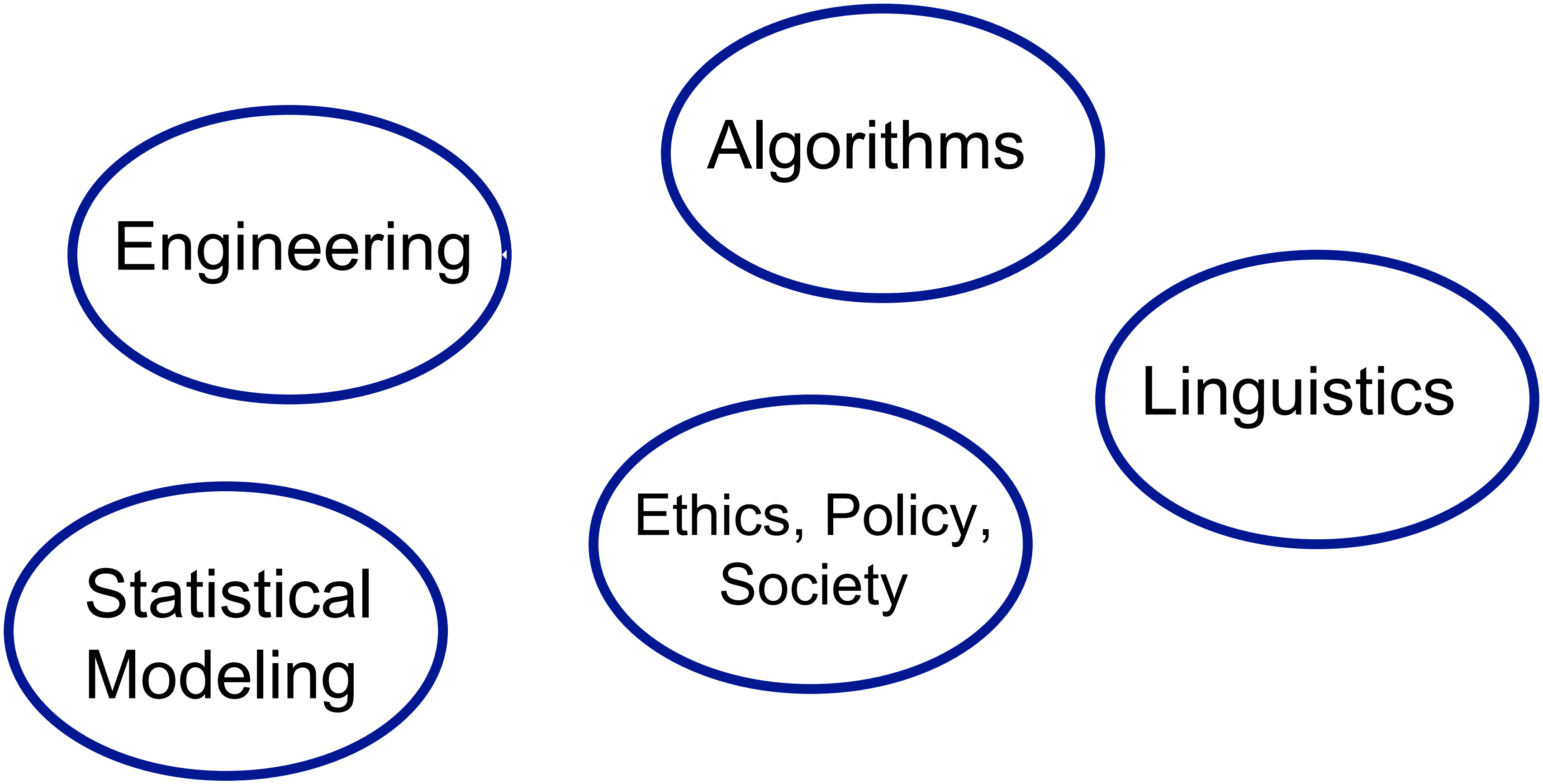
→

Yes, but …

Various meanings of

- Look up “Shirley Tilghman” in online phonebook.
- In consumer database, find “credit-worthy” consumers.
- Find web pages relevant to “computer music.”
- Among all cell phone conversations originating in Country X, identify suspicious ones.
- Search all religion and philosophy books of the world for meaning of life.

These are major scientific problems with many components



Algorithms

Engineering

Linguistics

Ethics, Policy, Society

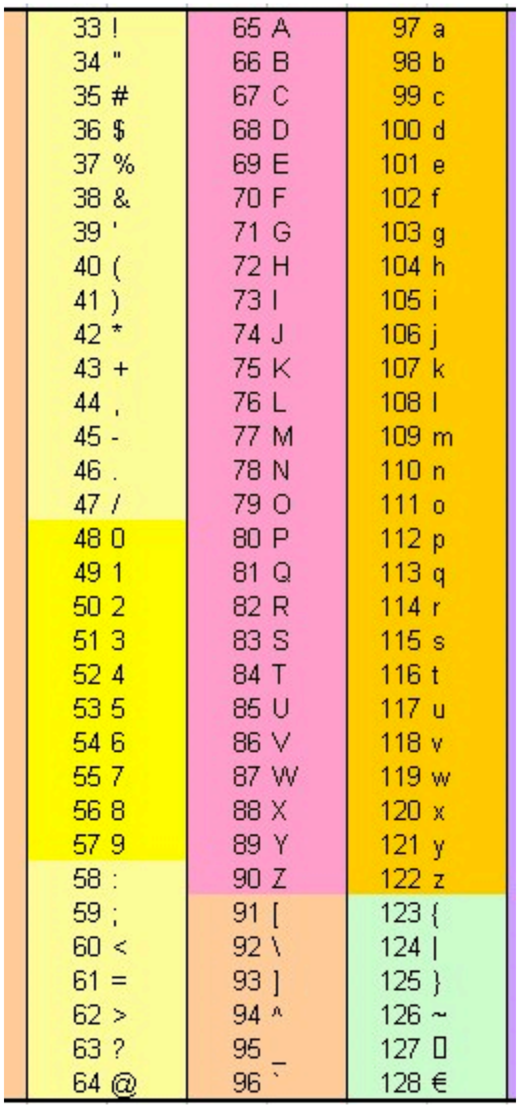
Statistical Modeling

Electronic Phonebook

- **ASCII:** Agreed-upon convention for representing letters with numbers
- Example:

