

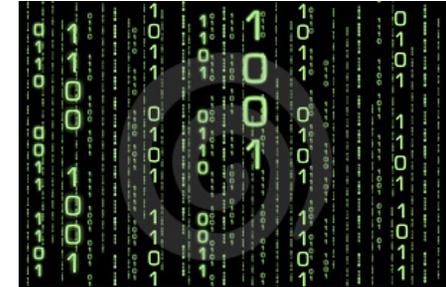
Seek and Ye shall Find

The continuum of computer “intelligence”

COS 116, Spring 2010

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.

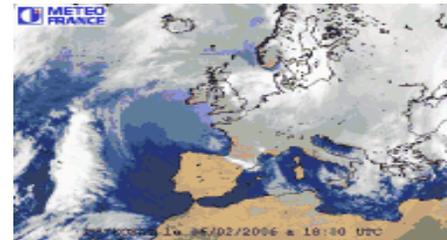
Ex: $25 = 16 + 8 + 1$
 $= 1 \times 2^4 + 1 \times 2^3 + 0 \times 2^2 + 0 \times 2^1 + 1 \times 2^0$
 $[25]_2 = 11001$

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 ...

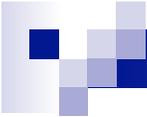
Various meanings of SEARCH



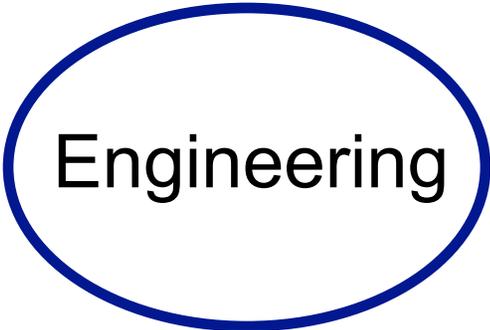
- Look up “Shirley Tilghman” in online phonebook.
- In consumer database, find “credit-worthy” consumers.
- Find web pages relevant to “computer music.”
- Among all cell phone conversations originating in Country X, identify suspicious ones.
- Search all religion and philosophy books of the world for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

“Data Mining”

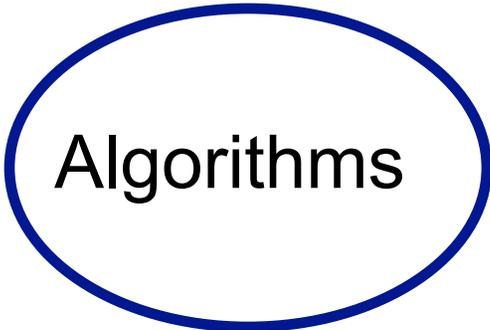
“Web Search”



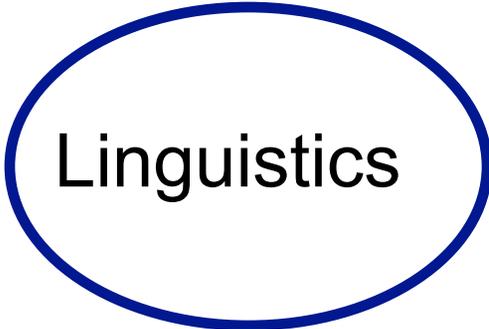
These are major scientific problems with many components



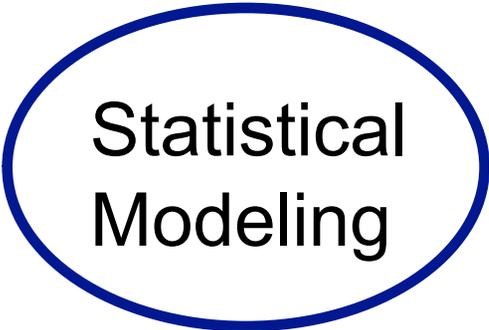
Engineering



Algorithms



Linguistics



Statistical
Modeling



Ethics, Policy,
Society



Discussion Time

How do you solve this task:

Sorted array of n numbers, find if it contains 58780

Binary search! First thing to check: “Is $A[n/2] < 58780$ ”?
(Whatever the answer, you halve the range.)

Question: What if the array of numbers is not sorted??

