

# Search Engines: Technology, Society, and Business

---

Course Summary

Marti Hearst

December 5, 2005

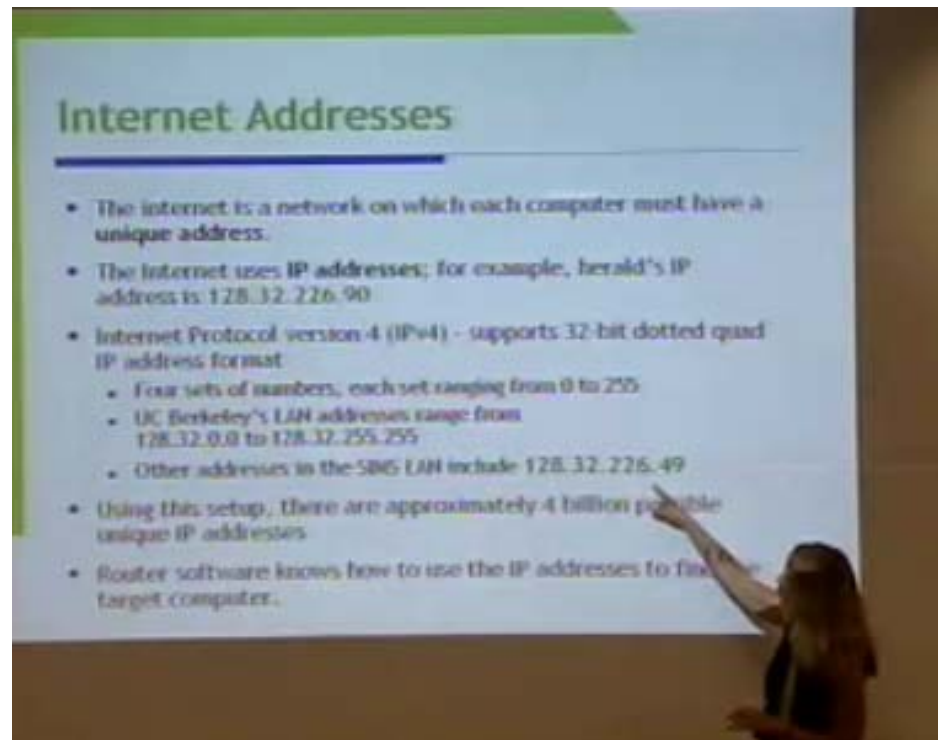
# Course Goals

---

- Gain an interdisciplinary understanding of search engines and related technologies.
  - How they work
  - How they affect communication
  - How they affect business
  - How they are changing our understanding of information and knowledge.
- Make the techy parts understandable for everyone.

# Intro to the Internet & WWW

- Prof. Hearst



# John Battelle

---

- The Search




## Search and Culture

- The Realization: My God....Google Knows What We Want...
- The Database of Intentions
- Ephemeral to Eternal
- First Use Case: Paid Search

# Dr. Jan Pedersen

- The Four Dimensions of Search Quality



 Freshness


- Problem:
  - Ensure that what is indexed correctly reflects current state of the web
- Impossible to achieve exactly
  - Revisit vs Discovery
- Divide and Conquer
  - A few pages change continually
  - Most pages are relatively static

YAHOO!

# Dr. Dan Rose

- User Experience Issues in Web Search




 **Vocabulary Problem**

- People use different words for the same thing.
  - <20% chance of choosing same word
  - Even “best” word has 65-85% failure