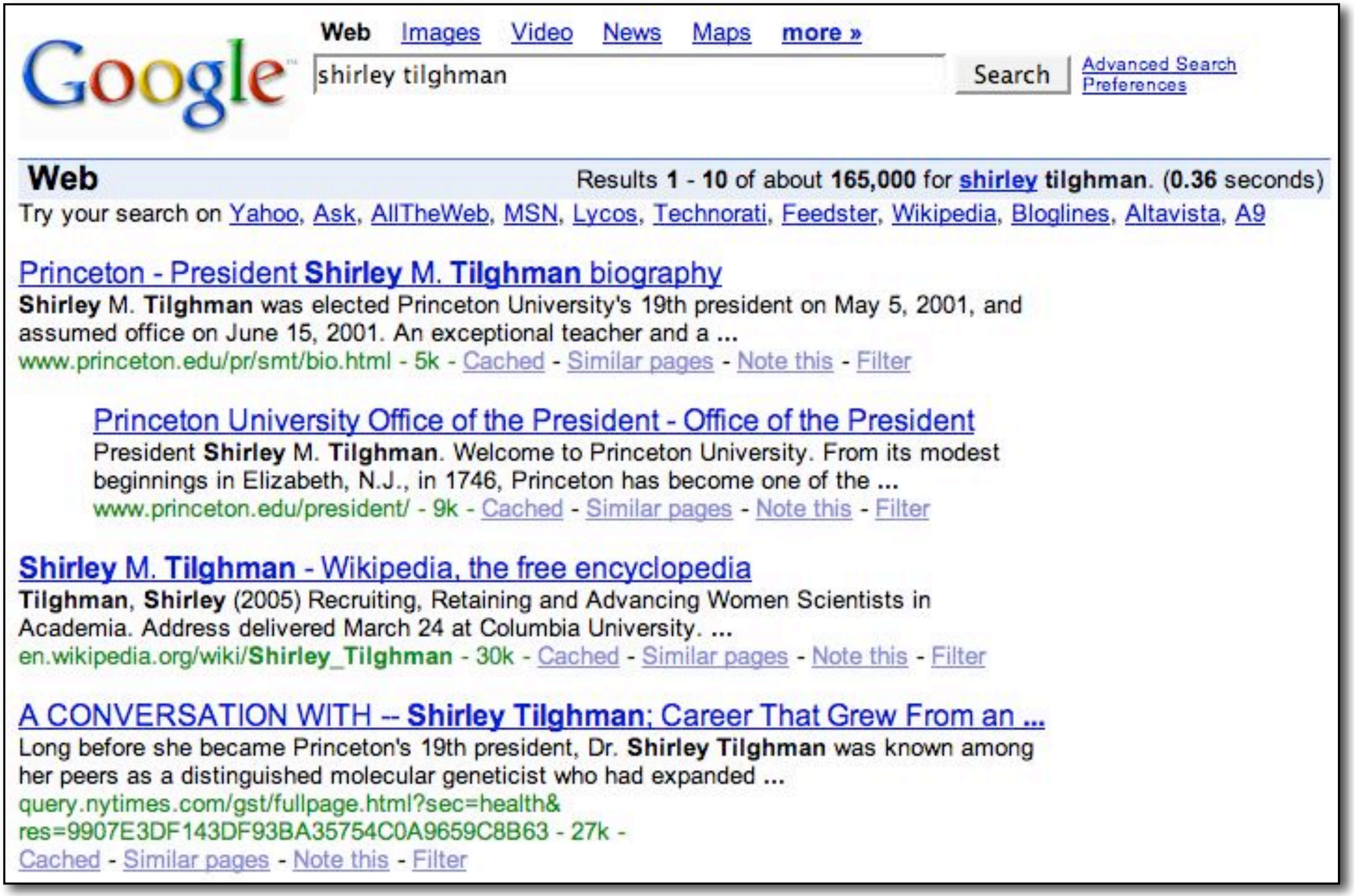
T	i	l	g	h	m	a	n	,	2	5	8	-	6	1	0	0
84	105	108	103	104	109	97	110	44	50	53	56	45	54	49	48	48

- Sorted Phonebook
  = sorted array of numbers
- Use binary search



33	!	65	A	97	a
34	"	66	B	98	b
35	#	67	C	99	c
36	$	68	D	100	d
37	%	69	E	101	e
38	&	70	F	102	f
39	'	71	G	103	g
40	(	72	H	104	h
41	)	73	I	105	i
42	*	74	J	106	j
43	+	75	K	107	k
44	,	76	L	108	l
45	-	77	M	109	m
46	.	78	N	110	n
47	/	79	O	111	o
48	0	80	P	112	p
49	1	81	Q	113	q
50	2	82	R	114	r
51	3	83	S	115	s
52	4	84	T	116	t
53	5	85	U	117	u
54	6	86	V	118	v
55	7	87	W	119	w
56	8	88	X	120	x
57	9	89	Y	121	y
58	:	90	Z	122	z
59	;	91	[	123	{
60	<	92	\	124	|
61	=	93	]	125	}
62	>	94	^	126	~
63	?	95	_	127	▯
64	@	96	`	128	€

Rest of the lecture: Web Search



Web Results **1** - **10** of about **165,000** for **shirley tilghman**. (**0.36** seconds)

Try your search on Yahoo, Ask, AllTheWeb, MSN, Lycos, Technorati, Feedster, Wikipedia, Bloglines, Altavista, A9

Princeton - President **Shirley** M. **Tilghman** biography
Shirley M. **Tilghman** was elected Princeton University's 19th president on May 5, 2001, and assumed office on June 15, 2001. An exceptional teacher and a ...
www.princeton.edu/pr/smt/bio.html - 5k - Cached - Similar pages - Note this - Filter

