Administrative

- Paper review
  - submit in PDF or plain text format
  - include name, email address, date
- Class meeting time, classroom
- Sign up for presentations

How to write a paper review?

- Journal/conference paper review
  - suitability, novelty, technical depth, writing
  - summary, strengths, weaknesses, reference
  - comments to editor/PC chair
- Paper reviews for this course
  - literature survey
  - don't just read, THINK

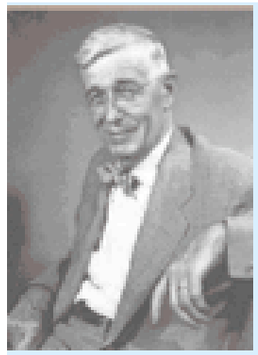
STONY BROOK
STATE UNIVERSITY OF NEW YORK

Data, Data, Data

CSE692 Fall 2007

A Digitized World

- PCs, laptops, Internet
- email, IM, VoIP, ...
- pdf, doc, ppt, xls, txt, ...
- audio, video, digital photos, ...
- World Wide Web, blogs, online news, ...
- medical data, scientific data, ...
- ...

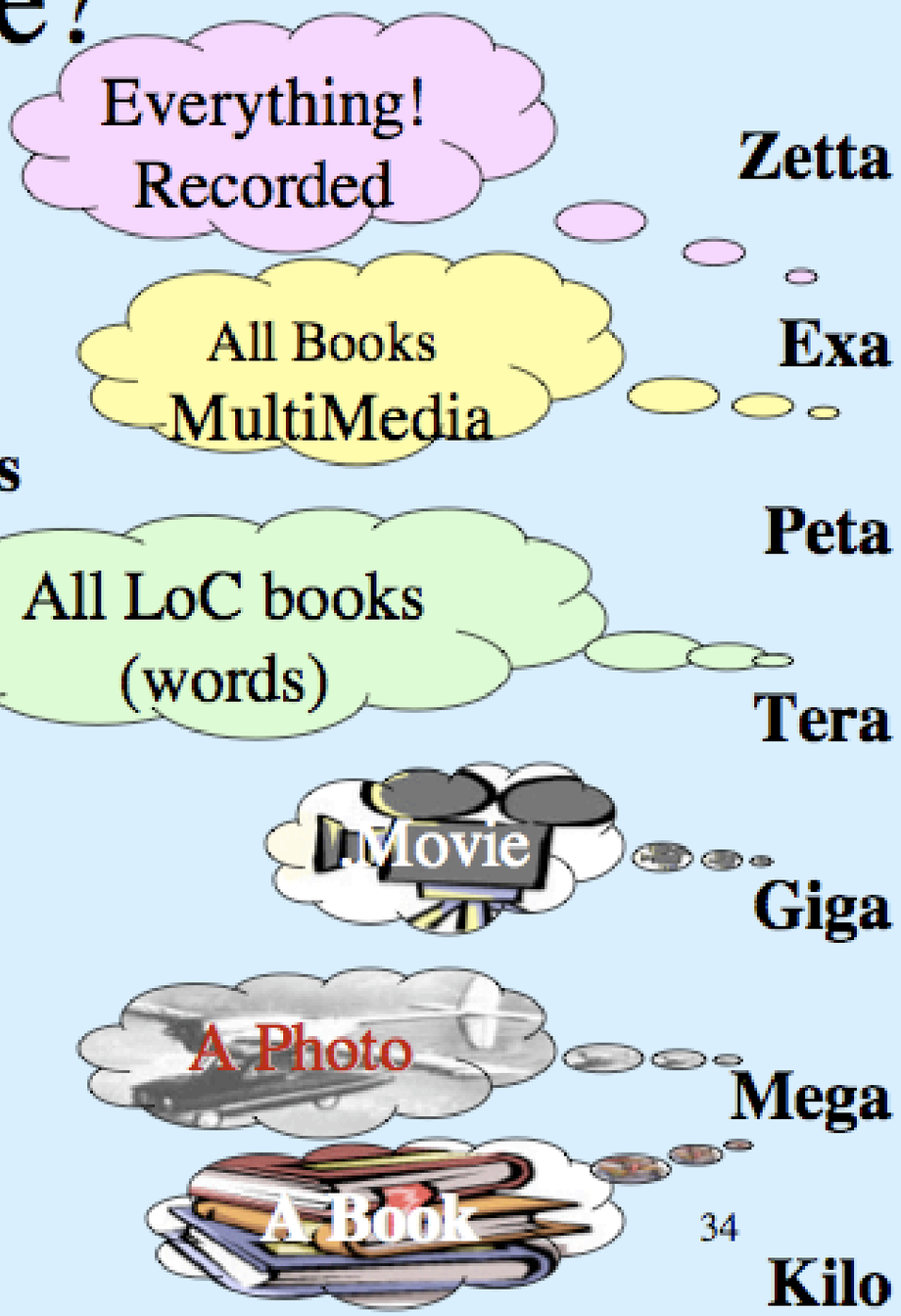


How Much Information Is there?

- **Soon everything can be recorded and indexed**
- **Most data never be seen by humans**
- **Precious Resource: Human attention**
  Auto-Summarization
  Auto-Search
  is key technology.

www.lesk.com/mlesk/ksg97/ksg.html

24 Yecto, 21 zepto, 18 atto, 15 femto, 12 pico, 9 nano, 6 micro, 3 milli



How Much Information? 2003

- Lead by Peter Lyman and Hal R. Varian
- School of Information Managements and Systems (SIMS), UC Berkeley
- Based on year 2002 data
- Previous 2000 study based on 1999 data
- Goal
  - Estimate how much new information is created each year

Approaches

- Storage media
  - print, film, magnetic, optical
- Information flow
  - telephone, radio, TV, Internet

How Much New Information?

Table 1.2: Worldwide production of original information, if stored digitally, in terabytes circa 2002. Upper estimates assume information is digitally scanned, lower estimates assume digital content has been compressed.

