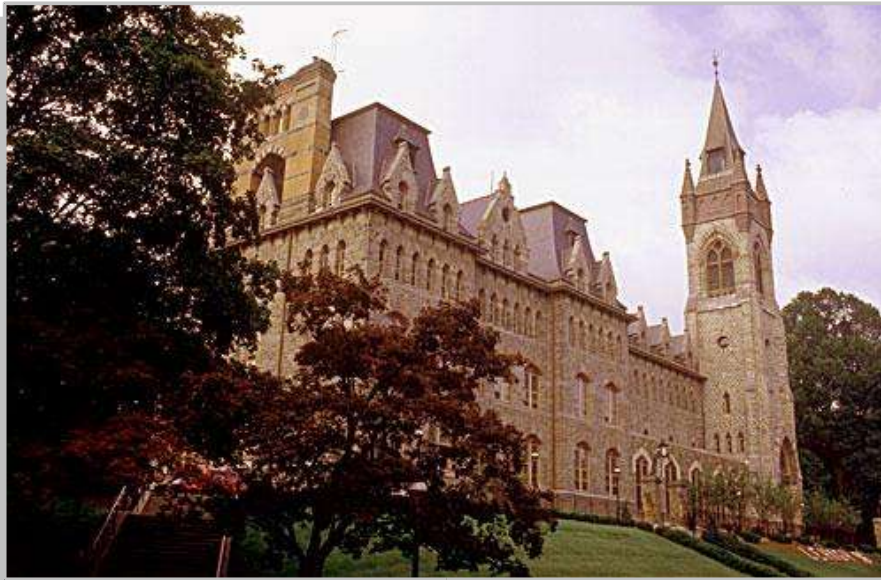


# Web Document Analysis: the Case of the Missing Dimension(s)



*Daniel Lopresti*

Computer Science & Engineering  
Lehigh University  
Bethlehem, PA 18015, USA  
[lopresti@cse.lehigh.edu](mailto:lopresti@cse.lehigh.edu)

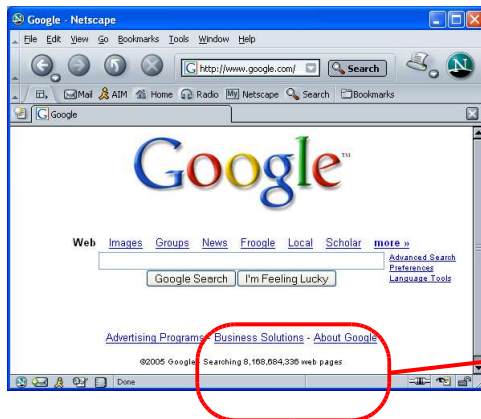


# Plan & Goals

- A talk that offers musings, suggestions, challenges, etc.
- Largely supported by intuition, less so by data.
- Hopefully provokes discussion, frames today's other talks.

Central question: What is the connection between Web documents and traditional document analysis?

Why care?



No, really, we're in pursuit of scientific excellence!!!

= \$\$\$ ... or £ or ¥ or ...

“Searching 8,168,684,336 web pages”

# Big Picture

- Documents are inherently 2-D (human view).
- Web pages are often processed as 1-D data stream (machine view).
- Goal is to capture content for whatever purpose. Success is ultimately judged based on human perceptions of relevance.

What's one obvious manifestation of this compromise?

“If people believe that they have searched the entire Internet when they run a search on a search engine, they are sadly mistaken – they are only seeing a subset of what is available.”

*Vint Cerf, Financial Times, 12/5/01*



# Big Picture View



“Invisible Web: Finding Hidden Content,” Diane Clark,  
[http://www.thealbertalibrary.ab.ca/netspeed/netspd2003/presentations/E2\\_Invisible\\_Web.ppt](http://www.thealbertalibrary.ab.ca/netspeed/netspd2003/presentations/E2_Invisible_Web.ppt)



# The Invisible Web

“Consists of material that general-purpose search engines either cannot or perhaps more importantly, will not include in their collections of webpages.”

*Chris Sherman, Search Engine Watch 2001*

*Author's  
biased view*

Why is there an invisible Web?

- Intentional exclusions: indexing policies, e.g., popularity.
- Financial considerations: user registration, pay for placement.
- Technical limitations: content contained in specialized databases, multimedia file formats.

“Invisible Web: Finding Hidden Content,” Diane Clark,  
[http://www.thealbertalibrary.ab.ca/netspeed/netspd2003/presentations/E2\\_Invisible\\_Web.ppt](http://www.thealbertalibrary.ab.ca/netspeed/netspd2003/presentations/E2_Invisible_Web.ppt)



# Types of Invisibility

- *Private Web*: password protected sites, “no robots” meta-tag.
- *Proprietary Web*: fee for use, free but registration required.
- *Opaque Web (also called “Gray Web”)*: pages excluded by crawl constraints, broken URL's, etc.
- *Invisible Web*: cannot be indexed for technical or other reasons.

Invisibility isn't necessarily all-bad and it may not be avoidable in every case, but it does seem to reflect missed opportunities.

“Invisible Web: Finding Hidden Content,” Diane Clark,  
[http://www.thealbertalibrary.ab.ca/netspeed/netspd2003/presentations/E2\\_Invisible\\_Web.ppt](http://www.thealbertalibrary.ab.ca/netspeed/netspd2003/presentations/E2_Invisible_Web.ppt)



# Rough Estimates

Claims by BrightPlanet, company aiming to deliver “deep” content:

## Visible or Surface Web:

- > 4 billion documents
- 19 terabytes (Tb)
- 100% publicly available
- Quality is often low

## Invisible or Deep Web:

- > 550 billion documents
- 7,500 terabytes (Tb)
- > 200,000 such sites exist
- 95% publicly available
- Quality 1,000x to 2,000x greater

BrightPlanet seems focused on data locked in databases.

Bright Planet, “The Deep Web: Surfacing Hidden Value,” Michael K. Bergman, September 2001.



# Document Analysis to the Rescue?

Certain Web documents are invisible because they can only be understood as 2-D entities.

Questions that seem most germane to us:

- How prevalent are different multimedia formats on the Web? How is this figure growing?
- How much of the information on Web pages is strongly 2-D?
- How much of the content on Web pages is locked in images?

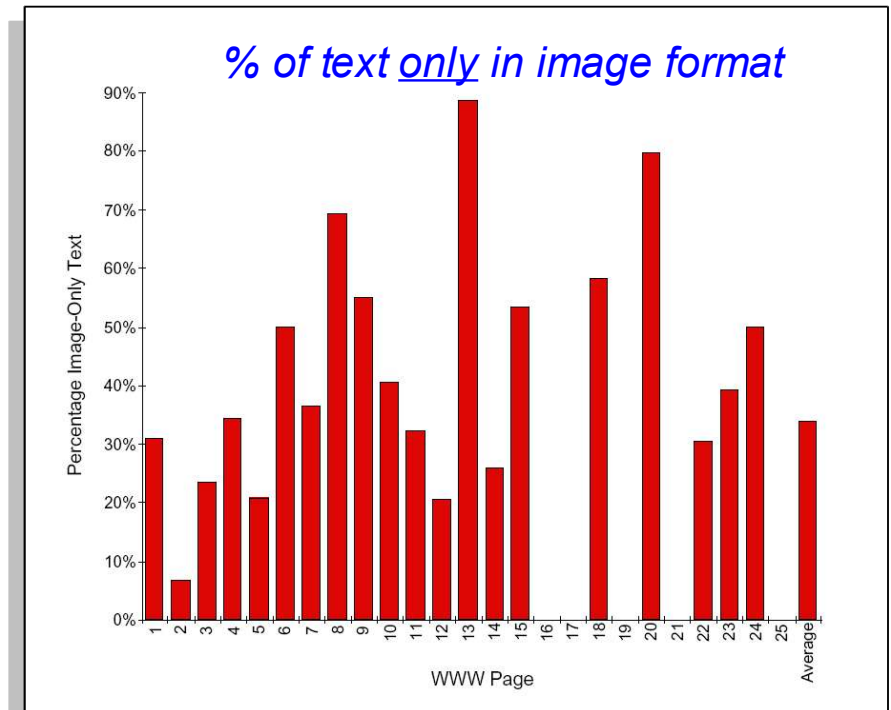
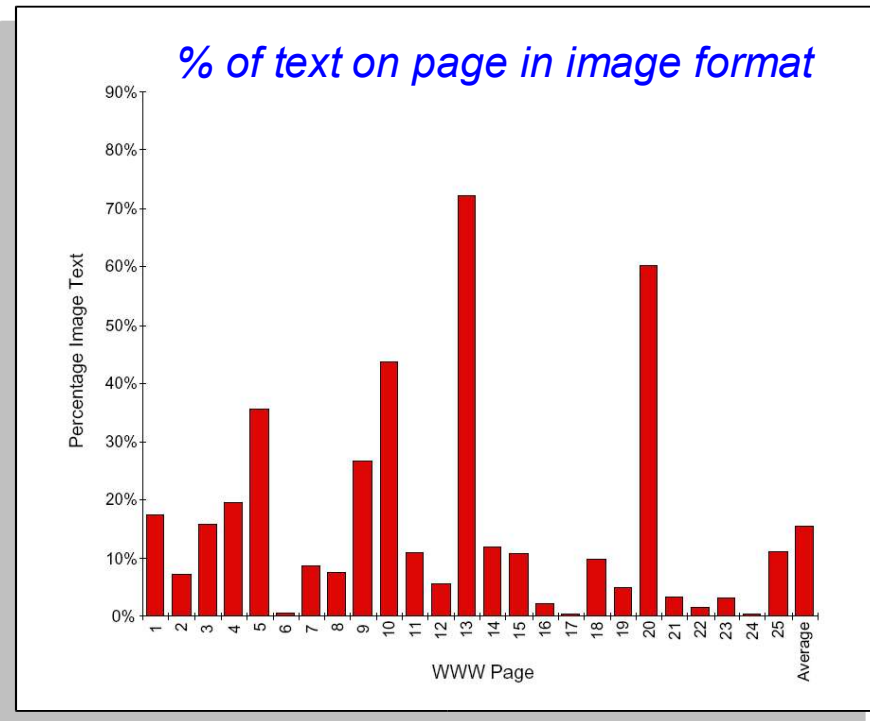
It would be nice to know current figures – something comparable to BrightPlanet analysis – but I had trouble finding any.