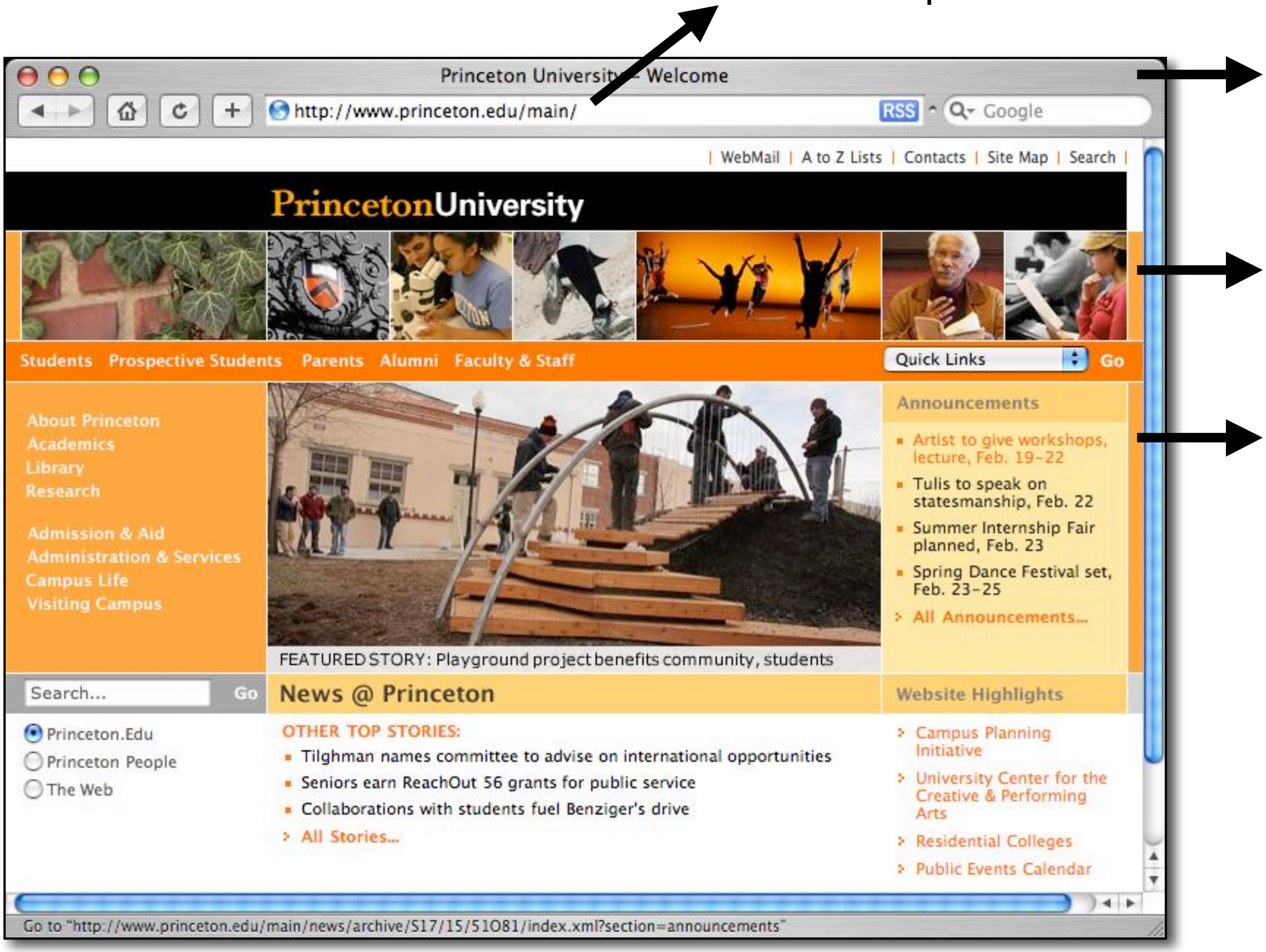
Princeton University Office of the President - Office of the President
President **Shirley** M. **Tilghman**. Welcome to Princeton University. From its modest beginnings in Elizabeth, N.J., in 1746, Princeton has become one of the ...
www.princeton.edu/president/ - 9k - Cached - Similar pages - Note this - Filter

Shirley M. **Tilghman** - Wikipedia, the free encyclopedia
Tilghman, **Shirley** (2005) Recruiting, Retaining and Advancing Women Scientists in Academia. Address delivered March 24 at Columbia University. ...
en.wikipedia.org/wiki/**Shirley**_**Tilghman** - 30k - Cached - Similar pages - Note this - Filter

A CONVERSATION WITH -- **Shirley Tilghman**; Career That Grew From an ...
Long before she became Princeton's 19th president, Dr. **Shirley Tilghman** was known among her peers as a distinguished molecular geneticist who had expanded ...
query.nytimes.com/gst/fullpage.html?sec=health&res=9907E3DF143DF93BA35754C0A9659C8B63 - 27k -
Cached - Similar pages - Note this - Filter

World Wide Web (simplified view)

URL: Unique address for each document



Princeton University - Welcome
http://www.princeton.edu/main/
RSS
Google
| WebMail | A to Z Lists | Contacts | Site Map | Search |
PrincetonUniversity
Students Prospective Students Parents Alumni Faculty & Staff
Quick Links
Go
About Princeton
Academics
Library
Research
Admission & Aid
Administration & Services
Campus Life
Visiting Campus
Search...
Go
○ Princeton.Edu
○ Princeton People
○ The Web
Announcements
Website Highlights
Campus Planning Initiative
University Center for the Creative & Performing Arts
Residential Colleges
Public Events Calendar

FEATURED STORY: Playground project benefits community, students

Announcements

- Artist to give workshops, lecture, Feb. 19-22
- Tulis to speak on statesmanship, Feb. 22
- Summer Internship Fair planned, Feb. 23
- Spring Dance Festival set, Feb. 23-25
- All Announcements...