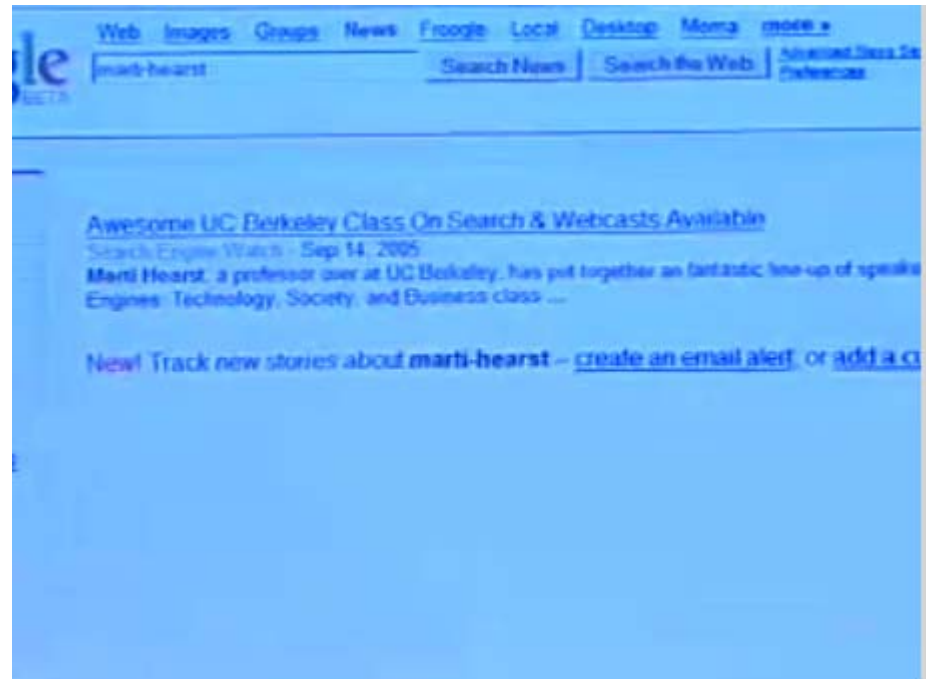
*“... The data show that no single access word, however well chosen, can be expected to cover more than a small proportion of users’ attempts...”*

Forman, G.W., et al (1987). “The Vocabulary Problem in Human-System Communication,” Communications of the ACM, 30(11): 964-971.



# Dr. Peter Norvig

- Google News, Print, Maps, & Earth



# Dr. Sep Kamvar

---

- Personalization and Search






# Dr. John Chuang

- Peer-to-peer Search



### Hierarchical networks

- This two-level hierarchy:
  - Super nodes are elected as "super nodes" or "alpha peers"
  - Each alpha peer serves as a central node for a portion of the network
  - If an alpha peer drops out, some users may not be able to connect to other alpha peers
- Advantages:
  - Works in the presence of network failure
  - Easy to scale
- Drawbacks:
  - Not all users can connect to super nodes
  - Super nodes may be overloaded
  - Super nodes may be slow
  - Super nodes may be expensive

A diagram illustrating a hierarchical network structure. It shows a central node (super node) connected to several smaller nodes (alpha peers). Each alpha peer is further connected to a group of smaller nodes (users). The diagram is set against a light blue background with a network logo in the top right corner.

# Sergey Brin

---

- Google and Life



# Dr. Hal Varian

- Search advertising



## Factors affecting revenue

$$\begin{aligned} \text{Monetization (RPM)} &= \frac{\text{Revenue}}{\text{Queries}} \times (1K) \\ &= \frac{\text{Revenue}}{\text{Clicks}} \times \frac{\text{Clicks}}{\text{Queries}} \\ &= \frac{\text{Revenue}}{\text{Clicks}} \times \frac{\text{Queries w/ Ads}}{\text{Queries}} \times \frac{\text{Ads}}{\text{Queries w/ Ads}} \times \frac{\text{Clicks}}{\text{Ads}} \\ &= \underbrace{\text{CPC}} \times \underbrace{\text{Coverage} \times \text{Depth}} \times \underbrace{\text{CTR per Ad}} \end{aligned}$$

Price	Quantity	Quality
-------	----------	---------

# Jason Schultz

- Search and Intellectual Property



Are we being diverted or informed?

**E** Electronic Frontier Foundation

# Dr. Sue Dumais

- Desktop Search



## SIS Demo

2767 rows returned

Document	Date	Path	Author	Mail To
<input checked="" type="checkbox"/> All (2767)	<input checked="" type="checkbox"/> All (2767)			
<input checked="" type="checkbox"/> Web Pages (7)	<input type="checkbox"/> Today (25)			
<input checked="" type="checkbox"/> Outlook (2625)	<input type="checkbox"/> Yesterday (17)			
<input checked="" type="checkbox"/> Files (139)	<input type="checkbox"/> Last 7 days (96)			
	<input type="checkbox"/> Last 30 days (488)			
	<input type="checkbox"/> Older than 30 days [...]			