# Web Image Text

Some older data for two specific questions:



“Document Analysis and the World Wide Web,” D. Lopresti and J. Zhou, Proceedings of the IAPR Workshop on Document Analysis Systems, October 1996, Malvern, PA, pp. 651-671.

“Locating and Recognizing Text in WWW Images,” D. Lopresti and J. Zhou, Information Retrieval, vol. 2, nos. 2/3, May 2000, pp. 177-206.



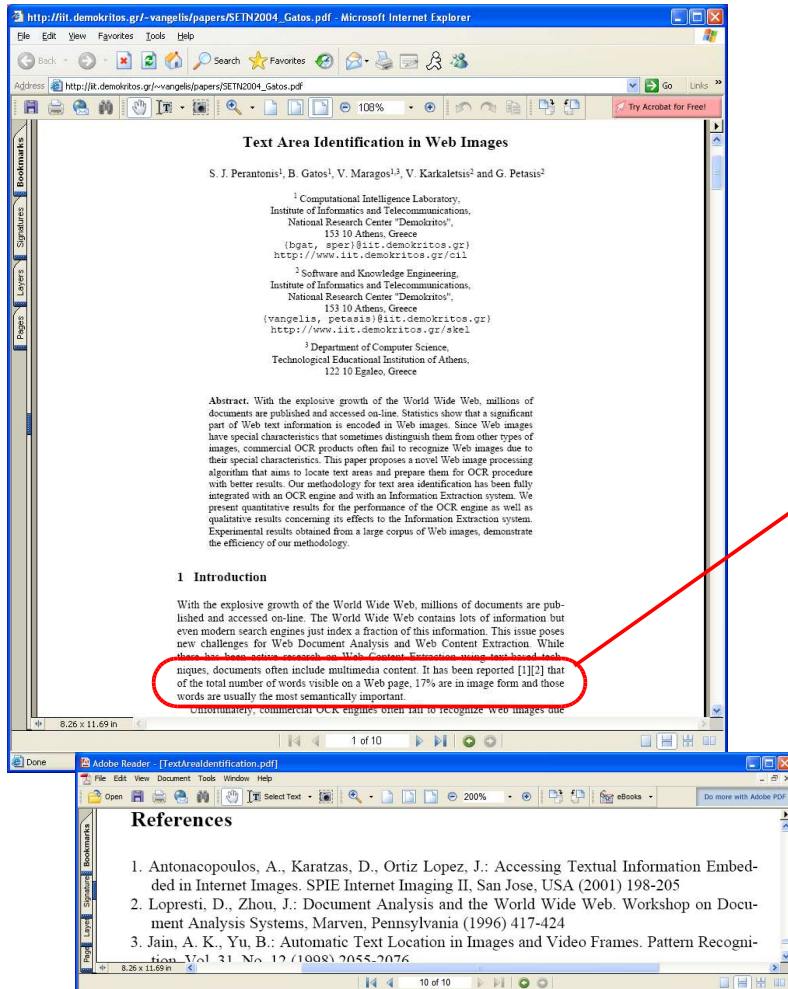
# Text in Web Images

A search for more current data didn't turn up much.

But from paper published in 2004:

“It has been reported [1][2] that of the total number of words visible on a Web page, 17% are in image form and those words are usually the most semantically important.”

“Text Area Identification in Web Images,” S. J. Perantonis, B. Gatos, V. Maragos, V. Karkaletsis and G. Petasis, Proceedings of the 3rd Hellenic Conference on Artificial Intelligence (SETN'04), Samos, Greece, May 5-8, 2004



# Simple Things That Don't Work

Tables are inherently 2-D, but HTML markup allows 1-D parsing.

The screenshot shows a Netscape browser window displaying the New York Mets sortable player stats page. The URL is `http://newyork.mets.mlb.com/NASApp/mlb/stats/sortable_player_stats.jsp?c_id=...`. The page title is "New York Mets : Sortable Player Stats". The browser's address bar and menu bar are visible. Below the browser window, a "Composer" window is open, showing the HTML source code of the page. The table header is highlighted with a red circle, and a red arrow points to it with the text "Note header misalignment".

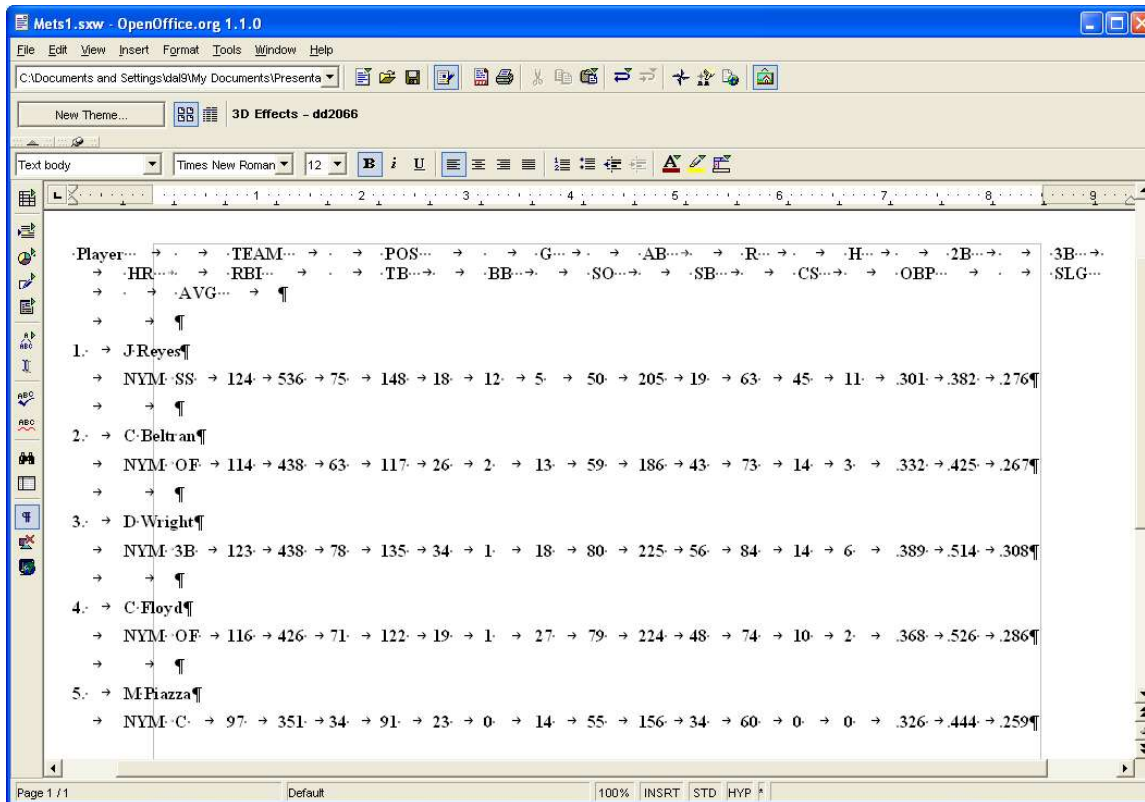
Player	TEAM	POS	G	AB	R	H	2B	3B	HR	RBI	TB	BB	SO	SB	CS	OBP	SLG	AVG	
<input type="checkbox"/>	1. J Reyes	NYM	SS	124	536	75	148	18	12	5	50	205	19	63	45	11	.301	382	276
<input type="checkbox"/>	2. C Beltran	NYM	OF	114	438	63	117	26	2	13	59	186	43	73	14	3	.332	425	267
<input type="checkbox"/>	3. D Wright	NYM	3B	123	438	78	135	34	1	18	80	225	56	84	14	6	.389	514	308
<input type="checkbox"/>	4. C Floyd	NYM	OF	116	426	71	122	19	1	27	79	224	48	74	10	2	.368	526	286
<input type="checkbox"/>	5. M Piazza	NYM	C	97	351	34	91	23	0	14	55	156	34	60	0	0	.326	444	259