News @ Princeton

OTHER TOP STORIES:

- Tilghman names committee to advise on international opportunities
- Seniors earn ReachOut 56 grants for public service
- Collaborations with students fuel Benziger's drive
- All Stories...

Go to "http://www.princeton.edu/main/news/archive/S17/15/51O81/index.xml?section=announcements"

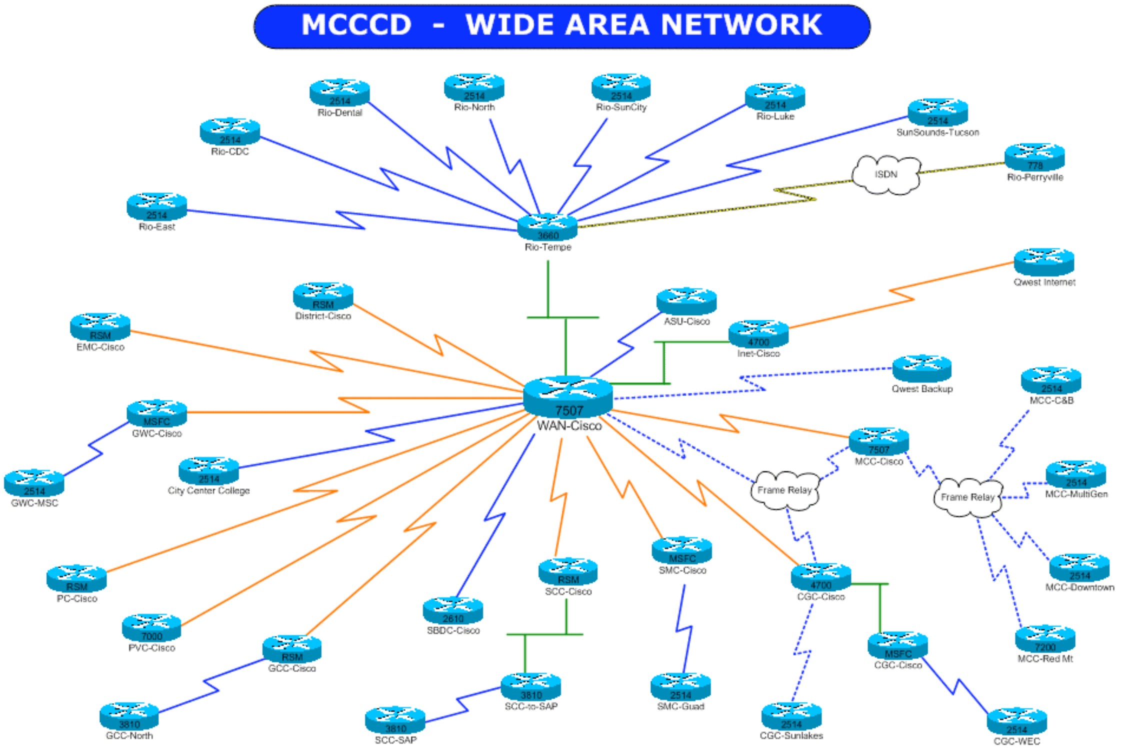
Browser

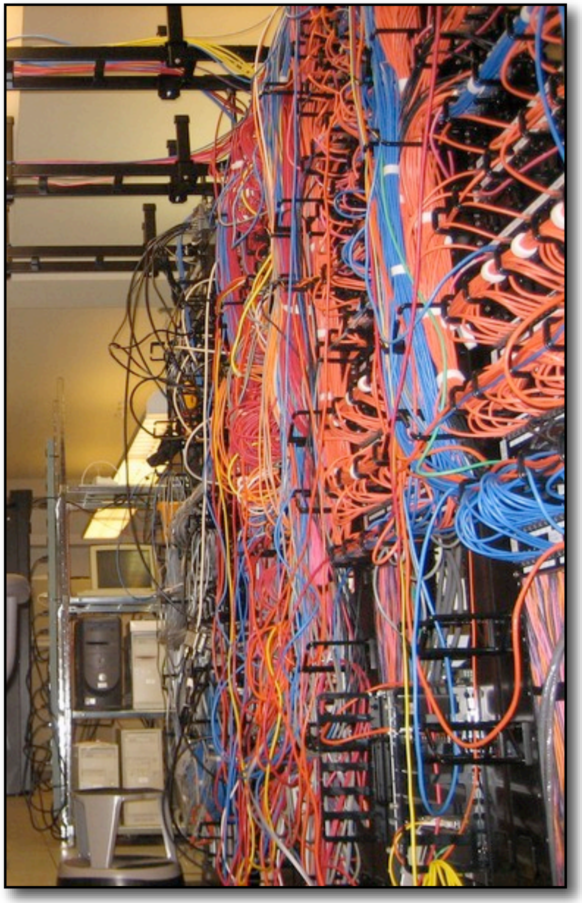
Web Page

Hyperlink

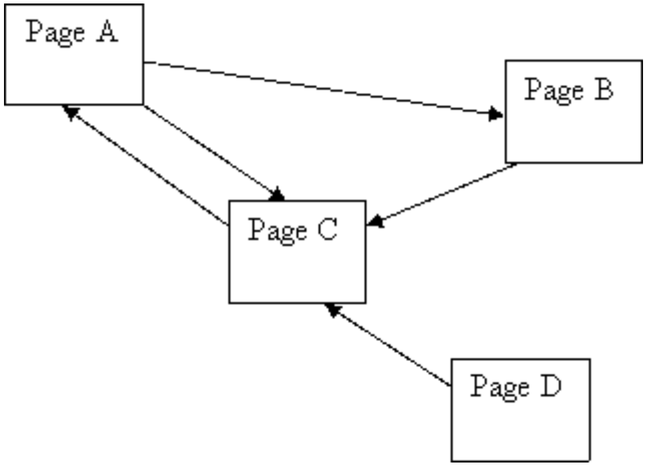
Future lecture:
Physical infrastructure of the Web

Routers, gateways, DNS, ...