Storage Medium	2002 Terabytes Upper Estimate	2002 Terabytes Lower Estimate	1999-2000 Upper Estimate	1999-2000 Lower Estimate	% Change Upper Estimates
Paper	1,634	327	1,200	240	36%
Film	420,254	76,69	431,690	58,209	-3%
Magnetic	5187130	3,416,230	2,779,760	2,073,760	87%
Optical	103	51	81	29	28%
TOTAL:	**5,609,121**	**3,416,281**	**3,212,731**	**2,132,238**	**74.5%**

Source: How much information 2003

How Much New Information?

Table 1.2: Worldwide production of original information, if stored digitally, in terabytes circa 2002. Upper estimates assume information is digitally scanned, lower estimates assume digital content has been compressed.

Storage Medium	2002 Terabytes Upper Estimate	2002 Terabytes Lower Estimate	1999-2000 Upper Estimate	1999-2000 Lower Estimate	% Change Upper Estimates
Paper	1,634	327	1,200	240	36%
Film	420,254	76,69	431,690	58,209	-3%
Magnetic	5187130	3,416,230	2,779,760	2,073,760	87%
Optical	103	51	81	29	28%
TOTAL:	**5,609,121**	**3,416,281**	**3,212,731**	**2,132,238**	**74.5%**

Source: How much information 2003

How Much New Information?

Table 1.2: Worldwide production of original information, if stored digitally, in terabytes circa 2002. Upper estimates assume information is digitally scanned, lower estimates assume digital content has been compressed.

Storage Medium	2002 Terabytes Upper Estimate	2002 Terabytes Lower Estimate	1999-2000 Upper Estimate	1999-2000 Lower Estimate	% Change Upper Estimates
Paper	1,634	327	1,200	240	36%
Film	420,254	76,69	431,690	58,209	-3%
Magnetic	5187130	3,416,230	2,779,760	2,073,760	87%
Optical	103	51	81	29	28%
TOTAL:	**5,609,121**	**3,416,281**	**3,212,731**	**2,132,238**	**74.5%**

Source: How much information 2003

How Much New Information?