Note header misalignment



# Try Pasting as “Text-Only”

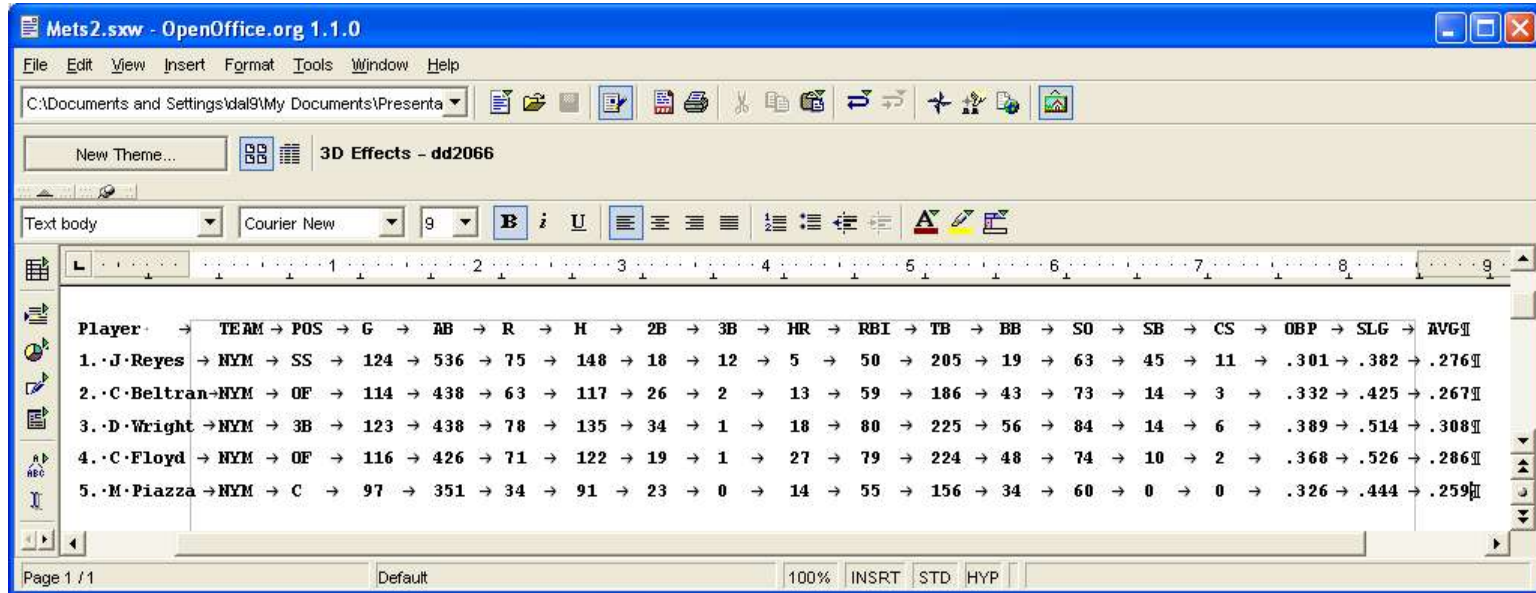
Sometimes a good word processor can save the day ...



... but even ignoring bad line breaks, we see spurious tabs and spaces, stuff that doesn't line up right, etc.



# After Much Manual Editing



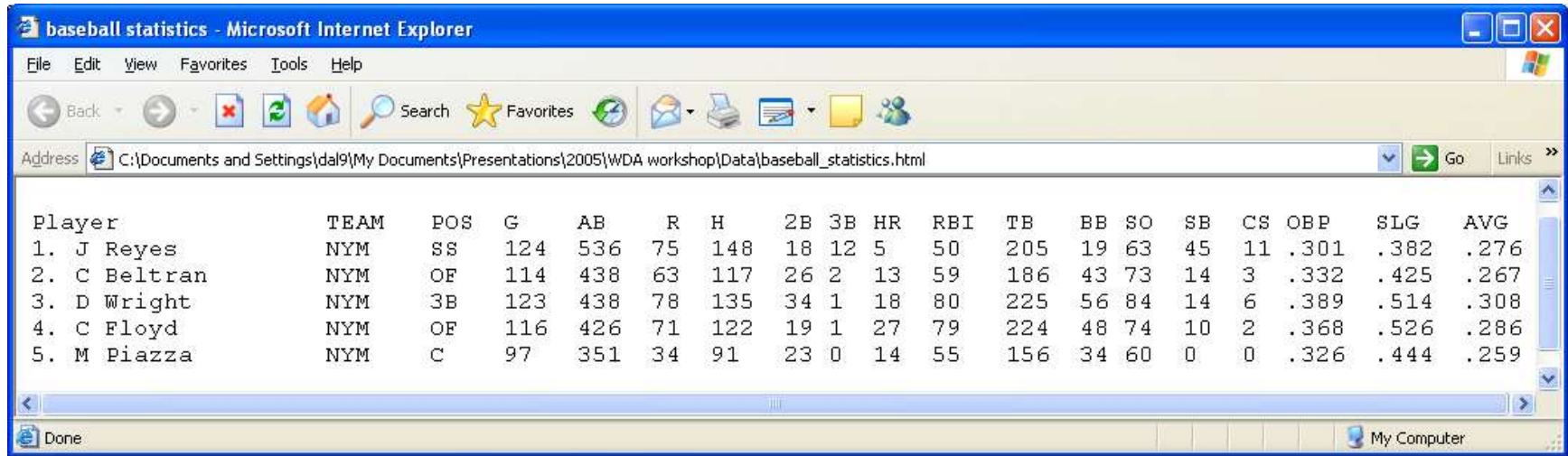
The screenshot shows the OpenOffice.org 1.1.0 interface with a table of baseball statistics. The table has 17 columns: Player, TEAM, POS, G, AB, R, H, 2B, 3B, HR, RBI, TB, BB, SO, SB, CS, OBP, SLG, and AVG. The data is as follows:

Player	TEAM	POS	G	AB	R	H	2B	3B	HR	RBI	TB	BB	SO	SB	CS	OBP	SLG	AVG
1. J. Reyes	NYM	SS	124	536	75	148	18	12	5	50	205	19	63	45	11	.301	.382	.276
2. C. Beltran	NYM	OF	114	438	63	117	26	2	13	59	186	43	73	14	3	.332	.425	.267
3. D. Wright	NYM	3B	123	438	78	135	34	1	18	80	225	56	84	14	6	.389	.514	.308
4. C. Floyd	NYM	OF	116	426	71	122	19	1	27	79	224	48	74	10	2	.368	.526	.286
5. M. Piazza	NYM	C	97	351	34	91	23	0	14	55	156	34	60	0	0	.326	.444	.259

- What I wanted from the start, but cost me about 10 minutes.
- Why should this be so hard?
- Tables look great on-screen, but browser doesn't truly understand their 2-D nature.

# What's the difference?

Table 1



baseball statistics - Microsoft Internet Explorer

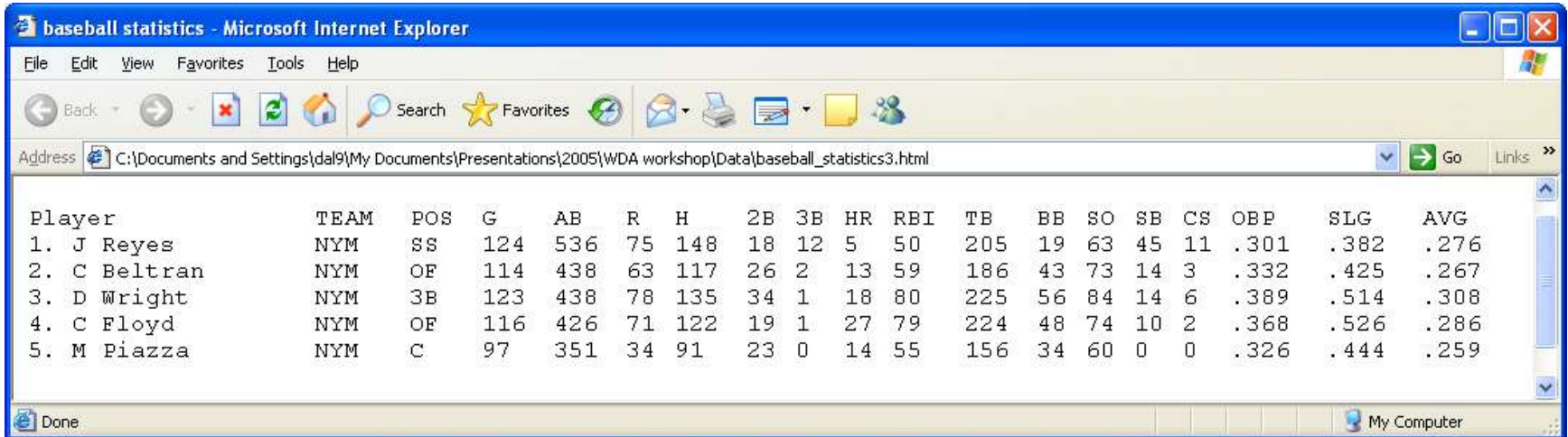
File Edit View Favorites Tools Help

Address C:\Documents and Settings\dal9\My Documents\Ppresentations\2005\WDA workshop\Data\baseball\_statistics.html

Player	TEAM	POS	G	AB	R	H	2B	3B	HR	RBI	TB	BB	SO	SB	CS	OBP	SLG	AVG
1. J Reyes	NYM	SS	124	536	75	148	18	12	5	50	205	19	63	45	11	.301	.382	.276
2. C Beltran	NYM	OF	114	438	63	117	26	2	13	59	186	43	73	14	3	.332	.425	.267
3. D Wright	NYM	3B	123	438	78	135	34	1	18	80	225	56	84	14	6	.389	.514	.308
4. C Floyd	NYM	OF	116	426	71	122	19	1	27	79	224	48	74	10	2	.368	.526	.286
5. M Piazza	NYM	C	97	351	34	91	23	0	14	55	156	34	60	0	0	.326	.444	.259

Done My Computer

Table 2



baseball statistics - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address C:\Documents and Settings\dal9\My Documents\Ppresentations\2005\WDA workshop\Data\baseball\_statistics3.html