Logical Structure of the Web

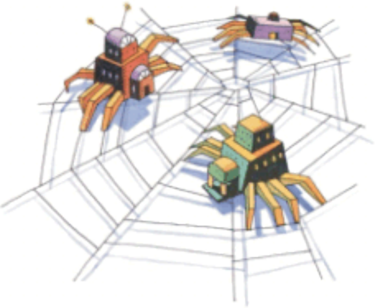


"Directed graph"

"edges" = link from one node to another

- **Important:** This logical structure is created by independent actions of 100s of millions of users

1st step for search engines: create snapshot of the web



- **Webcrawler:** "browser on autopilot"
- Maintains array of web pages it has seen
- 2 types of pages: "visited", "fully explored"
- Do forever

```
{
  Pick any webpage marked "visited" from array.
  Mark it "fully explored."
  Open all its linked pages in browser.
  Save them in array and mark them "visited."
}
```

Better: just the pages not "fully explored" yet.

First Web Crawler

From: bp@cs.washington.edu (Brian Pinkerton)
Newsgroups: comp.infosystems.announce
Subject: The WebCrawler Index: A content-based Web index
Date: 11 June 1994 21:33:42 GMT
Organization: University of Washington

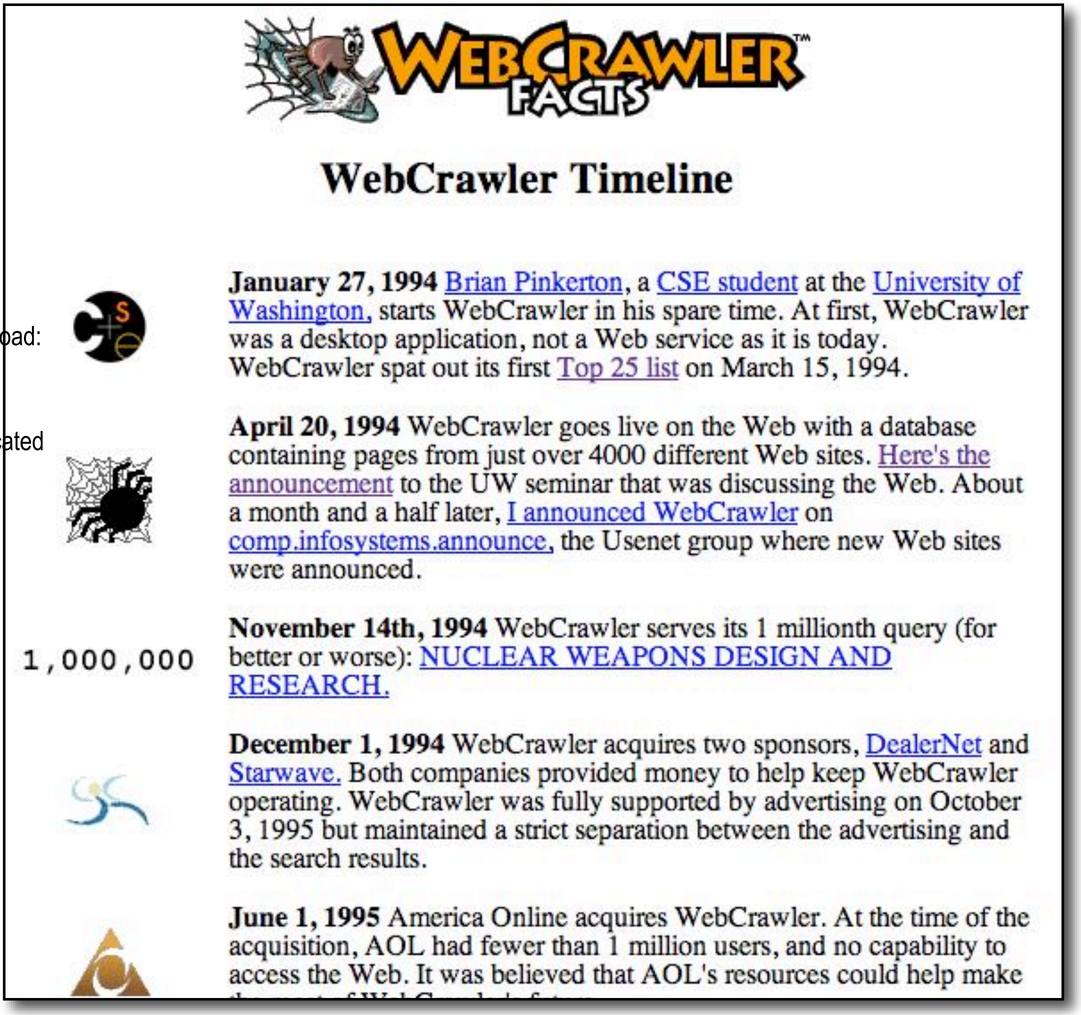
The WebCrawler Index is now available for searching! The index is broad: it contains information from as many different servers as possible. It's a great tool for locating several different starting points for exploring by hand. The current index is based on the contents of documents located on nearly 4000 servers, world-wide.

Check it out at:

http://www.biotech.washington.edu/WebCrawler/WebQuery.html

Other information is available from there, including a description of the WebCrawler (the robot itself), and a list of the 25 most frequently referenced sites on the Web.

Brian Pinkerton
Dept of Computer Science and Engineering
University of Washington



WebCrawler Timeline

January 27, 1994 Brian Pinkerton, a CSE student at the University of Washington, starts WebCrawler in his spare time. At first, WebCrawler was a desktop application, not a Web service as it is today. WebCrawler spat out its first Top 25 list on March 15, 1994.

April 20, 1994 WebCrawler goes live on the Web with a database containing pages from just over 4000 different Web sites. Here's the announcement to the UW seminar that was discussing the Web. About a month and a half later, I announced WebCrawler on comp.infosystems.announce, the Usenet group where new Web sites were announced.