Future

Updated: Stuff I've Seen... Fast mail ind... 11/4/2002 1:00 PM mailbox - susan dumais/sent items Susan Dumais Marc Olson, Wil Kennedy, Jensen Henrik, ...  
When: Monday, November 04, 2002 1:00 PM (GMT-08:00) Pacific Time (US & Canada); To: Marc Olson, Wil Kennedy, Jensen Henrik, ...  
Fido: We are waiting on a prototype called Stuff I've Seen (SIS). SIS provides an integrated index of all the things you look

Today

stuff i've seen - outlook 11/1/2002 5:18 PM c:\personal\papers\misc\misc Susan Dumais  
Stuff I've Seen Susan Dumais, Ed Cuhell, V Code, Gavin Jones, Ramon Sam, Microsoft Research Search Fodge, and Tomonow SIS Details Unified index of stuff you've seen Web  
pages, office docs, email, and more. Full-text index of content plus metadata attributes (e.g., creation time, author).

RE: Local store vs. Server hits in SIS 11/1/2002 5:16 PM mailbox - susan dumais/sent items Susan Dumais Edvard Cuhell  
Ed - Can you send the file (I can't seem to cut and past the figures below into ppt. Thanks. Sus 6 - Original Message - From: Edvard Cuhell Sent: Friday, November 01,  
2002 3:34 PM To: Susan Dumais; Adrian Klein; R. Co-SIV; Lutz; R. object; R. oval; Steve vs. Server hits in SIS

Local store vs. Server hits in SIS 11/1/2002 3:33 PM mailbox - susan dumais/inbox Edvard Cuhell Susan Dumais; Adrian Klein  
Simon also from our topic. It all real time aggregated status of items. 433 was total hits (not. also 6 672 was located on the server. On a previous base, I determined the  
of email opened that was local and the others from server. Then I did a histogram of these values for

Stuff I've Seen... Fast mail indexing and ... 11/1/2002 3:00 PM mailbox - susan dumais/sent items Susan Dumais Marc Olson, Wil Kennedy, Jensen Henrik, ...  
When: Friday, November 01, 2002 3:00 PM (GMT-08:00) Pacific Time (US & Canada); To: Marc Olson, Wil Kennedy, Jensen Henrik, ...  
called Stuff I've Seen (SIS). SIS provides an integrated index of all the things you look at, including files, web pages, email

RE: Things have quieted down again 11/1/2002 10:47 AM mailbox - susan dumais/sent items Susan Dumais Eugene Samsonov; Ramon Samin [E...  
It's simply not as easy as "is search". I get 12 on I/O read, and 174 on I/O write (the search amount varies a bit from time to time). Incremental crawl is all that is happening

RE: Things have quieted down a... 11/1/2002 10:47 AM mailbox - susan dumais/inbox Susan Dumais Eugene Samsonov; Ramon Samin [E...  
It's simply not as easy as "is search". I get 12 on I/O read, and 174 on I/O write (the exact amount varies a bit from time to time). Incremental crawl is all that is happening

# Dr. Mark Najork

- Web Spam



## Examples of synthetic content



Monetization

Random words

Well-formed sentences stitched together

Links to keep crawlers going

# Dr. Doug Tygar

---

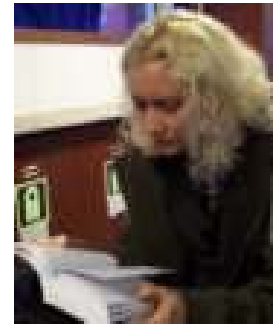
- Search and Privacy



# Dr. Alma Whitten

---

- Search Engine Architecture





# Dr. Eric Brewer

---

- Reflections on Starting a Search Engine Company



# Avi Rappaport

- Enterprise Search



## Choose, Implement, Maintain

- Buy or download, don't build
  - Quality search and other features are non-trivial
  - Homegrown systems rarely satisfy
- Effort to implement varies wildly
  - Number of documents, complexity, interfaces
  - Resources, servers, network
  - Enterprise information needs
- Maintenance
  - Keep system working, index current, scale up
  - Add new data sources
  - Change as new needs appear
  - Log analysis

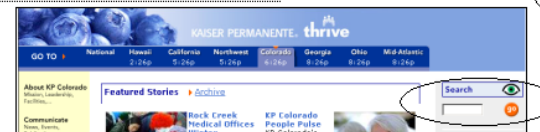
# Jennifer English

- Enterprise Search



Can the user find the search box?