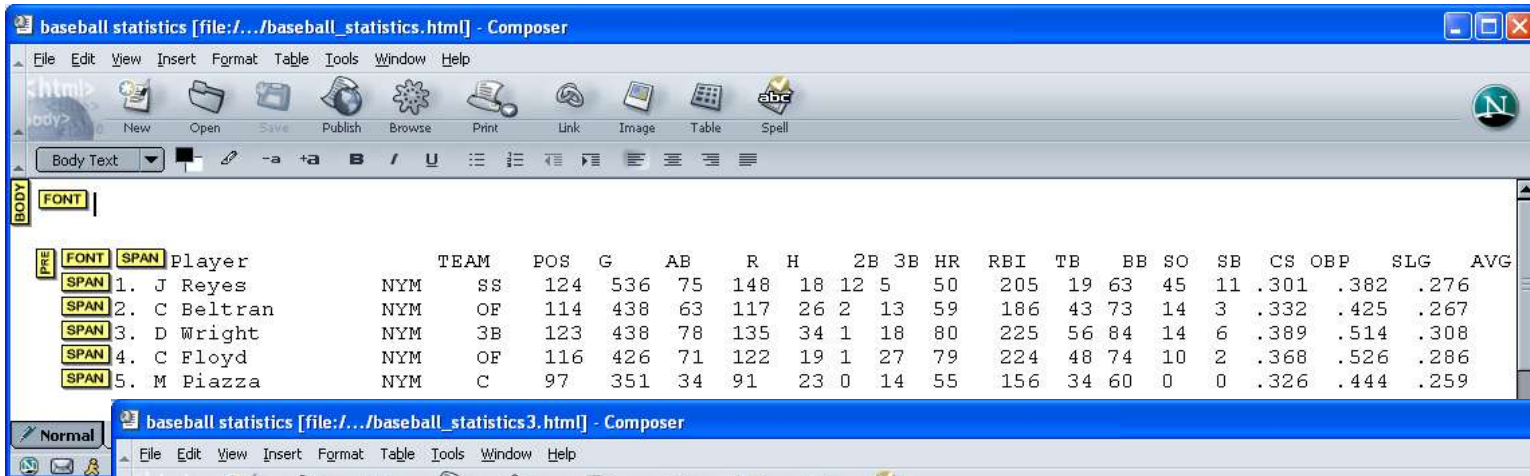
Player	TEAM	POS	G	AB	R	H	2B	3B	HR	RBI	TB	BB	SO	SB	CS	OBP	SLG	AVG
1. J Reyes	NYM	SS	124	536	75	148	18	12	5	50	205	19	63	45	11	.301	.382	.276
2. C Beltran	NYM	OF	114	438	63	117	26	2	13	59	186	43	73	14	3	.332	.425	.267
3. D Wright	NYM	3B	123	438	78	135	34	1	18	80	225	56	84	14	6	.389	.514	.308
4. C Floyd	NYM	OF	116	426	71	122	19	1	27	79	224	48	74	10	2	.368	.526	.286
5. M Piazza	NYM	C	97	351	34	91	23	0	14	55	156	34	60	0	0	.326	.444	.259

Done My Computer

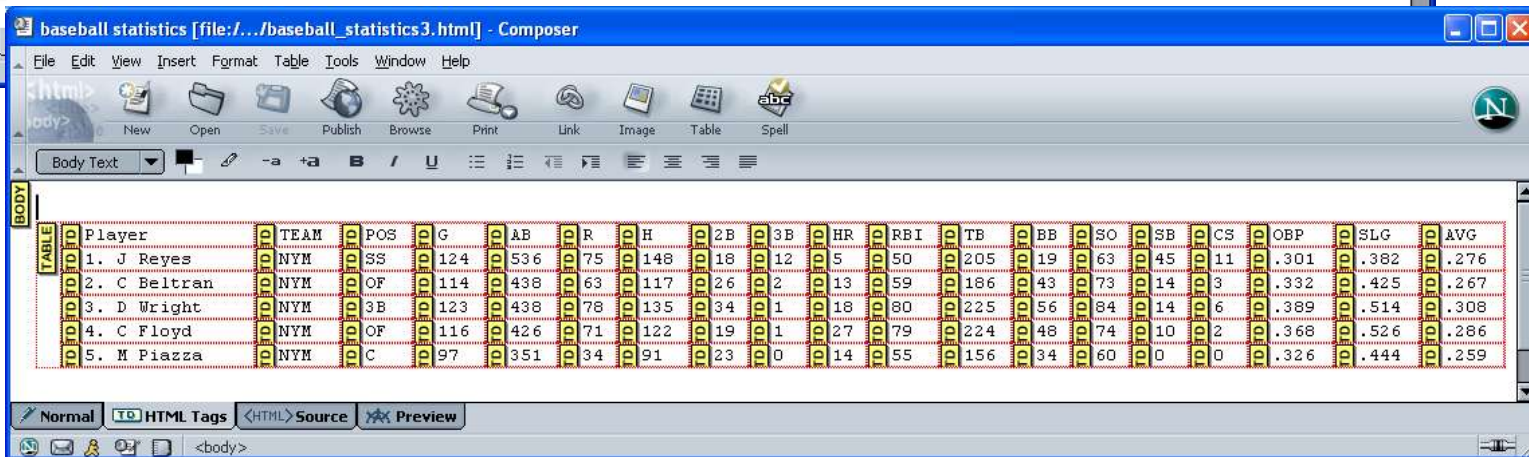


# Examine HTML Tags

Not a Table



Table



They look identical to user, but browser doesn't know first is a table.



# More Table Problems

The converse is use of the HTML <table> tag for non-table layouts.

The screenshot shows a web browser window with the URL <http://www.crmdaily.com/stocks/>. The page title is "CRMDaily: Real-time CRM Industry News from Around the World - Microsoft Internet Explorer". The browser's address bar shows the URL. The page content includes a navigation menu, a "Section Updated Today" section, and a "The CRMDaily Stock Snapshot" table. The table lists various companies and their stock performance. A red circle labeled "Non-genuine Table" points to the "The CRMDaily Stock Snapshot" table. Another red circle labeled "Genuine Table" points to a table below it, which is a "MARKET WATCH" table.

Company	Last	Change	52 Week High/Low	Market Cap
APAC (APAC)	2.91	-0.04	9.37 2.45	\$141 Mil
Applix (APLX)	1.40	-0.11	6.31 1.32	\$15 Mil
Blue Martini (BLUE)	1.50	-0.05	77.62 1.45	\$95 Mil
BroadVision (BVISN)	3.12	-0.22	39.75 2.50	\$853 Mil
Calico (CLIC)	0.14	0.00	10.75 0.10	\$4 Mil
Convergys (CVG)	30.00	-0.68	52.25 25.14	\$5,126 Mil
Davox (DAVX)	9.00	+0.05	14.87 6.12	\$115 Mil
Delano (DTEC)	0.17	-0.04	17.75 0.17	\$6 Mil
eGain (EGAN)	1.85	-0.01	14.31 1.53	\$65 Mil
eLoyalty (ELOY)	0.46	-0.02	16.18 0.40	\$23 Mil

Work on HTML table detection by Wang and Hu:

	L	T	LT	LTW-VS	LTW-NB	LTW-KNN
R (%)	87.24	90.80	94.20	94.25	95.46	89.60
P (%)	88.15	95.70	97.27	97.50	94.64	95.94
F (%)	87.70	93.25	95.73	95.88	95.05	92.77

L: Layout features only.

T: Content type features only.

LT: Layout and content type features.

LTW-VS: Layout, content type and vector space based word group features.

LTW-NB: Layout, content type and naive Bayes based word group features.

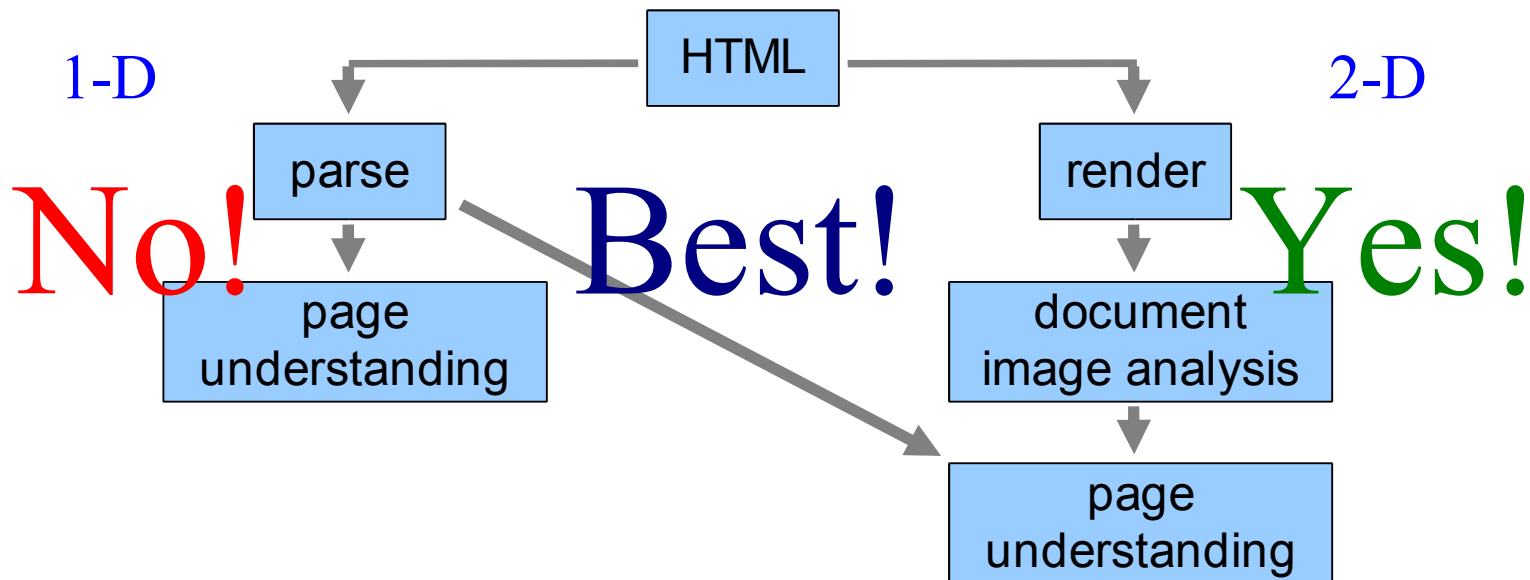
LTW-KNN: Layout, content type and kNN based word group features.

“Detecting Tables in HTML Documents,” Y. Wang and J. Hu, Proceedings of the IAPR Workshop on Document Analysis Systems, August 2002, Princeton, NJ.