Table 1.2: Worldwide production of original information, if stored digitally, in terabytes circa 2002. Upper estimates assume information is digitally scanned, lower estimates assume digital content has been compressed.

Storage Medium	2002 Terabytes Upper Estimate	2002 Terabytes Lower Estimate	1999-2000 Upper Estimate	1999-2000 Lower Estimate	% Change Upper Estimates
Paper	1,634	327	1,200	240	36%
Film	420,254	76,69	431,690	58,209	-3%
Magnetic	5187130	3,416,230	2,779,760	2,073,760	87%
Optical	103	51	81	29	28%
TOTAL:	**5,609,121**	**3,416,281**	**3,212,731**	**2,132,238**	**74.5%**

Source: How much information 2003

Magnetic Storage Media

Table 1.6: Worldwide production of magnetic original content, if stored digitally using standard compression methods, in terabytes circa 2002.

Storage Medium	Type of Content	Terabytes/Yr Upper Estimate	Terabytes/Yr Lower Estimate	1999 Report Upper Estimate	1999 Report Lower Estimate	% Change Upper Estimates
Magnetic	Videotape	1,340,000	1,340,000	1,420,000	1,420,000	-6%
	Audiotape	128,800	128,800	182,000	182,000	-30%
	Digital tape	250,000	250,000	250,000	250,000	0
	MiniDV	1,265,000	1,265,000	N/A	N/A	N/A
	Floppy disc	80	80	70	70	14%
	Zip	350	350	1,690	1,690	-79%
	Audio MD	17,000	17,000	N/A	N/A	N/A
	Flash	12,000	12,000	N/A	N/A	N/A
	Hard Disk	1,986,000	403,000	926,000	220,000	114%
	TOTAL	**4,999,230**	**3,416,230**	**2,779,760**	**2,073,760**	**80%**

Source: How much information 2003

Magnetic Storage Media

Table 1.6: Worldwide production of magnetic original content, if stored digitally using standard compression methods, in terabytes circa 2002.

Storage Medium	Type of Content	Terabytes/Yr Upper Estimate	Terabytes/Yr Lower Estimate	1999 Report Upper Estimate	1999 Report Lower Estimate	% Change Upper Estimates
Magnetic	Videotape	1,340,000	1,340,000	1,420,000	1,420,000	-6%
	Audiotape	128,800	128,800	182,000	182,000	-30%
	Digital tape	250,000	250,000	250,000	250,000	0
	MiniDV	1,265,000	1,265,000	N/A	N/A	N/A
	Floppy disc	80	80	70	70	14%
	Zip	350	350	1,690	1,690	-79%
	Audio MD	17,000	17,000	N/A	N/A	N/A
	Flash	12,000	12,000	N/A	N/A	N/A
	Hard Disk	1,986,000	403,000	926,000	220,000	114%
	TOTAL	**4,999,230**	**3,416,230**	**2,779,760**	**2,073,760**	**80%**

Source: How much information 2003

Magnetic Storage Media

Table 1.6: Worldwide production of magnetic original content, if stored digitally using standard compression methods, in terabytes circa 2002.

Storage Medium	Type of Content	Terabytes/Yr Upper Estimate	Terabytes/Yr Lower Estimate	1999 Report Upper Estimate	1999 Report Lower Estimate	% Change Upper Estimates
Magnetic	Videotape	1,340,000	1,340,000	1,420,000	1,420,000	-6%
	Audiotape	128,800	128,800	182,000	182,000	-30%
	Digital tape	250,000	250,000	250,000	250,000	0
	MiniDV	1,265,000	1,265,000	N/A	N/A	N/A
	Floppy disc	80	80	70	70	14%
	Zip	350	350	1,690	1,690	-79%
	Audio MD	17,000	17,000	N/A	N/A	N/A
	Flash	12,000	12,000	N/A	N/A	N/A
	Hard Disk	1,986,000	403,000	926,000	220,000	114%
	TOTAL	**4,999,230**	**3,416,230**	**2,779,760**	**2,073,760**	**80%**

Source: How much information 2003

Magnetic Storage Media

Table 1.6: Worldwide production of magnetic original content, if stored digitally using standard compression methods, in terabytes circa 2002.