Looking up “Shirley Tilghman” in Electronic Phonebook

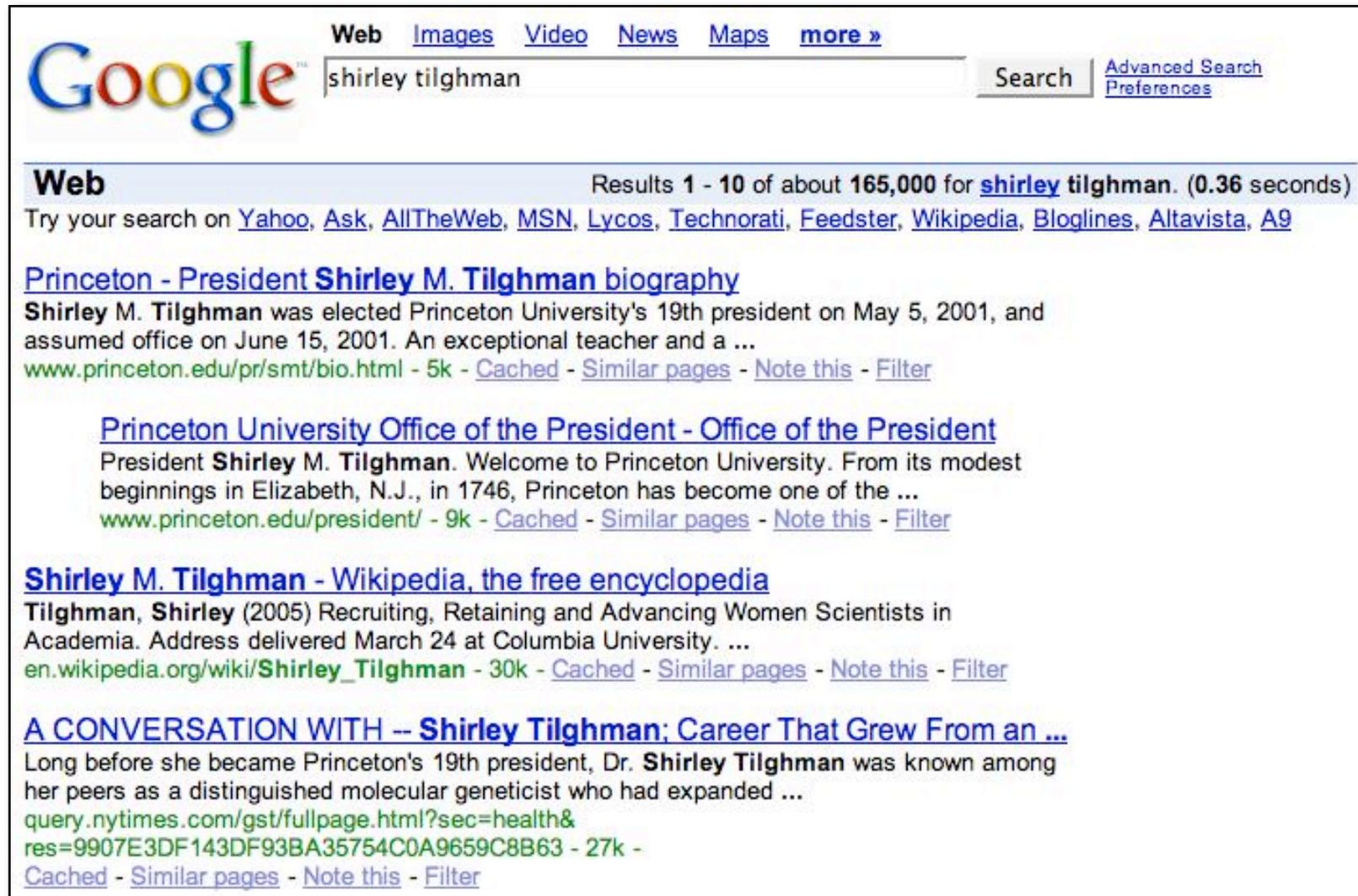
- **ASCII:** Agreed-upon convention for representing letters with numbers
- Example: **Ideas??**

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 (prev. slide)

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



The image shows a screenshot of a Google search results page. At the top left is the Google logo. To its right are navigation links for 'Web', 'Images', 'Video', 'News', 'Maps', and 'more'. Below these is a search input field containing the text 'shirley tilghman'. To the right of the input field is a 'Search' button and links for 'Advanced Search' and 'Preferences'. Below the search bar, the results are categorized under 'Web'. The first result is titled 'Princeton - President Shirley M. Tilghman biography' and includes a snippet of text about her election as Princeton University's 19th president in 2001. The second result is titled 'Princeton University Office of the President - Office of the President' and includes a snippet of text welcoming visitors to Princeton University. The third result is titled 'Shirley M. Tilghman - Wikipedia, the free encyclopedia' and includes a snippet of text about her recruitment and retention of women scientists. The fourth result is titled 'A CONVERSATION WITH -- Shirley Tilghman: Career That Grew From an ...' and includes a snippet of text about her career as a molecular geneticist. Each result includes a URL, a size indicator (e.g., '5k'), and links for 'Cached', 'Similar pages', 'Note this', and 'Filter'.