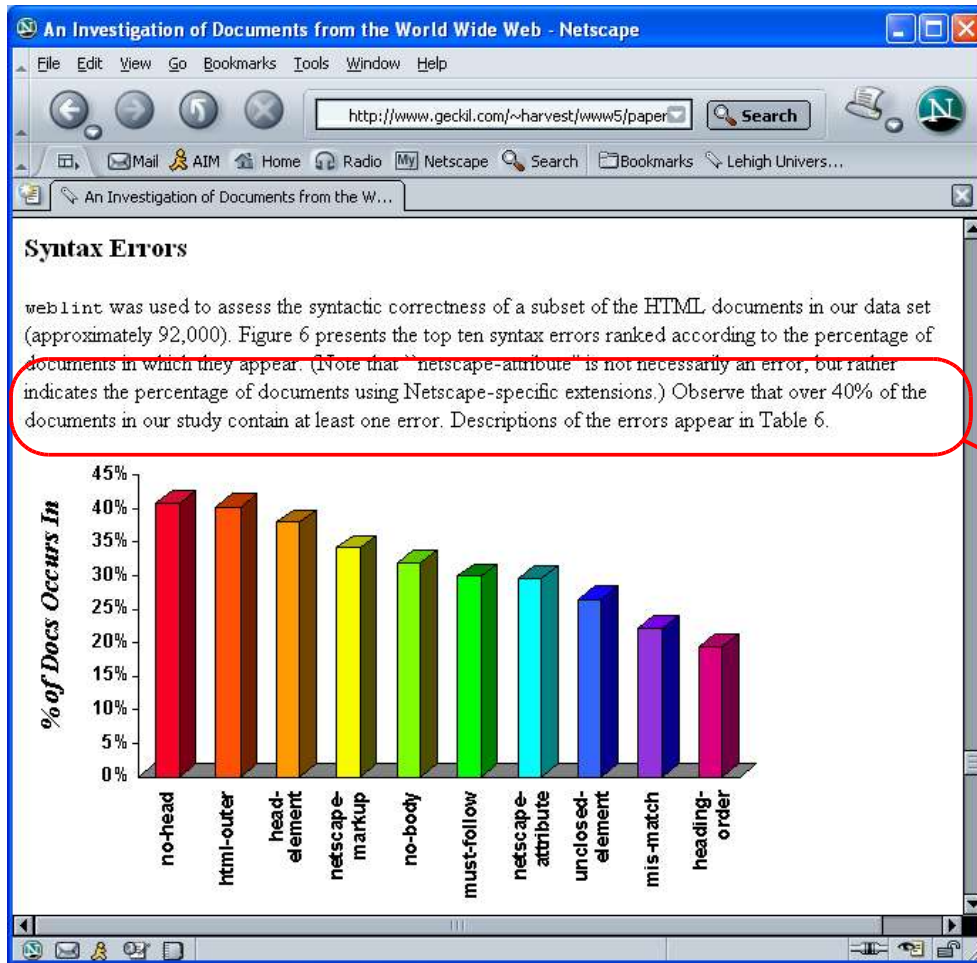
# Moral

It is a table if and only if it looks like a table.



- Counter-arguments:
- “Parsing is easier and usually works.”
  - “What about the Semantic Web?”

# Bad HTML



The world isn't perfect: browsers must always be prepared to deal errors in their input.

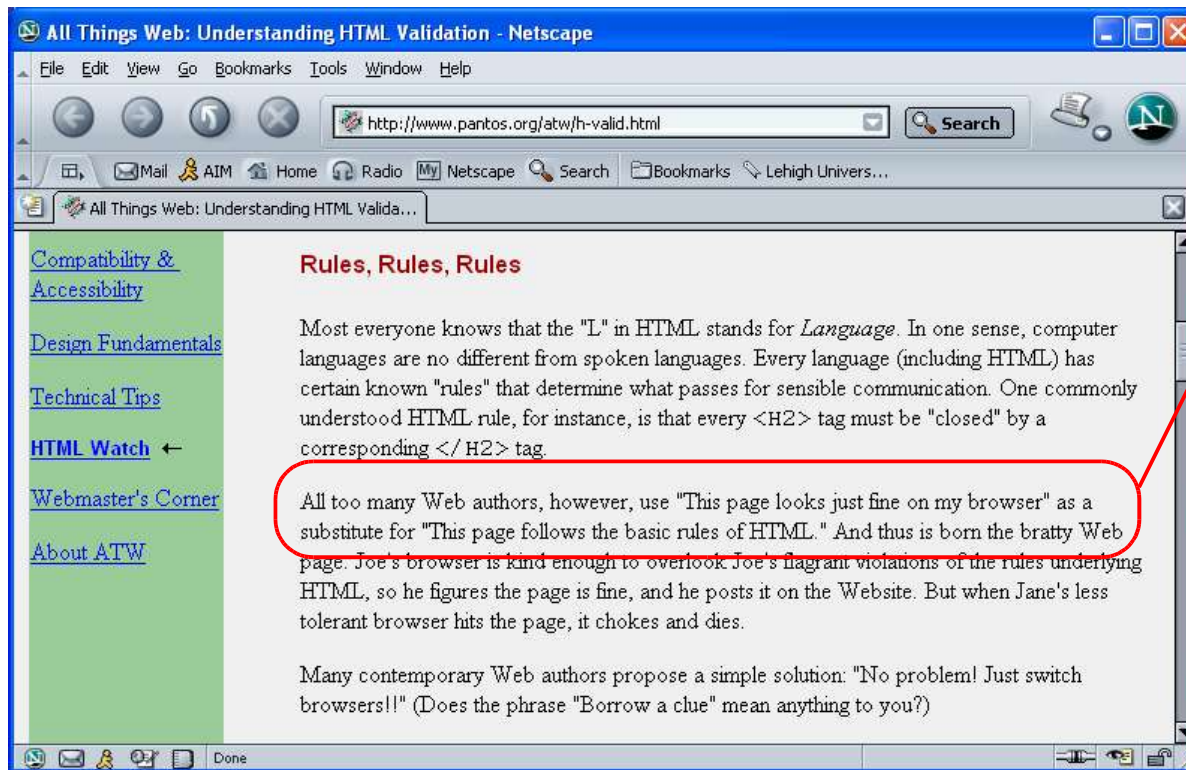
“... over 40% of the documents ... contain at least one error ...”

“An Investigation of Documents from the World Wide Web” Allison Woodruff, Paul M. Aoki, Eric Brewer, Paul Gauthier, Lawrence A. Rowe

<http://www.geckil.com/~harvest/www5/papers/P7/Overview.html>

# Bad HTML

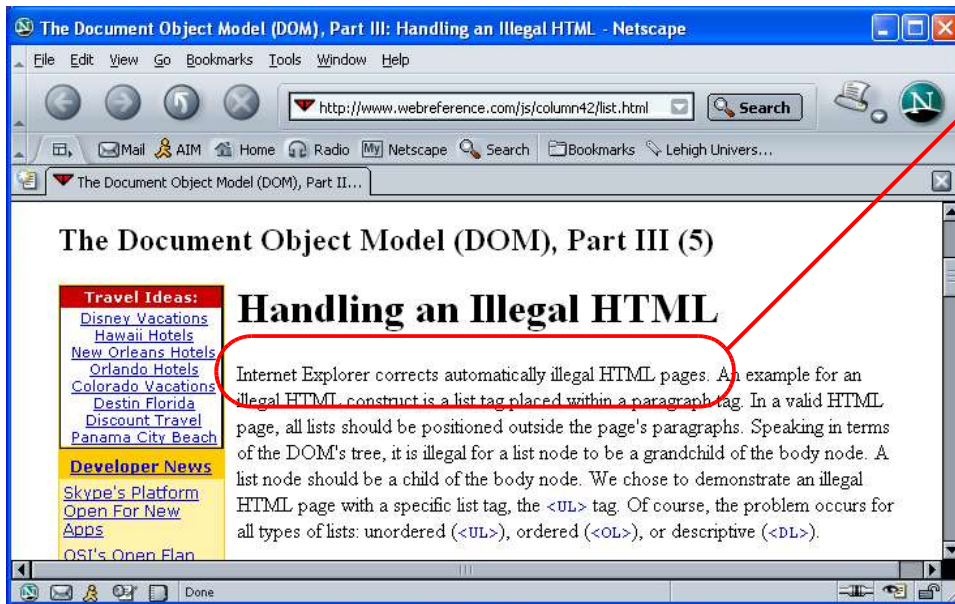
How a browser deals with errors, not the formal definition of the markup language, determines what a user sees.



“... many Web authors ... use “This page looks just fine on my browser” as a substitute for “This page follows the basic rules of HTML.” ”

# Bad HTML

What happens when people break the rules?