1,000,000

November 14th, 1994 WebCrawler serves its 1 millionth query (for better or worse): NUCLEAR WEAPONS DESIGN AND RESEARCH.

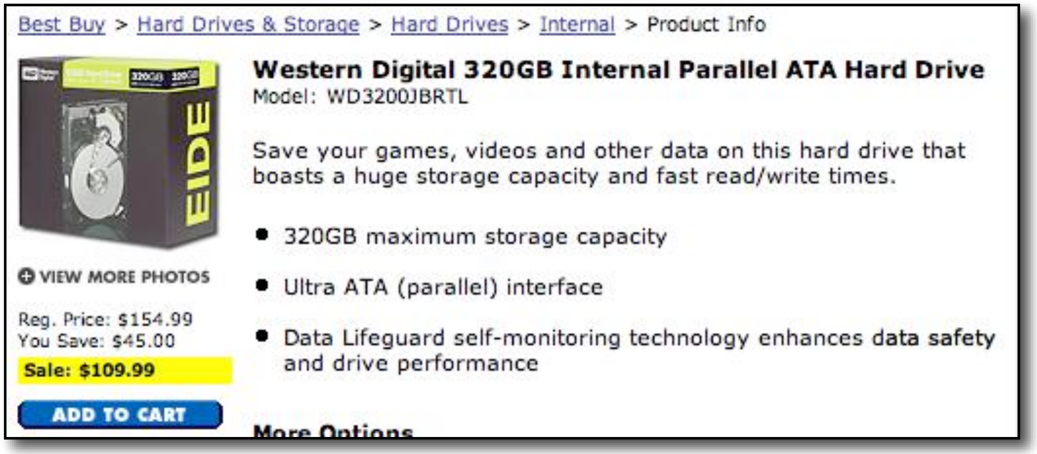
December 1, 1994 WebCrawler acquires two sponsors, DealerNet and Starwave. Both companies provided money to help keep WebCrawler operating. WebCrawler was fully supported by advertising on October 3, 1995 but maintained a strict separation between the advertising and the search results.

June 1, 1995 America Online acquires WebCrawler. At the time of the acquisition, AOL had fewer than 1 million users, and no capability to access the Web. It was believed that AOL's resources could help make

[http://thinkpink.com/bp/WebCrawler/History.html]

Still Feasible Today?

- About 15 billion web pages today.
- Say 10 kb (10,000 bytes) of data per page
- 15 X 10^{13} bytes to store the web
- ≈ 150, 000 Gb
- ≈ 500 hard disks
- ≈ $50,000 in ‘07



Best Buy > Hard Drives & Storage > Hard Drives > Internal > Product Info

Western Digital 320GB Internal Parallel ATA Hard Drive
Model: WD3200JBRTL

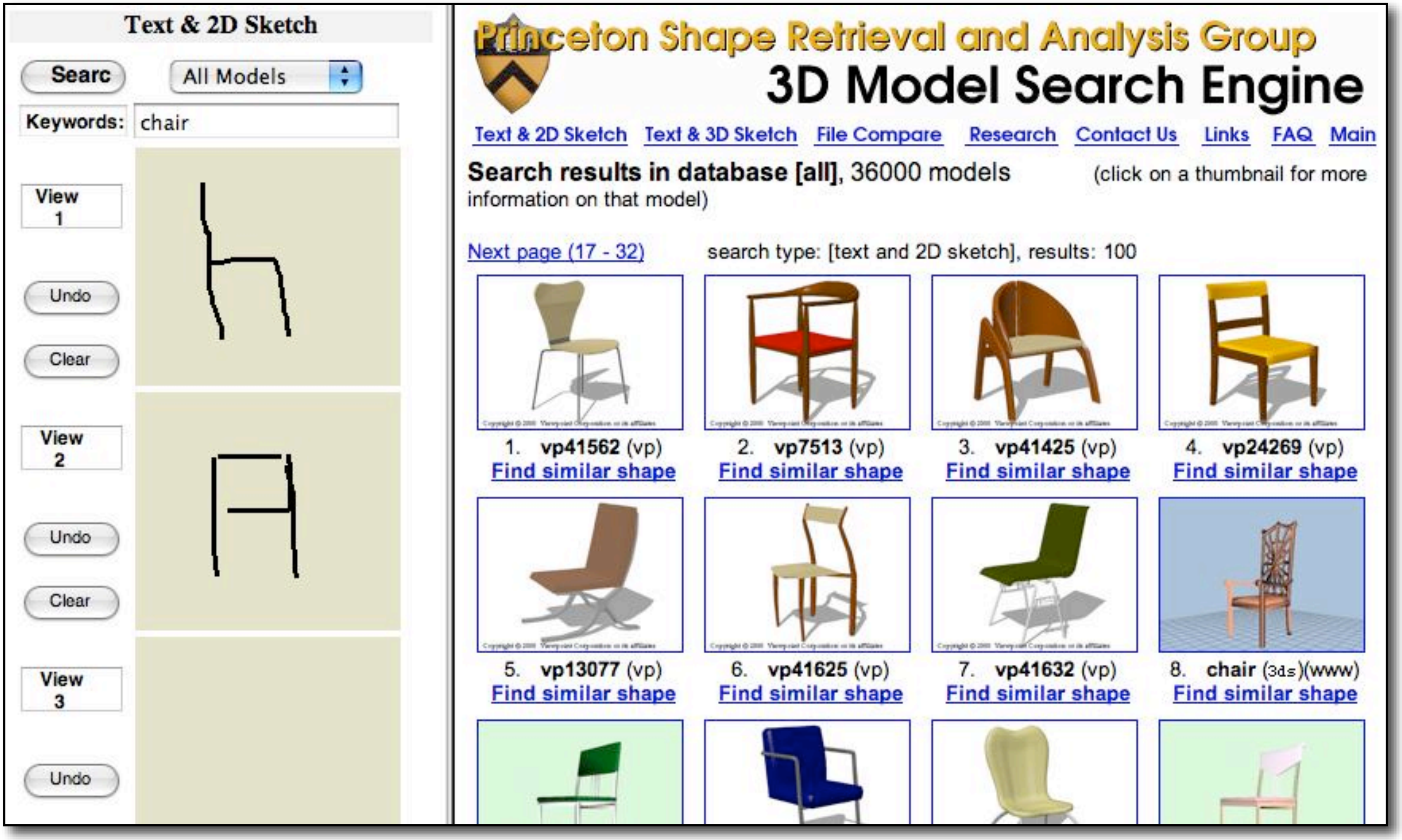
Save your games, videos and other data on this hard drive that boasts a huge storage capacity and fast read/write times.

