

To Add

- Slide on Amazon & AWS
 - Mech turk,
 - EC2, 8/06
 - S3
 - http://en.wikipedia.org/wiki/Amazon_Web_Services
- History of cloud computing
 - Inktomi, - SOSP 1997 "Cluster-Based Scalable Network Services"
 - GFS, SOSP 2003
 - map reduce, patent filed 6/04
 - EC2
 - Hadoop - 10/05
 - Bigtable - OSDI 11/06
 - Pig etc

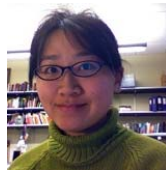
CSE 454

Advanced Internet & Web Services



CSE 454 Advanced Internet & Web Services

- Prof: Dan Weld
 - Most lectures, concepts, perspective.
- TA: Sandra Fan
 - Project details
- Expectations:
 - Project (multiple parts, *on time!*)
 - Reading (papers, web - no formal text)
 - Class participation / development
- Caveat: Life on the cutting edge



10/3/2010 4:24 PM

3

My Background

- Research on Intelligent Internet Systems [1991-
 - Internet Softbot
 - Discover Award Finalist '95
 - Webcrawler
 - By Brian Pinkerton
 - Metacrawler & Shopbot
 - Basis for Netbot Inc.
 - Mulder
 - First automated WWW question answerer
 - KnowItAll
 - Massive, autonomous information extraction
 - Intelligence in Wikipedia Project



10/3/2010 4:24 PM

4

Background Continued

- Co-founded
 - Netbot (Jango)
 - AdRelevance
 - Nimble Technology
 - Asta Networks
- Leaves of absence
 - VP Engineering at Netbot
 - Venture Partner w/ Madrona Venture Group.
- Incredible shortage of software engineers!
- Dearth of training



(r)



Your Background?

- Year in Program?
- Classes?
 - 444, 446, 451, 461, 473, 490H
- Concepts?
 - Threads, race condition, deadlock
 - Naive Bayes classifier
 - Hybrid hash join algorithm
 - Precision, recall
- Programming Background?
 - Ruby, .NET, XML, admin own webserver

10/3/2010 4:24 PM

6

454 Topics

- Information Retrieval
- Search Engines
 - Crawling, Indexing, Query Processing, Ranking
 - Pagerank, Interfaces
- Text Categorization & Clustering
- Information Extraction
 - Machine Learning
- Internet Advertising
- Security, Cryptography, Malware
- Social Networks
- Temporal Web
- Special Topics

Course Outcomes

- After this course, you should know:
 - How search engines work
 - How to build information extraction systems
 - How to ensure a web site scales
 - How Amazon generates personalized recommendations
 - Cryptography fundamentals
 - Other cool stuff
- Focus: search! (why?)

10/3/2010 4:24 PM

8

Why Search?

- A billion or so searches per day...
- Boost to productivity
 - Intellectual & economic
- Search is (*still*) 'hot'
 - Google, Amazon, Ebay, Farecast
 - Search for/in books, products, music, people, ...
- Fascinating research problem.
- You can learn to be a something of a search expert in one quarter!

10/3/2010 4:24 PM

9

What is "Information Extraction"

As a task: Filling slots in a database from sub-segments of text.

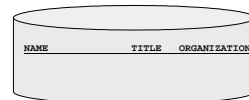
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels—the coveted code behind the Windows operating system—to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying...



Slides from Cohen & McCallum

What is "Information Extraction"

As a task: Filling slots in a database from sub-segments of text.

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels—the coveted code behind the Windows operating system—to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying...



NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft...

Slides from Cohen & McCallum

Why Information Extraction

- Next-Generation Search
 - People
 - Zoominfo
 - Flipdog
 - Intelius
 - Research Papers
 - Citeseer
 - Google scholar
 - Product search
- Question Answering

10/3/2010 4:24 PM

12

Example

The screenshot shows the ZoomInfo website with search filters for 'Person Name' set to 'Daniel weld'. The results table lists 16 people, with Daniel S. Weld being the primary focus. His roles include Venture Partner at Madrona Venture Group LLC, Associate Editor at Access Foundation Inc, and various positions at the University of Washington and Northwestern University.

10/3/2010 4:24 PM

13

...Continued

This section continues the profile of Daniel S. Weld, detailing his employment history at Madrona Venture Group LLC, his references, and his extensive list of affiliations and awards. It highlights his role as a Venture Partner and his involvement in various academic and industry organizations.