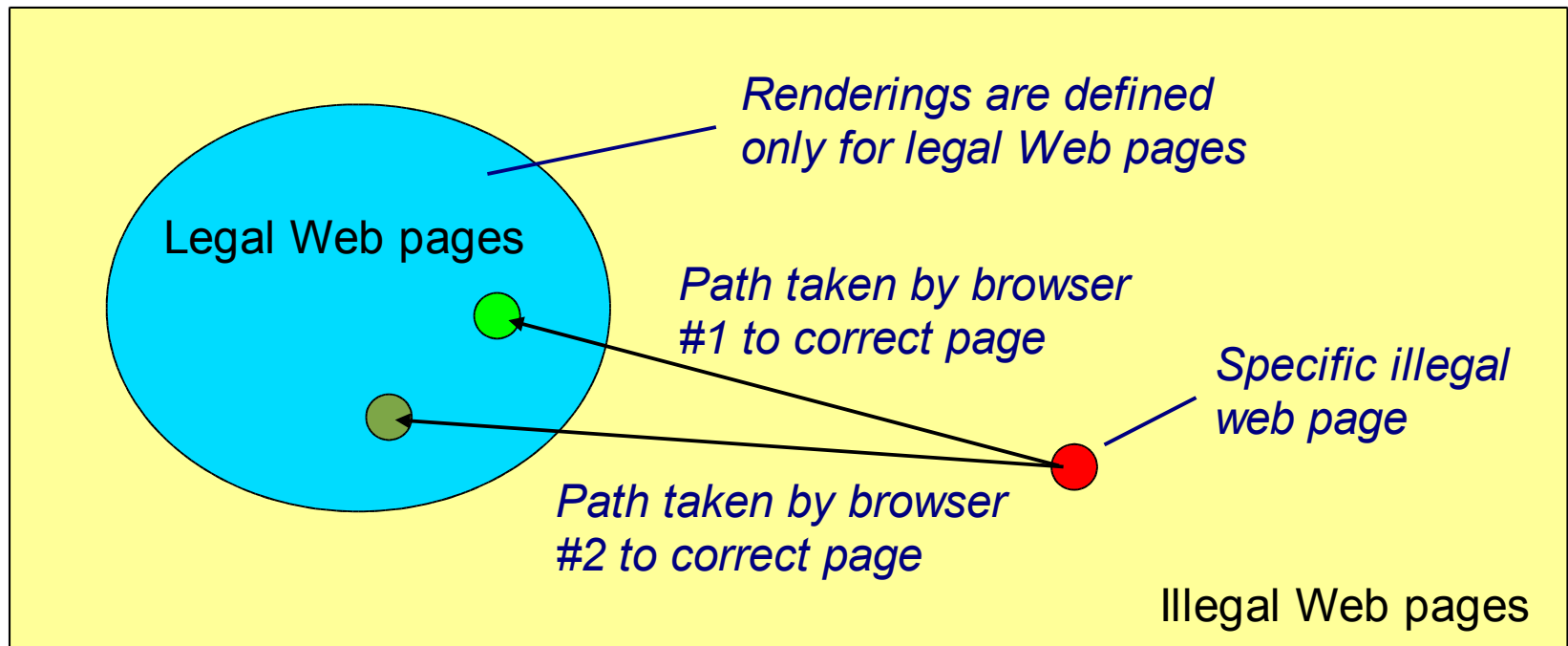
“Internet Explorer corrects automatically illegal HTML pages.”

What does “correct” mean?  
How could it possibly know the author's intent?

It can't. It's just making some sort of guess.

# Bad HTML

Universe of all possible Web pages



Corollary: you can't know what a Web page is going to look like until you render it in a specific browser.

# Midpoint Summary

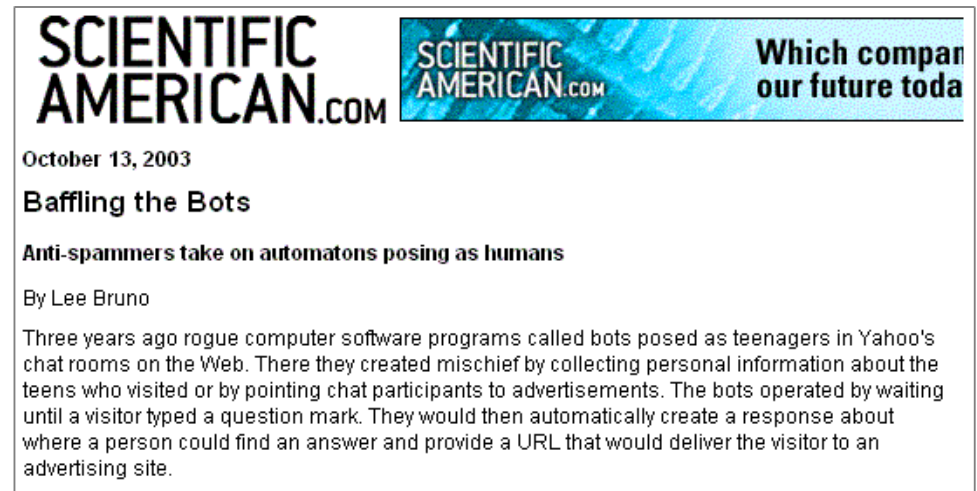
- While parsing HTML as a 1-D data stream buys you a lot, it would be better to combine this with a 2-D analysis of the rendered image.
- Motivation comes not only from need to recognize image text on the Web, but also to process physical / logical structure of tables and to handle malformed HTML.
- This is good news for us as document analysis researchers!



# Protecting Web Services

Internet has become vehicle for distributing valuable content. Malicious programs (“bots”) attempt to exploit online services intended for human users.

Idea: create a pattern recognition task easy for humans to solve but hard for machines.



**SCIENTIFIC AMERICAN.COM** **SCIENTIFIC AMERICAN.COM** Which compares our future today

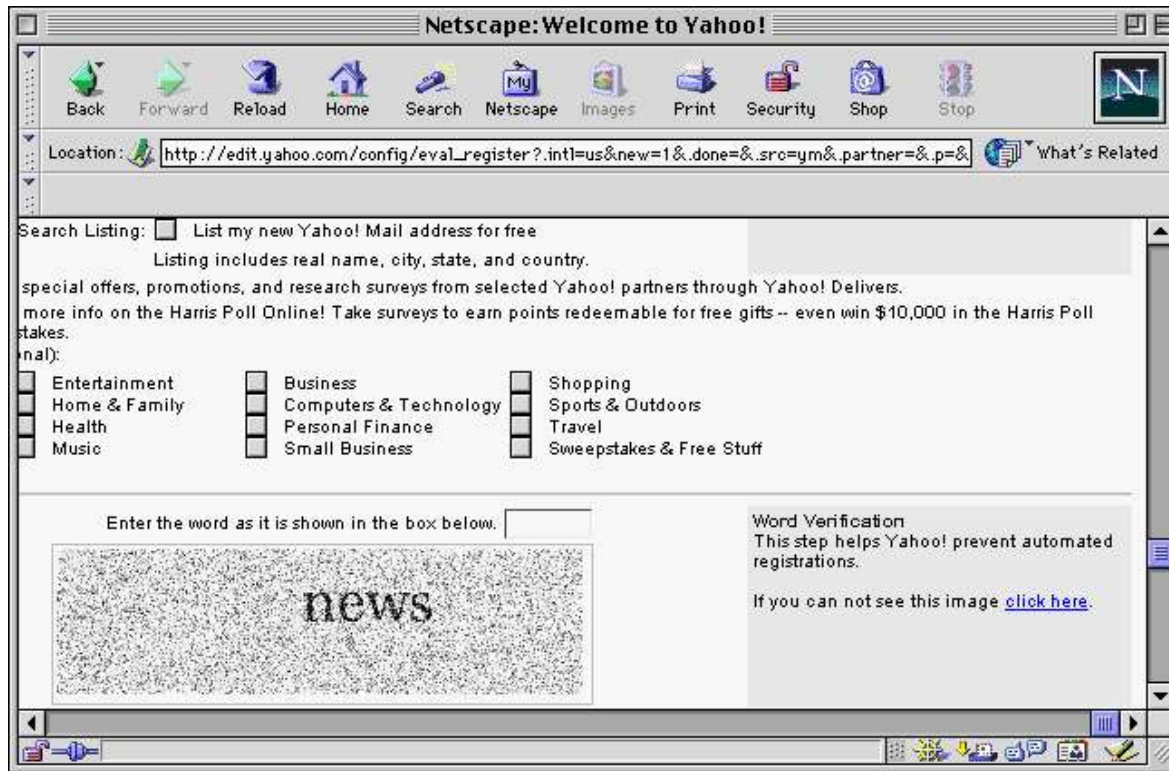
October 13, 2003  
**Baffling the Bots**  
**Anti-spammers take on automatons posing as humans**  
By Lee Bruno

Three years ago rogue computer software programs called bots posed as teenagers in Yahoo's chat rooms on the Web. There they created mischief by collecting personal information about the teens who visited or by pointing chat participants to advertisements. The bots operated by waiting until a visitor typed a question mark. They would then automatically create a response about where a person could find an answer and provide a URL that would deliver the visitor to an advertising site.



# Protecting Web Services

Yahoo's method for protecting free email service. User must solve simple character recognition task:

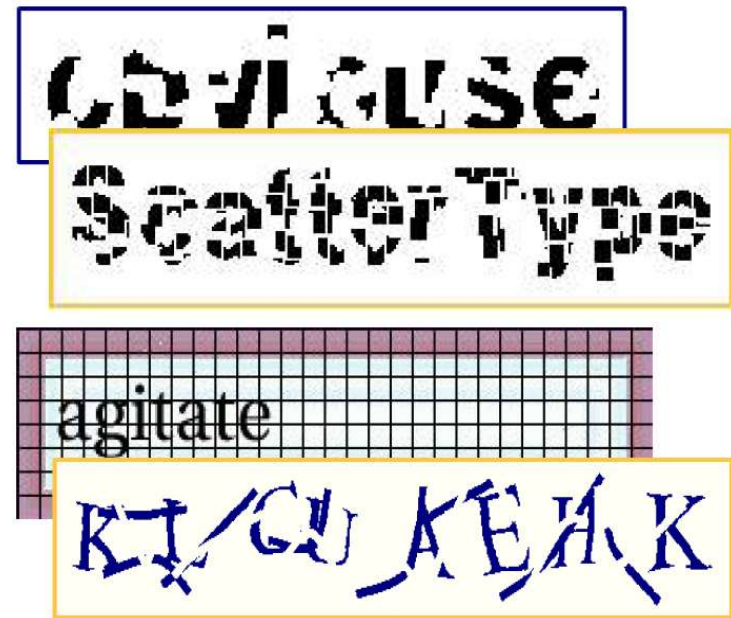


*CAPTCHA* =  
“Completely  
Automated  
Public Test to tell  
Computers and  
Humans Apart”



# If You Can't Join 'em, Beat 'em

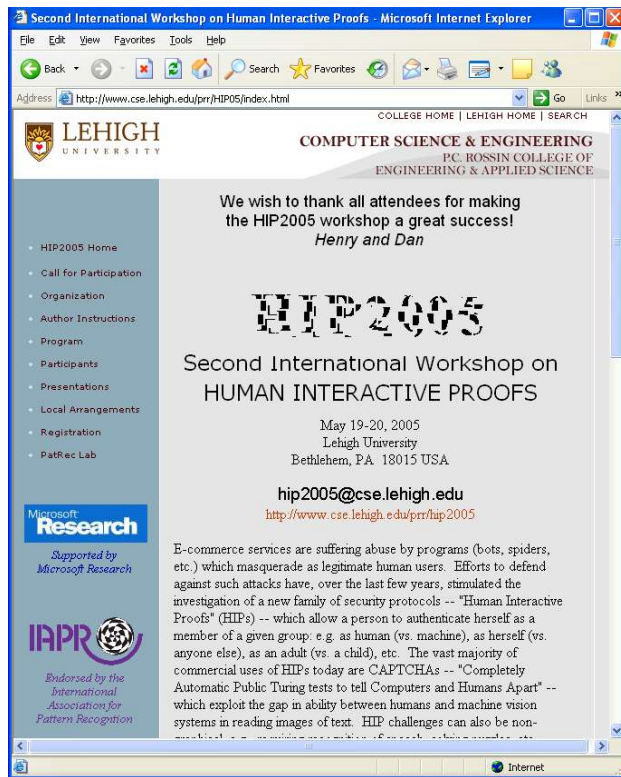
This turns the problem around 180 degrees: we want to create Web pages that computers find impossible to analyze.