Storage Medium	Type of Content	Terabytes/Yr Upper Estimate	Terabytes/Yr Lower Estimate	1999 Report Upper Estimate	1999 Report Lower Estimate	% Change Upper Estimates
Magnetic	Videotape	1,340,000	1,340,000	1,420,000	1,420,000	-6%
	Audiotape	128,800	128,800	182,000	182,000	-30%
	Digital tape	250,000	250,000	250,000	250,000	0
	MiniDV	1,265,000	1,265,000	N/A	N/A	N/A
	Floppy disc	80	80	70	70	14%
	Zip	350	350	1,690	1,690	-79%
	Audio MD	17,000	17,000	N/A	N/A	N/A
	Flash	12,000	12,000	N/A	N/A	N/A
	Hard Disk	1,986,000	403,000	926,000	220,000	114%
	TOTAL	**4,999,230**	**3,416,230**	**2,779,760**	**2,073,760**	**80%**

Source: How much information 2003

How Much New Info Flows?

Table 1.9: Summary of electronic flows of new information in 2002 in terabytes.	
Medium	**2002 Terabytes**
Radio	3,488
Television	68,955
Telephone	17,300,000
Internet	532,897
TOTAL	**17,905,340**

Source: How much information 2003

- voice telephone traffic
- Internet traffic
- VoIP, HDTV, IPTV, P2PTV, ...

The Size of Internet

- surface web, deep web, blogs
- email: 31 billion daily, about 1/3 spam
- P2P file sharing
- IM

Table 1.13: The size of the Internet in terabytes.	
Medium	**2002 Terabytes**
Surface Web	167
Deep Web	91,850
Email (originals)	440,606
Instant messaging	274
TOTAL	**532,897**