10/3/2010 4:24 PM

...Continued Some More

This section continues the profile of Daniel S. Weld, detailing his education (Bachelor's and Ph.D. from Yale University), his research interests in database and data warehousing, and his professional experience at Madrona Venture Group LLC and the University of Washington.

10/3/2010 4:24 PM

15

CiteSeer vs. Scholar

The screenshot shows search results from CiteSeer for the query 'Daniel Weld'. It lists several academic papers, including 'A Software-Based Interface to the Internet' and 'An Approach to Probabilistic Planning', with links to the full documents and citation counts.

Grading

- 85% Project (Staged in Parts)
 - Part artifact
 - Part writeup
 - Clear and concise explanation / justification
 - Experimentation
 - Part presentation
- 15% Class participation

10/3/2010 4:24 PM

17

Capstone Projects

- Done in Group
 - Why?
- Topics
 - Roll your own
 - Or see me

10/3/2010 4:24 PM

18

Start with Concrete Problem

- Text Classification
- Corpus of Wikipedia pages
 - E.g., scientist, writer, author, university
- You'll use machine learning to construct
 - Program which outputs the 'type' of the page
- Details online
 - Done in pairs
 - Due Tues 10/12

Project Possibilities

- Extract Facts from Wikipedia
 - Or recipes, or ...?
- Build Ontology of Products & Attributes
- Mine product reviews for attribute valence
- Or suggest something different

Timeline

- Assemble into pairs by over weekend
 - Needed for PS1
- Propose a project idea in class on Tues
- Final teams and projects settled by 10/12

Last Quarter's Projects

- Topocycle
 - Google map-style website for planning bike rides
- Craigslist Rank & Search
- Tastecliq
 - Online service for discussion & recommendation of media (eg movies)
- Instroodle
 - Centralized site to help students pick which classes to take and choose between professors
- Paperazzi
 - Visual search engine for research papers

Previous Quarter's Projects

- Craigslist++
- University Search
- Twitter Feedrank
- Apartment Listing & Aggregation
- Webcam Identification & Search
- Trail / Hike Search
- Seattle Event Finder
- Automatic Stock Investor

What This Course Is Not

... there is a difference between training and education. If computer science is a fundamental discipline, then university education in this field should emphasize enduring fundamental principles rather than transient current technology.
-Peter Wegner, *Three Computing Cultures*. 1970.

- We won't:
 - Teach you how to be a web master
 - Teach all the latest x-buzzwords in technology
 - XML/SOAP/WSDL
 - (okay, may be a little).
 - Teach web/javascript/java/jdbc... programming

Warning

- No textbook
- Large project component
- Poorly documented, unstable systems
- Field changes quickly
 - Each year is essentially a new course
- Need students to help debug class!

10/3/2010 4:24 PM

25

Ancient History

- Pre-history: Dewey Decimal system
 - Bizarre medieval rituals performed by hand
- 1960: Ted Nelson → Xanadu
 - Hypertext vision of WWW
 - Why did it fail?
 - Focus on copyright issues
 - Still a thorny problem
 - Focus on stable, bidirectional links
 - "Trying to fix HTML is like trying to graft arms and legs onto hamburger" -- Ted Nelson



1961 Kleinrock paper on packet switching

Contrast with phone lines - circuit switched.

10/3/2010 4:24 PM

26

Paleolithic Era

- 1965 Gordon Moore proposes law
- 1966 Design of ARPAnet
- 1968 Doug Engelbart:
The first WIMP
- 1969 First ARPAnet message
UCLA -> SRI
- 1970 ARPAnet spans country, has 5 nodes
- 1971 ARPAnet has 15 nodes
- 1972 First email programs, FTP spec



10/3/2010 4:24 PM

27

The Personal Computer Era

- 1974 Intel launches 8080;
TCP design
- 1975 Gates/Allen write Basic - Altair 8800
- 1976 Jobs/Wozniak form Apple Computer
111 hosts on ARPAnet
- 1979 Visicalc
- 1981 Microsoft has 40 employees;
IBM PC
- 1984 Launch of Macintosh
- 1986 Microsoft goes public

10/3/2010 4:24 PM

28

Internet Ramps Up

- 1983 ARPAnet uses TCP/IP, Design of DNS
1000 hosts on ARPAnet
- 1985 Symbolic.com first registered domain name
- 1989 100,000 hosts on Internet
- 1990 Cisco Systems goes public
Tim Berners-Lee creates WWW at CERN

10/3/2010 4:24 PM

29

Web Search Pre-History

- 1950s: "Information Retrieval" (IR) term coined
- 1960s-70s: SMART system, vector space model,
 - Gerald Salton (Cornell) father of IR
- 1980s: Proprietary document DBs
 - (Lexis-Nexis, Medline)
- 1990: Archie (index file names, anon. ftp)
- 1991: Gopher (menus, links to servers)
- 1992: Veronica (index of menu items on gophers)
- 1993: Jughead (keyword + boolean search)
 - Rapid evolution, but what is missing?