Note: Henry Baird and some of his students are actively working on ScatterType and seek volunteers to attack it. Visit <http://arcturus.cse.lehigh.edu/CAPTCHAs>

# Lehigh HIP Workshop, May 2005

## Web page



<http://www.cse.lehigh.edu/pr/HIP05/index.html>

## Proceedings



# CAPTCHA's

Again, goal is to prevent automated attacks on Web services:

- Exploits observation that humans are still much better than computers at many pattern recognition tasks.
- Paradigm is variant of well known Turing Test.

The two criteria that matter most:

- Is test effective at keeping out machines?
- Is test tolerable to humans?

Implications:

- Need very large supply of different challenges.
- Must be cognizant of human reaction to CAPTCHA's.



# Points to Ponder

- Machines won't stay stupid forever. Range of problems they can solve is growing reasonably rapidly – it certainly isn't shrinking.
- Humans evolve at a more modest pace. What does this suggest about our ability to assimilate new pattern recognition tasks?

While today there are a large number of generative CAPTCHA's to choose from, someday we may run out of tests that meet both criteria (hard for machines, tolerable to humans). Should we be concerned?

Cause for hope: variety of pattern recognition tasks in real world is almost endless. Note apparent disconnect, however, between natural tasks and synthetic versions we use for CAPTCHA's.



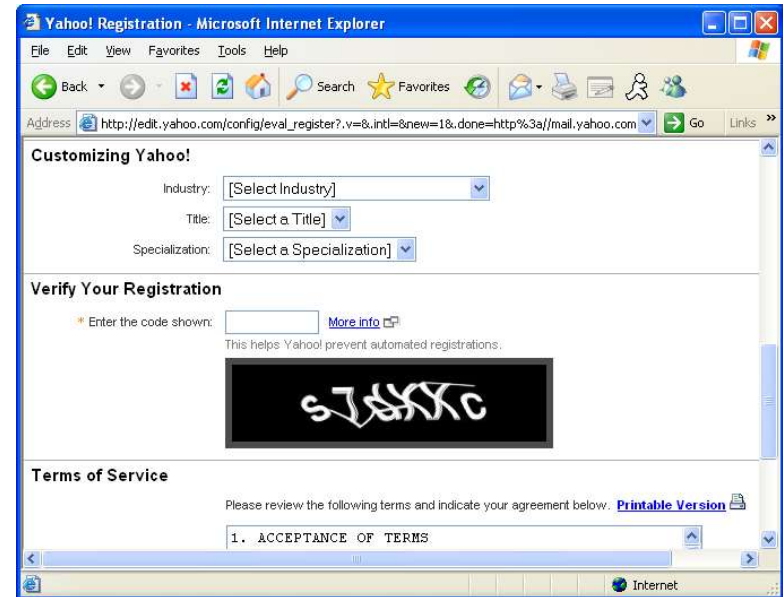
# Natural vs. Synthetic #1

of virtuous and enlightened men to clip the  
inhabitants of Harrisburgh among this number  
is only to bear testimony to the zealous and efficient  
exertions which they have made towards the defence  
of the Law. *Washington*  
Octo. 4. 1774.

testimony

What word do you see in the box?

*George Washington Papers at the Library of Congress*  
<http://memory.loc.gov/ammem/gwhtml/gwhome.html>



What "word" do you see in the box?

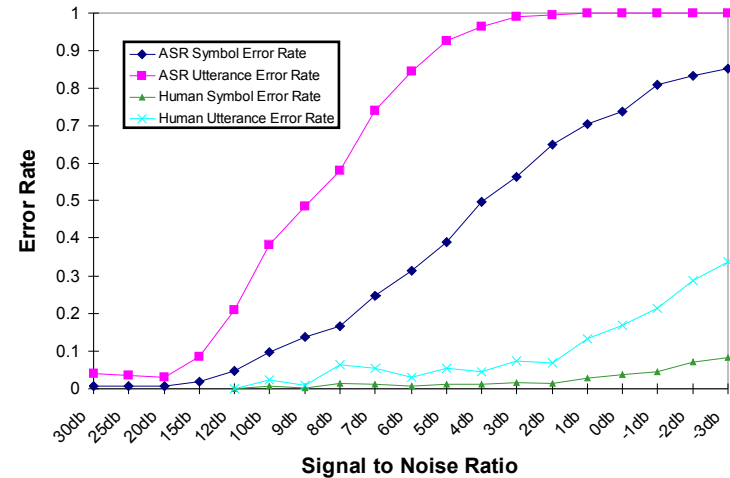
*Yahoo! free email account registration page*  
<http://mail.yahoo.com/>

# Natural vs. Synthetic #2



What is Bobby Thomson's average?

*"The Shot Heard Round the World," Russ Hodges*  
[http://www.baseballhalloffame.org/exhibits/online\\_exhibits/1951/1951\\_story.htm](http://www.baseballhalloffame.org/exhibits/online_exhibits/1951/1951_story.htm)



What number do you hear spoken?

*"Human Interactive Proofs for Spoken Language Interfaces,"*  
D. Lopresti, C. Shih, and G. Kochanski, *Workshop on Human Interactive Proofs, January 2002, Palo Alto, CA*  
<http://www.cse.lehigh.edu/~lopresti/Publications/2002/hip02.pdf>

# What's Fundamental Here?

Recalling two primary criteria, two secondary criteria are:

- Is test easy to generate?
- Is test easy to grade?

These don't seem as fundamental as criteria listed earlier:

- In deploying CAPTCHA's, all we require is very large supply of different tests. No one said we have to generate them ourselves.
- Likewise, no one said we have to grade them ourselves if we can get someone else knowledgeable (and trustworthy) to do it.



# The Case for Natural CAPTCHA's

Range of pattern recognition tasks we face every day is far greater than what has been fielded as CAPTCHA's so far.

Above statement remains true even if we confine our attention to what's available on the Internet.

Might humans be more accepting of natural tasks – ones we have had lots of experience with – than synthetic ones?

An alternative to generating CAPTCHA's: harvest them.





# Where Do CAPTCHA's Grow?

Describe the weather  
in this scene.

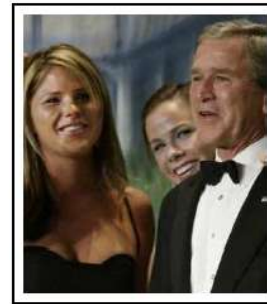
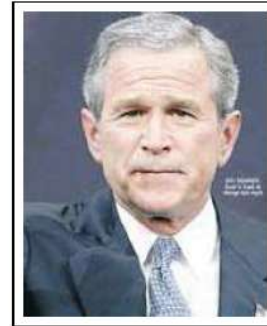
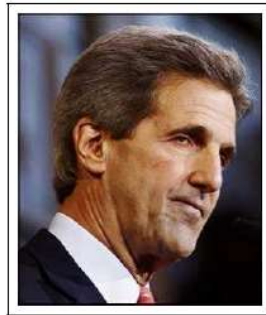


From WABC Central Park WebCam, [http://abclocal.go.com/kabc/features/cams/082102\\_central\\_Park\\_cam.html](http://abclocal.go.com/kabc/features/cams/082102_central_Park_cam.html)



# Where Do CAPTCHA's Grow?

Which photos show the same person?



# Where Do CAPTCHA's Grow?

How many cars do you see in this image?

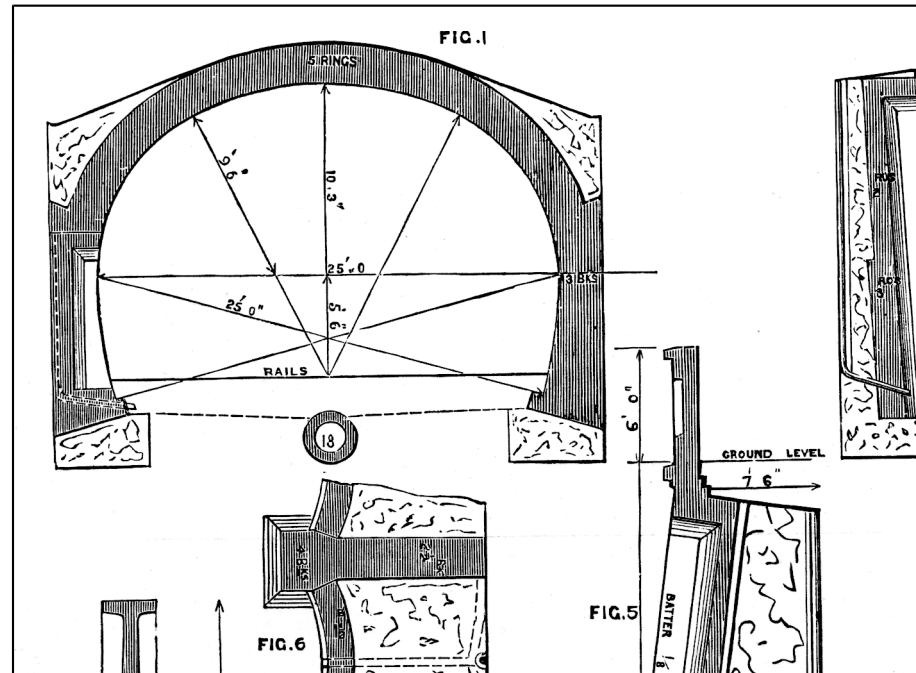


From WCPO Cincinnati Ohio Skycam, [http://webcambiglook.com/cinn\\\_skycam.html](http://webcambiglook.com/cinn\_skycam.html)



# Where Do CAPTCHA's Grow?

Draw a box around a text string in this image.