Source: How much information 2003

Internet Email Spam

Email Spam Percentage

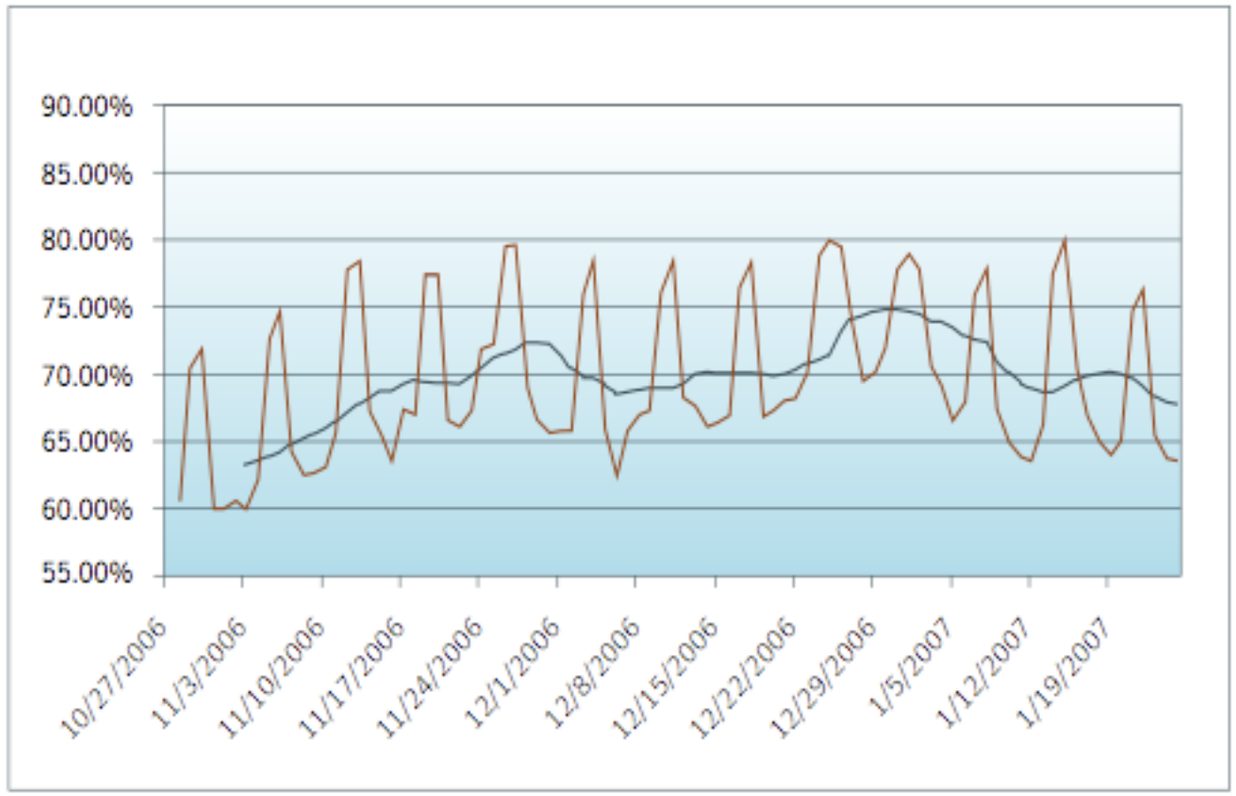
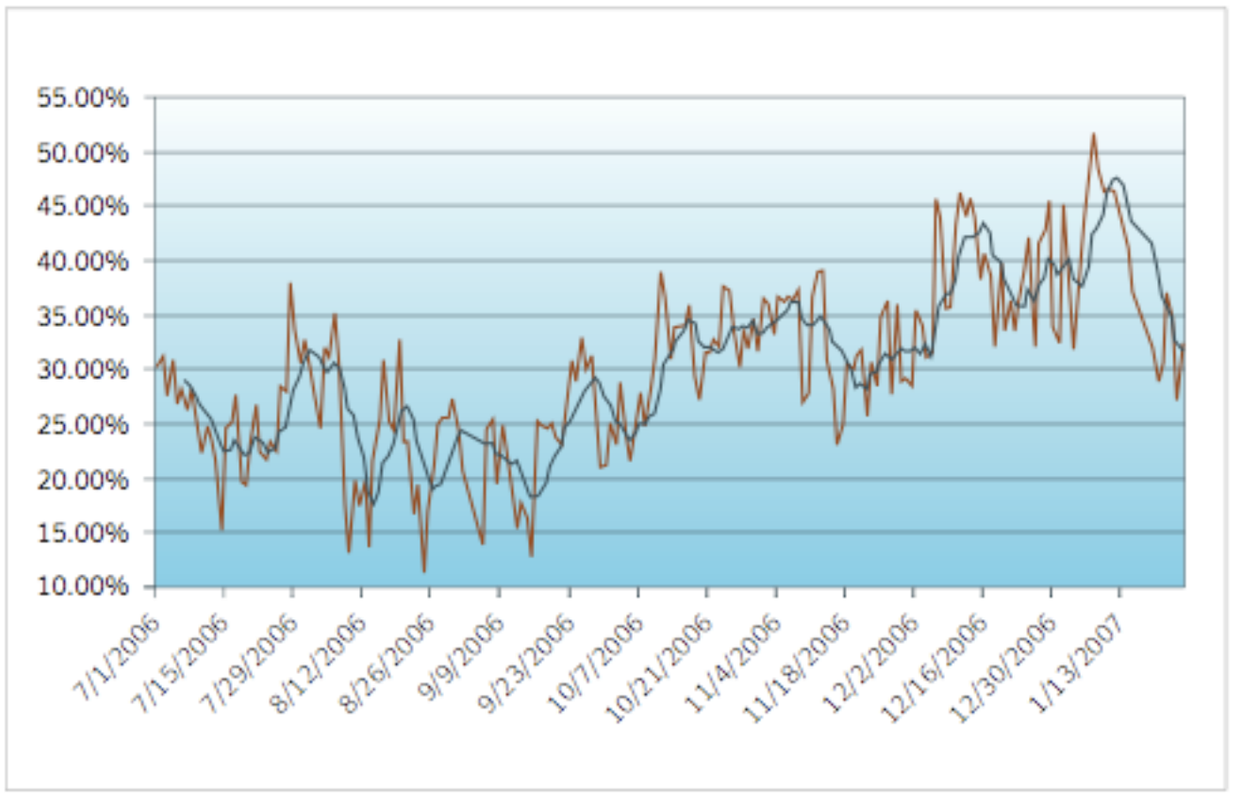