10/3/2010 4:24 PM

30

Modern History of Search

- 1993: WWW Wanderer (first crawler)
- 1994: WebCrawler, Lycos (1st widely-used SEs)
 - WebCrawler was a UW class project by Brian Pinkerton
- 1994: Yahoo directory (Stanford; founded '95)
Amazon founded
Netscape founded (90% mkt share → 1%)
- 1995: Ebay
MetaCrawler (1st major meta-SE)
 - UW Master's thesis by Erik Selberg

10/3/2010 4:24 PM

31

Discovery of the Biz Model

- 1996: Flash by Macromedia
later acquired by Adobe
- 1997: goto.com
"sponsored links" pay-per-click
AskJeeves
manually-powered question answering
Netbot
comparison-shopping search
- 1998: Open directory launched
Google, pagerank algorithm
Paypal founded

Turn of the Millennium

- 1999:  becomes dominant browser
Napster starts operation
Search Engines → portals (Yahoo, Excite)
"Search is a commodity"
- 2000: Flipdog
Commercial information extraction
- 2001: Bittorrent protocol (soon 35% of internet)
Ascendance of Google
"Search is nirvana"
- 2002: IE peaks at 90% market share



10/3/2010 4:24 PM

33

Approaching the Present

- 2003: Skype released
- 2004: Facebook founded
Social news (Digg)
- 2005: Youtube founded
 - 9.5 B videos shown per month
 - 33 months after founding!
- 2006: Twitter founded
- 2007: Google Streetview
Apple iPhone
- 2009: Facebook 200M users



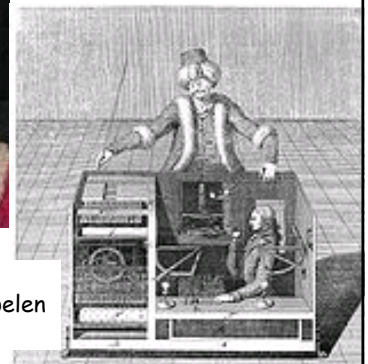
Future of the Net

- Domination of Mobile Devices (cellphone, etc)
- Link-Spamming (Arms race to bias SE ranking)
- Local Search, Digital Earth
- Image & Video search
- Social news (Digg / Twitter)
- Crowd Sourcing
- What else?

10/3/2010 4:24 PM

35

Mechanical Turk




Built in 1770 by
Wolfgang von Kempelen

10/3/2010 4:24 PM

amazonmechanical turk
beta Artificial Intelligence

- **Launched in Nov '05**
 - Initially: detect duplicate product pages
- **100k workers in 100 countries by 3/07**
 - 34k HITs on 3/28/08
- **Search for Jim Gray**
 - 12k searchers



10/3/2010 4:24 PM 37

Death of the Web

- **Pages vs Apps**
 - Can't search apps
 - Still use HTTP, but closed protocols
- **HTML5?**

Observations

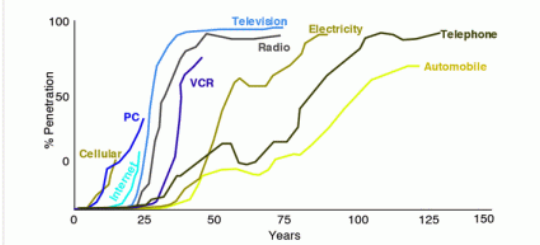
- **Internet/Web *evolved*** - it wasn't created
- **Scalability beats structure**
 - search engines over directories
 - Web over hypertext
- **"We are 10 seconds from the Big Bang"**
 - John Doerr

10/3/2010 4:24 PM 39

Adoption

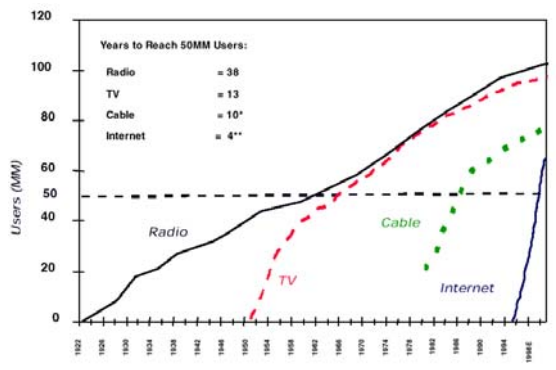
Facilitating Innovation the pace of innovation is increasing

- Newer technologies taking hold at double or triple previous rates



The graph shows the percentage penetration of various technologies over time. The x-axis represents years from 0 to 150, and the y-axis represents % Penetration from 0 to 100. Technologies shown include Cellular, PC, Television, Radio, VCR, Electricity, Telephone, and Automobile. The graph illustrates that newer technologies are being adopted much faster than older ones.