- 320GB maximum storage capacity
- Ultra ATA (parallel) interface
- Data Lifeguard self-monitoring technology enhances data safety and drive performance

Reg. Price: $154.99
You Save: $45.00
Sale: $109.99

More Options

Princeton Shape Search Engine



Keywords: chair

Princeton Shape Retrieval and Analysis Group

3D Model Search Engine

Search results in database [all], 36000 models (click on a thumbnail for more information on that model)

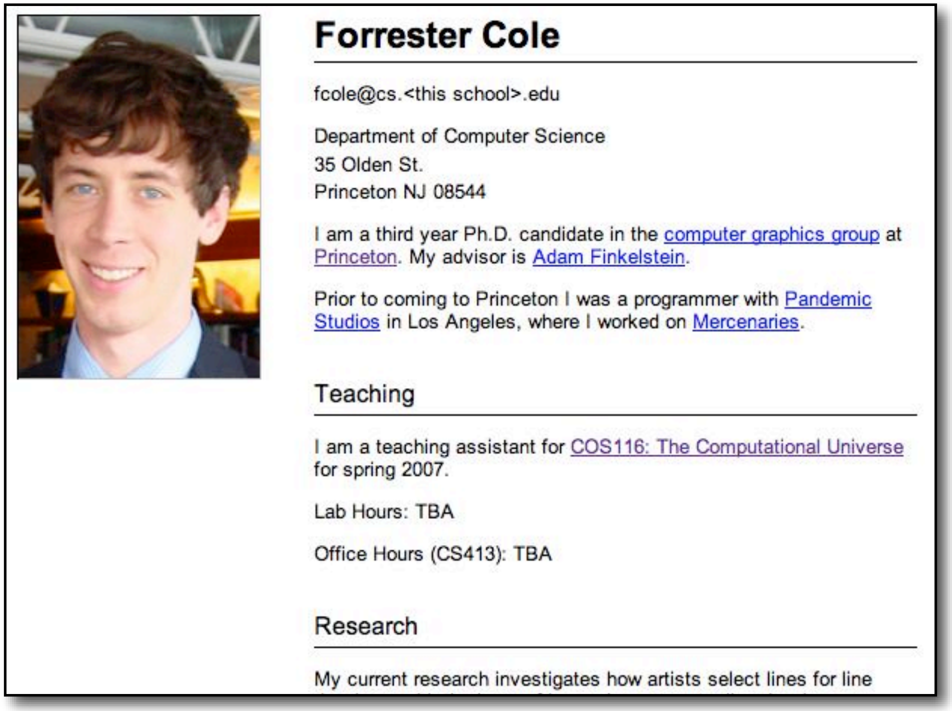
Next page (17 - 32) search type: [text and 2D sketch], results: 100

1. **vp41562** (vp) Find similar shape
2. **vp7513** (vp) Find similar shape
3. **vp41425** (vp) Find similar shape
4. **vp24269** (vp) Find similar shape
5. **vp13077** (vp) Find similar shape
6. **vp41625** (vp) Find similar shape
7. **vp41632** (vp) Find similar shape
8. **chair** (3ds)(www) Find similar shape

[http://shape.cs.princeton.edu/search.html]

Finding Forrester

How does Google
find Forrester Cole…?



Forrester Cole

fcole@cs.<this school>.edu

Department of Computer Science
35 Olden St.
Princeton NJ 08544

I am a third year Ph.D. candidate in the computer graphics group at Princeton. My advisor is Adam Finkelstein.

Prior to coming to Princeton I was a programmer with Pandemic Studios in Los Angeles, where I worked on Mercenaries.

Teaching

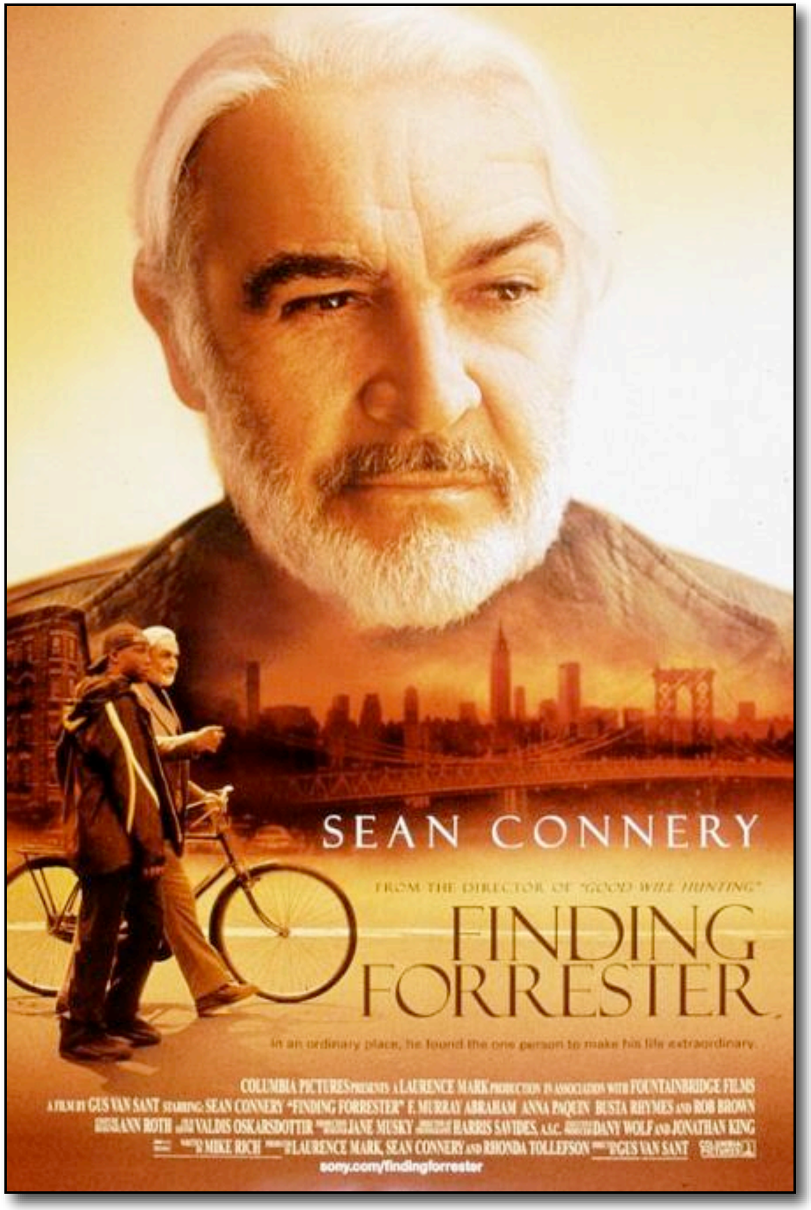
I am a teaching assistant for COS116: The Computational Universe for spring 2007.