Image Spam Percentage



symantec The State of Spam: A Monthly Report --- February 2007

Internet Usage: USA

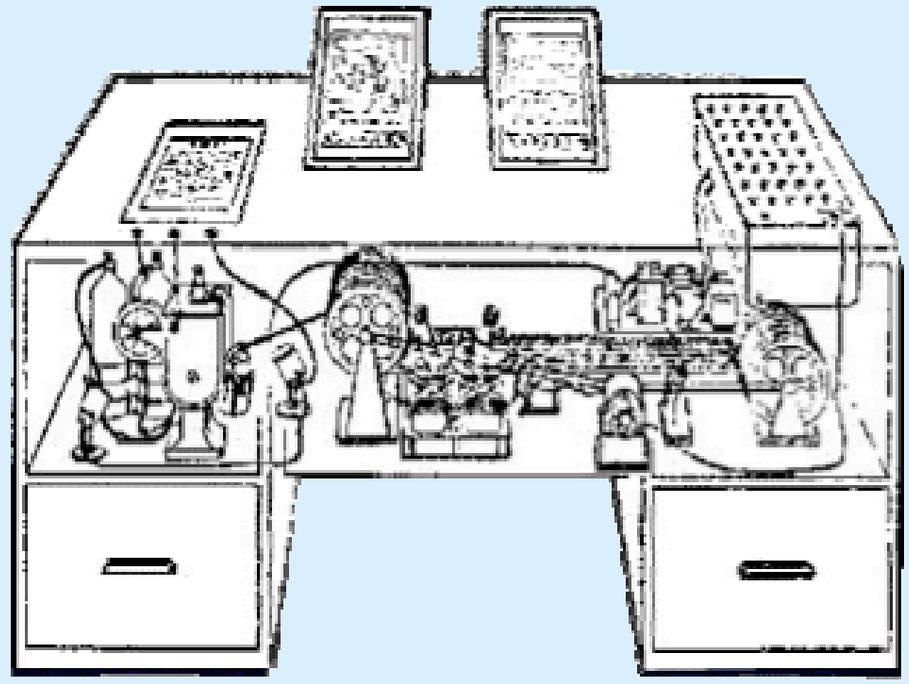
- 25h 25m home, 74h 26m work (per month)
- send email (52%)
- get news (32%)
- use a search engine to find info (29%)
- surf the web (23%)
- do research for work (19%)
- check weather (17%), IM (14%)

Vannevar Bush (1890-1974)

"As We May Think" *The Atlantic Monthly*, **July 1945**

http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm

- Memex
  All human knowledge
  in Memex
  "a billion books"
  hyper-linked together
- **Record everything you see**
  - **camera glasses**
  - **"a machine which types when talked to"**
- **Navigate by**
  text search
  following links
  associations.