- Inconsistent placement.
- Size – text box usually too small to accommodate a reasonable query.
- Wording around search box/button is inconsistent.
- Text entry boxes are “ugly” – designers want them to be as small and inconspicuous as possible



# Bradley Horowitz

---

- Mass Media and Micro Media Search at Yahoo

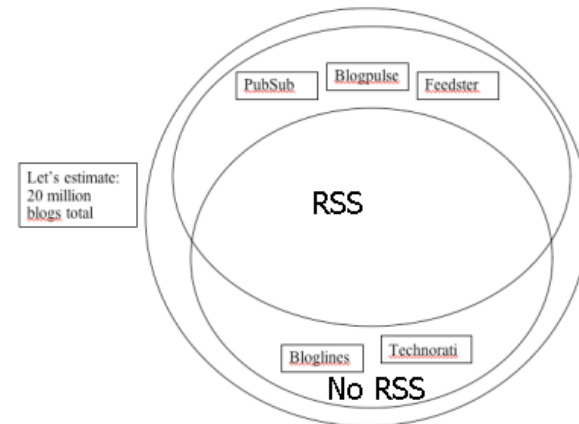


# Mary Hodder

- Searching the Live Web



The Blogosphere: what are you searching?

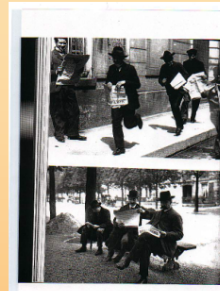


# Dr. Geoff Nunberg

- How Search Shapes Cyberspace



## Formal Correlates of Metrical Space: The Book as a Public Presence



The newspaper reader, observing exact replicas of his own paper being consumed by his subway, barbershop, or residential neighbors, is continually reassured that the imagined world is visibly rooted in everyday life...creating that remarkable confidence of community in anonymity which is the hallmark of modern nations. —Benedict Anderson, *Imagined Communities*.



# Dr. Marti Hearst

- Faceted Metadata in Search Interfaces



**Objects (group results)**

- [Clothing](#) (12)
- [Containers](#) (4)
- [Food and meals](#) (18)
- [Fuel](#) (2)
- [Lighting](#) (1)
- [Musical instruments](#) (3)
- [Vehicles](#) (8)
- [Weapons](#) (66)
- [Writing tools](#) (4)

**Themes: all > military > war (group results)**

- [Battle](#) (28)
- [Combat](#) (15)
- [Duel](#) (1)
- [Fighting](#) (14)
- [War](#) (19)

**Shapes, Colors, and Scenes (group results)**

- [Color](#) (11)
- [Decoration](#) (8)
- [Metal](#) (5)
- [Scene](#) (24)
- [Shape](#) (1)

**Artists (group results)**

- [Anonymous](#) (2)
- [Baur, johann wilhelm, 1600 - 1640](#) (2)
- [Beham, barthel, 1502 - 1540](#) (1)
- [Berthault, pierre gabriel, 1748 - 1819](#) (3)
- [Bry, johann-theodore de, 1561 - 1623](#) (1)
- [Burgkmair, the elder, hans, 1473 - 1531](#) (4)
- [Callot, jacques, 1592 - 1635](#) (12)
- [Chesham, francis, 1749 - 1806](#) (1)
- [Courtois, jacques \(le bourguignon\), 1621 - 1676](#) (4)
- [Daddi, bernardo, 1512 -](#) (1)

Grid of artwork thumbnails:

- Capricci Di Vari... Baur 1635
- Capricci Di Vari... Baur 1635
- Duel de Faust et... Delacroix 1827
- Five Men Fightin... Daddi 1532
- Four Soldiers Fi... Ottavianni 18th century
- Killing the Buff... Anonymous 19th century
- Krieg, from Vom ... Klinger 1885
- Les Caprices Callot 1617
- Les Caprices Callot 1617
- Les Caprices Callot 1617
- Les Caprices Callot 1617
- Les Caprices Callot
- Les Caprices Callot
- Les Caprices Callot
- Les Caprices Callot

# What is the Future of Search?

---



# Administrivia

---

# Final Projects

---

- Turn them in using online links
- **HARD DEADLINE!**
- Undergrads: due Friday Dec 9, 7pm  
<http://www.sims.berkeley.edu/courses/is141/f05/discussion.html>
- Grads: due Saturday Dec 10, 5pm  
<http://www.sims.berkeley.edu/courses/is290-2/f05/assignments.html>

# Course Evaluations

---

- This is the SIMS form
- First page is instructor evaluation
- Back of page is course evaluation
- PLEASE CIRCLE THE CLASS YOU ARE IN AT THE TOP
  - 141 vs 290-2
- Instructor does NOT see these until after she turns in the grades.
- Turn in the form to a TA, who will then check you off for attendance.
  - TAs will not accept forms until 10 minutes after they are distributed.

# Let's Thank Our TAs!

---

Helen Kim and Fredrik Wallenberg

# Thank you!

---

And Happy Searching!