Web Images Video News Maps more »

Google™ shirley tilghman Search Advanced Search Preferences

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](#)

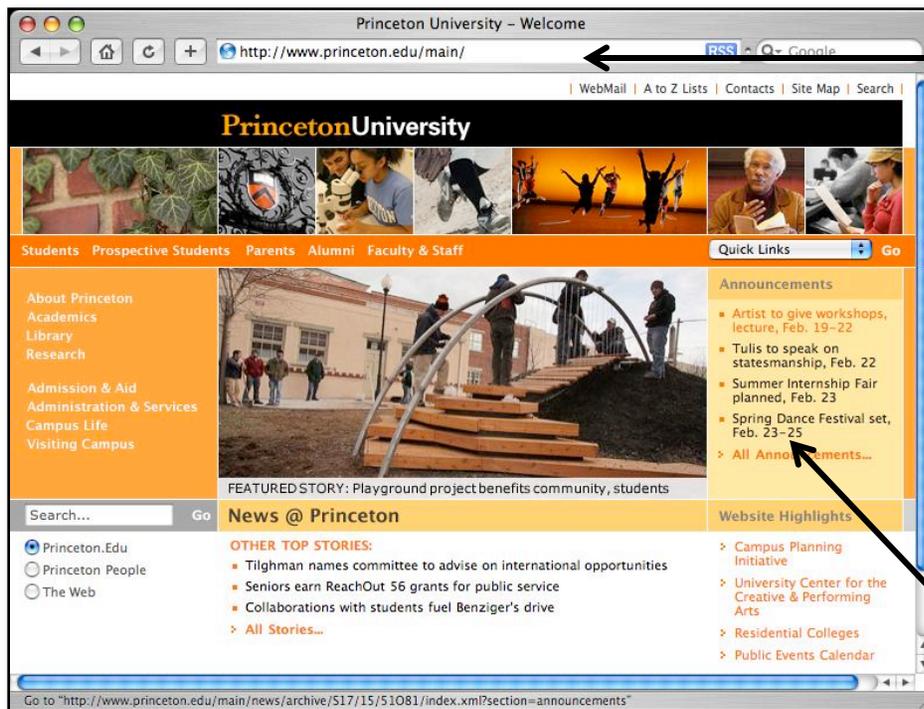
[Princeton University Office of the President - Office of the President](#)
President **Shirley M. Tilghman**. Welcome to Princeton University. From its modest beginnings in Elizabeth, N.J., in 1746, Princeton has become one of the ...
www.princeton.edu/president/ - 9k - [Cached](#) - [Similar pages](#)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

What is World Wide Web?

Files residing on “servers” that are connected to internet.

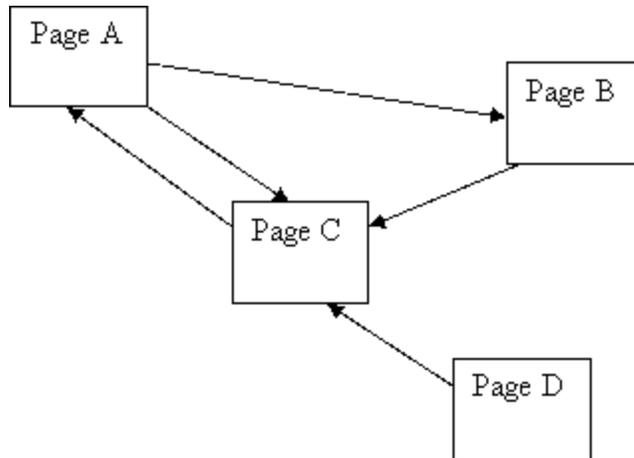


URL (uniform resource locator); basically an “address”

A file “index.html” in “public_html” directory on some server belonging to PU.

“hyperlinks”:
URL of other files; could be on another server.

Logical Structure of the Web

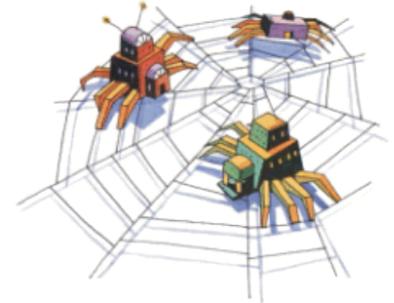


“Directed graph”

“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?



Insights from Googlers into our products, technology, and the Google culture.

We knew the web was big...