- **Direct electrical path to human nervous system?**

29

Memex is Here! (or near)

- The Internet is growing fast.
- Most scientific literature is online somewhere.
  - it doubles every 10 years!
- Most literature is online (but copyrighted).
- Most Library of Congress visitors: web.
- A problem Bush anticipated:
  Finding answers is hard.

30

Why information moves to cyberspace.

- **Low rent: 10x cheaper**

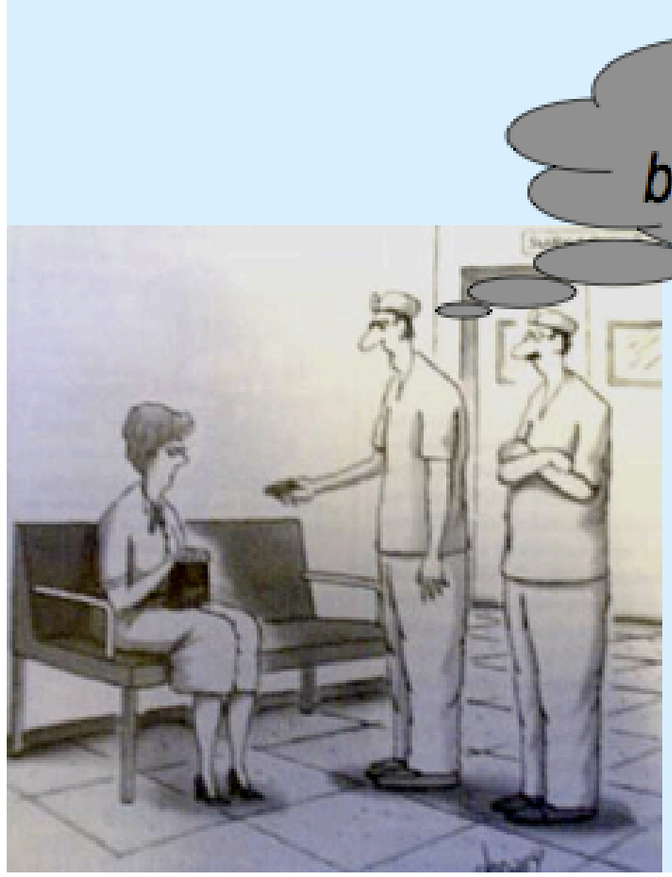
100 letters on disk:	10¢	in file cabinet	500¢
1 picture: on disk:	10¢	printed	40 ¢

- **Easy access and search:**
  - Robot can find all docs matching a predicate
  - Access from anywhere
  - Human costs 15$/hr

31

Personal Memex

6. Remember what is seen and heard
and quickly return any item on request.



Human input data		/hr	/lifetime
read	text	100 KB	25 GB
Hear	speech @ 10KBps	40 MB	10 TB
See	TV@ .5 MB/s	2 GB	8 PB

33

MyLifeBits

- A lifetime store of everything
- An experiment in lifetime storage
  - Gordon Bell

- A software research effort
  - Jim Gemmell and Roger Lueder
  - SQL server, recording tools, screensaver, mapped UI, SenseCam, ...
- Gorden Bell's SIGMOD 2005 Keynote

Summary

- Massive amounts of digital data
  - More are being created daily
- Data management and exploration
  - Storage: availability, scalability
  - Search: structured, semi-structured, unstructured
  - Exploration: mining, statistics, predicts
- Efficient systems design

Discussions (1)

- MyLifeBits: tries to store everything
  - storage is cheap
  - may turn out to be needed/important later
  - enhances associations, make search easier
- But,
  - waste of storage, useless data
  - slow down the whole system
  - how to filter, how to forget...

Discussions (2)

- MyLifeBits/Memex
  - personal store
- Why not share?
  - What to share?
    - data, annotations, classification, ...
  - How to share?
    - database integration, ...
  - Possible issues?

Discussions (3)

- File folders, databases, now what?
- Associations
- Annotations
  - ESP game
  - Peekaboom
  - Phetch