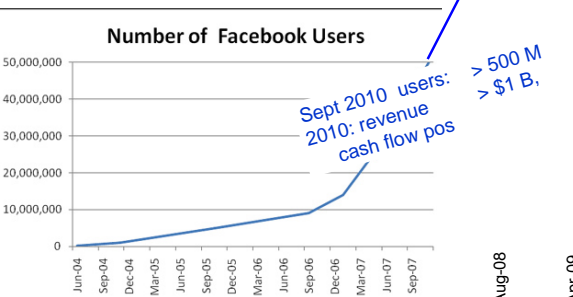
Accelerating



The graph shows the number of users in millions from 1922 to 1996 for Radio, TV, Cable, and Internet. A dashed horizontal line is drawn at 50 million users. A table indicates the years to reach 50 million users for each technology.

Technology	Years to Reach 50MM Users
Radio	= 38
TV	= 13
Cable	= 10*
Internet	= 4**

And now?



The graph shows the number of Facebook users from June 2004 to April 2009. The y-axis represents the number of users from 0 to 50,000,000. A blue line shows exponential growth. Handwritten notes indicate: "Sept 2010 users: > 500 M" and "2010: revenue cash flow pos".

Facebook: Tool of the Devil?

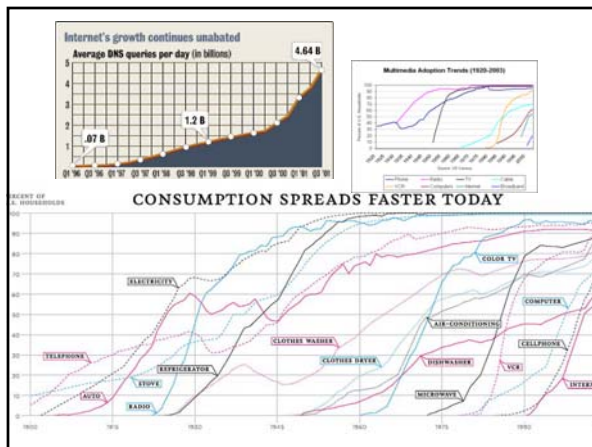
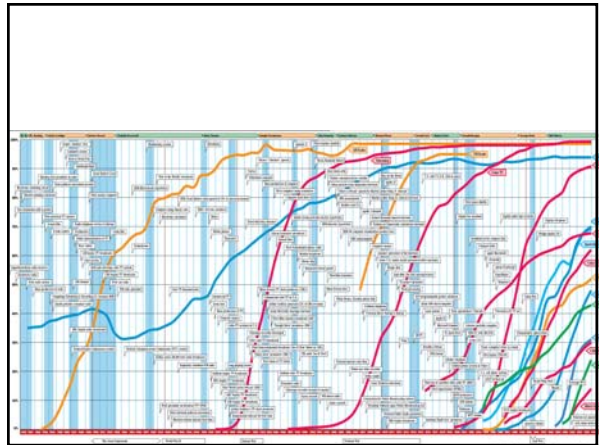


For Next Time

- Add yourself to mailing list
 - We'll send out a **key email** tomorrow
 - Be sure to get it !
- Form a group of 2 people
 - Think about ps1
 - Brainstorm project idea

10/3/2010 4:24 PM

44



33 months after founding

Top U.S. Online Video Properties* by Videos Viewed
November 2007
Total U.S. - Home/Work/University Locations
Source: comScore Video Metrix

Property	Videos Viewed (MM)	Share (%) of Videos
Total Internet	9,491	100.0%
Google Sites	2,966	31.3%
Fox Interactive Media	419	4.4%
Yahoo! Sites	328	3.5%
Viacom Digital	245	2.6%
Time Warner Network	184	1.9%
Microsoft Sites	181	1.9%
Disney Online	96	1.0%
ABC.com	88	0.9%
ESPN	87	0.9%
Break	47	0.5%

10/3/2010 4:24 PM

48