7/25/2008 10:12:00 AM

We've known it for a long time: the web is big. The first Google index in 1998 already had 26 million pages, and by 2000 the Google index reached the one billion mark. Over the last eight years, we've seen a lot of big numbers about how much content is really out there. Recently, even our search engineers stopped in awe about just **how** big the web is these days -- when our systems that process links on the web to find new content hit a milestone: 1 trillion (as in 1,000,000,000,000) unique URLs on the web at once!

How do we find all those pages? We start at a set of well-connected initial pages and follow each of their links to new pages. Then we follow the links on those new pages to even more pages and so on, until we have a huge list of links. In fact, we found even more than 1 trillion

Still Feasible Today?



Western Digital - Caviar Black 1TB Internal Serial ATA Hard Drive for Desktops

Model: WD10000LSRTL | SKU: 8909595

Serial ATA interface; integrated dual processors; data transfer rates up to 3 Gbps

★★★★☆ 4.5 of 5 (97 reviews)

[Check Shipping & Availability](#) ▶

\$99.99

add to cart 

bestbuy.com 2/18/2010



Still Feasible Today?

- More than 1 trillion web pages now
- 1 terabyte = 10^{12} byte disk cost \$100
- Say 10 kb (10,000 bytes) of data per page
- 1 petabyte = 10^{15} bytes to store the web
- \approx 1,000 disks
- \approx \$100,000 in 2010



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

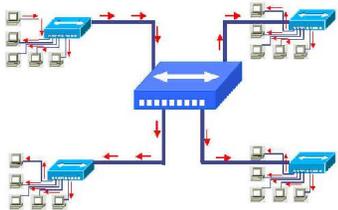
Web link confers “approval”

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”

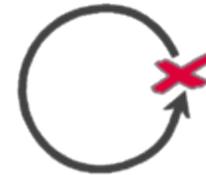
Typically Authorities point to hubs and hubs point to authorities

Circular Definition?



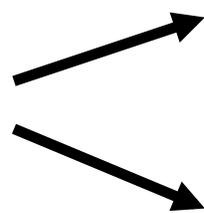
Circular Definition – see Definition, Circular

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.