Lab Hours: TBA

Office Hours (CS413): TBA

Research

My current research investigates how artists select lines for line


SEAN CONNERY
FROM THE DIRECTOR OF "GOOD WILL HUNTING"
FINDING FORRESTER
In an ordinary place, he found the one person to make his life extraordinary.
COLUMBIA PICTURES PRESENTS A LAURENCE MARK PRODUCTION IN ASSOCIATION WITH FOUNTAINBRIDGE FILMS
sony.com/findingforrester

Searching for “computer music”

Ideas?

- Identify all pages that contain “computer music”.
- Sort according to number of occurrences of “computer music” in the page.
- Human staff computes answers to all possible questions.

Some pitfalls

- “Spamming” by unscrupulous websites
- Synonymy (car, auto, vehicle …)
- Polysemy (jaguar: car or cat?)

Solution

IBM's CLEVER – 1996

Google's PAGERANK – 1997

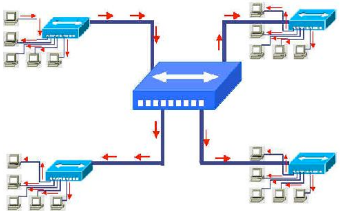
Take advantage of the link structure of the web

Web link confers "approval"

CLEVER

Authorities: Sites that are viewed “with respect” by many

- New York Times
- International Computer Music Association



Hubs: Clearinghouses of information

- “My favorite computer music links”

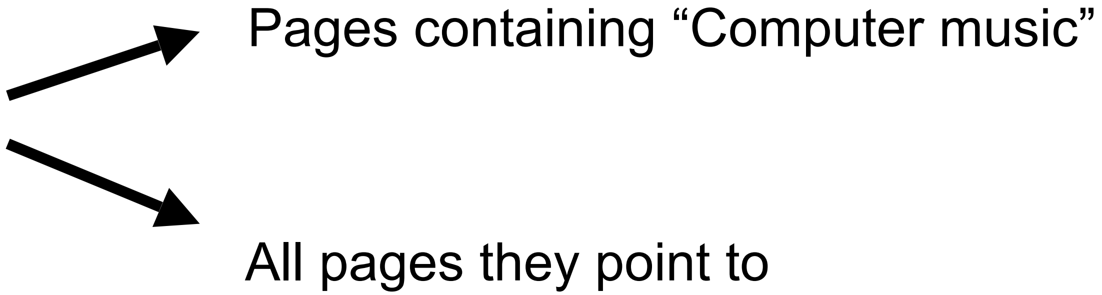
Typically Authorities point to hubs and hubs point to authorities

Circular Definition?

Circular Definition – *see* Definition, Circular

Breaking Circularity

- Iterative algorithm
- Start with



Pages containing “Computer music”

All pages they point to

- At every step each page has:
  - “Hub Score”
  - “Authority Score”

} Initially all 1

Score Calculation

- Do forever

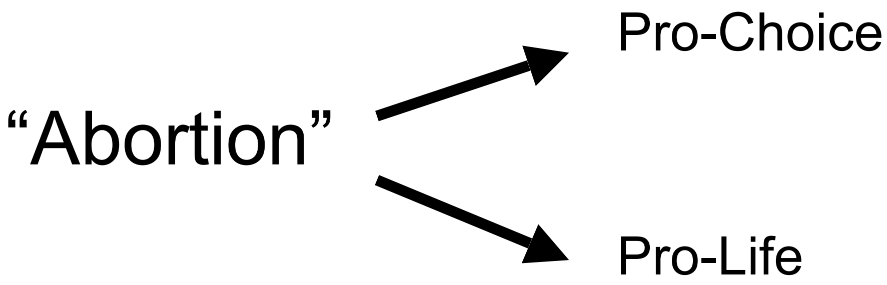
```
{
  Next Hub Score for page          ⇐ Sum of current Authority
                                     Scores of pages that link
                                     to it.

  Next Authority Score for page    ⇐ Sum of current Hub
                                     Scores of pages that link
                                     to it.
}
```

Fact The scores converge.

(Proof uses Linear Algebra, Eigenvalues)

- By Product – Algorithm reveals clusters

Example:



“Abortion”

Pro-Choice

Pro-Life

- Data Mining – Process of finding answers that are not in the data and must be inferred.

Example: “How is a person who shops at
Whole Foods & REI likely to vote?”

Concerns

From **users**:

- Privacy
- Privacy
- Privacy

From **Computer scientists**:

- Formalize privacy
- How to safeguard privacy while allowing legitimate computations

Next Time…

Digital Audio / Music (Perry Cook)