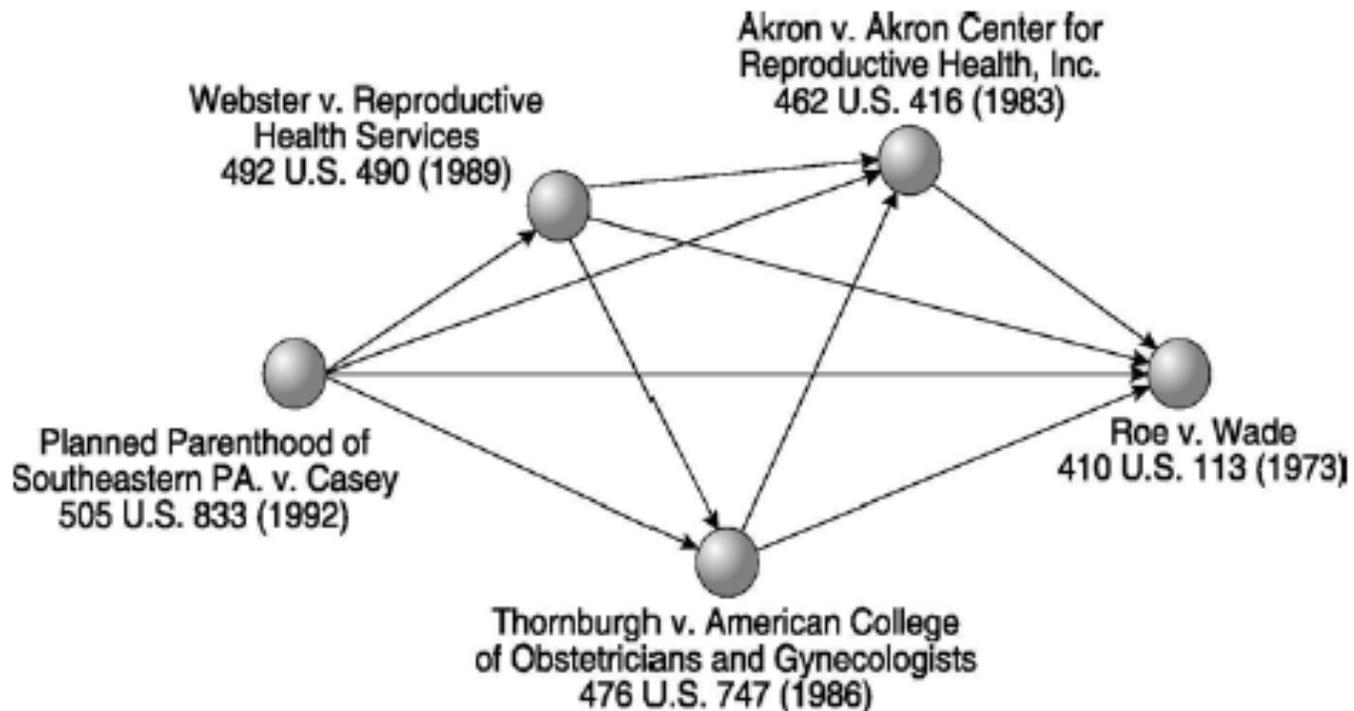
}

Fact The scores converge.

(Proof uses Linear Algebra, Eigenvalues)

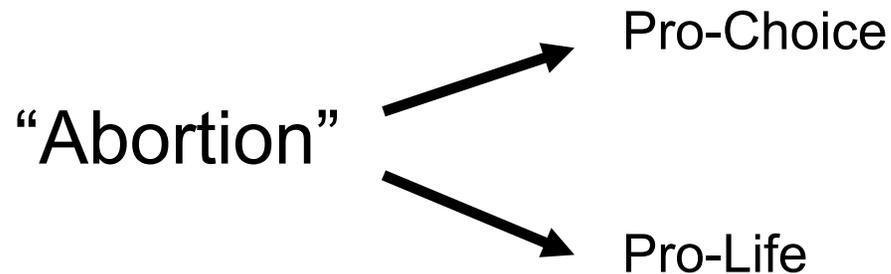
[Fowler and Jeon, '05]

FIGURE 1. Network of Selected Landmark Abortion Decisions



- By product of CLEVER algorithm—
it reveals **clusters**

Example:



- **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

NEWS | ALUMNI

Former Tigers reach finals of \$1 million Netflix competition

By ILYA SABNANI
STAFF WRITER

Published: Monday, February 18th, 2008

 [Print this story](#)  [Email this story](#)

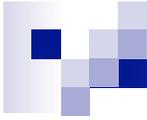
[Respond to this Story](#)

Three friends from the Class of 2007 reached the finals of the Netflix Challenge, a competition held by the internet DVD rental service with the goal of improving its method of predicting customer movie preferences.

Team leader David Weiss '07 and teammates Lester Mackey '07 and David Lin '07 won the "progress prize," an honor that came with a cash prize of \$50,000.



“Netflix Prize seeks to substantially improve the accuracy of predictions about how much someone is going to love a movie based on their movie preferences” (top prize: \$1M)



Trends in web search

Algorithms to “guess” what user generating the query had in mind (using AI, Psychology, User History, News tracking).

Seamless integration with e-commerce, and click-based revenue harvesting (interesting meeting point of economics and computer science)

“Semantic web”: Allow users to attach “meaning” to web-based documents; allowing search engines to make sense of them.

Shape of things to come:

Text & 2D Sketch

Search

Keywords:

View 1 

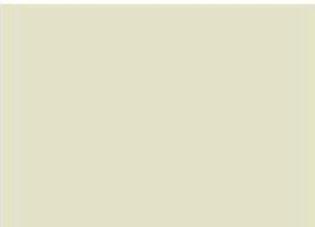
Undo

Clear

View 2 

Undo

Clear

View 3 

Undo

Princeton Shape Retrieval and Analysis Group

3D Model Search Engine

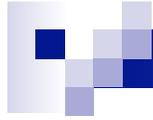
[Text & 2D Sketch](#) [Text & 3D Sketch](#) [File Compare](#) [Research](#) [Contact Us](#) [Links](#) [FAQ](#) [Main](#)

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

 <small>Copyright © 2000, Princeton Corporation or its affiliates</small>	 <small>Copyright © 2000, Princeton Corporation or its affiliates</small>	 <small>Copyright © 2000, Princeton Corporation or its affiliates</small>	 <small>Copyright © 2000, Princeton Corporation or its affiliates</small>
1. vp41562 (vp) Find similar shape	2. vp7513 (vp) Find similar shape	3. vp41425 (vp) Find similar shape	4. vp24269 (vp) Find similar shape
 <small>Copyright © 2000, Princeton Corporation or its affiliates</small>	 <small>Copyright © 2000, Princeton Corporation or its affiliates</small>	 <small>Copyright © 2000, Princeton Corporation or its affiliates</small>	 <small>Copyright © 2000, Princeton Corporation or its affiliates</small>
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>]



Next week...

What computers *cannot* do.