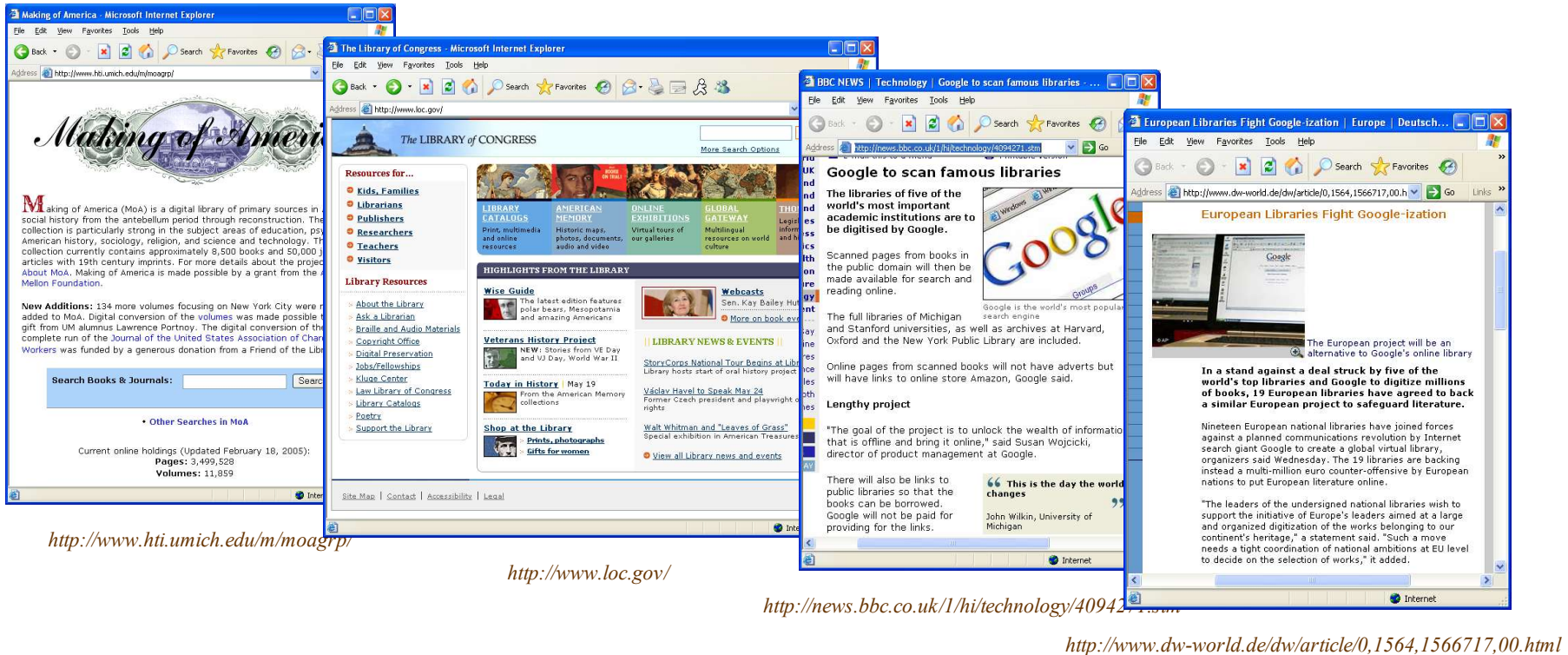


From the Lehigh University Library Digital Bridges project, <http://bridges.lib.lehigh.edu/>



# Where Do CAPTCHA's Grow?

One obvious answer: in digital libraries.



Google's project alone totals an estimated 4.5 billion pages.





# Something is Missing

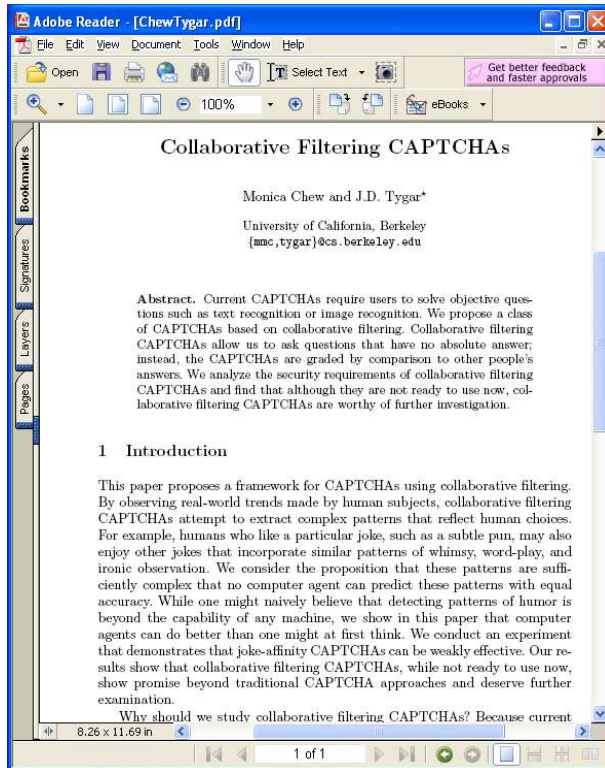
To grade response to a CAPTCHA challenge, we need to know “correct” answer (or, rather, how a human would respond).

- Google is scanning books with intention of making them searchable online, of course. Hence, we might expect a textual transcription will be available somewhere.
- From standpoint of CAPTCHA's, this would seem to be bad news: it gives away answers.
- Note, though, that providing a transcription is just one of many pattern recognition tasks associated with material in question.

If we didn't generate the CAPTCHA, how do we get the answer?



# Collaborative Filtering



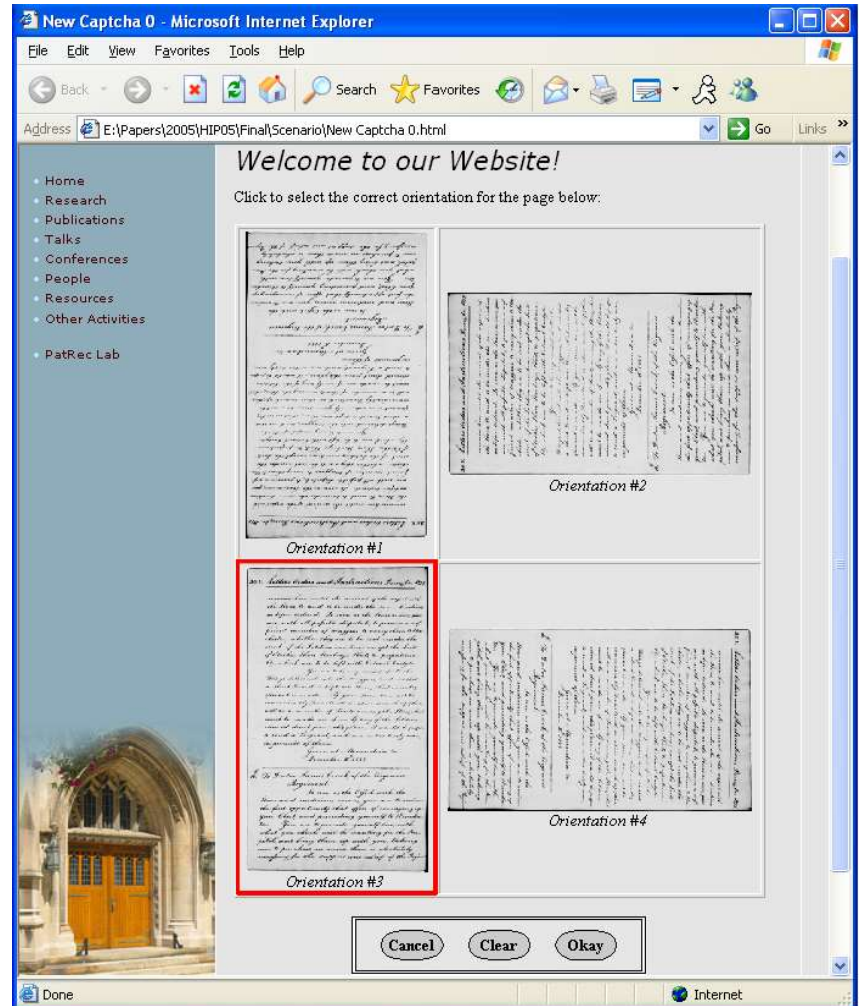
- Bootstrap from tests with known answers.
- Require users to solve more than one CAPTCHA.
- Collect responses to new candidate CAPTCHA's from proven humans to grow collection of available tests.

*“Collaborative Filtering CAPTCHA's,” Monica Chew and Doug Tygar, Human Interactive Proofs: Second International Workshop, Springer LNCS Volume 3517, May 2005, pp. 66-81.*



# Scenario 1

Click to select the correct orientation for this page.



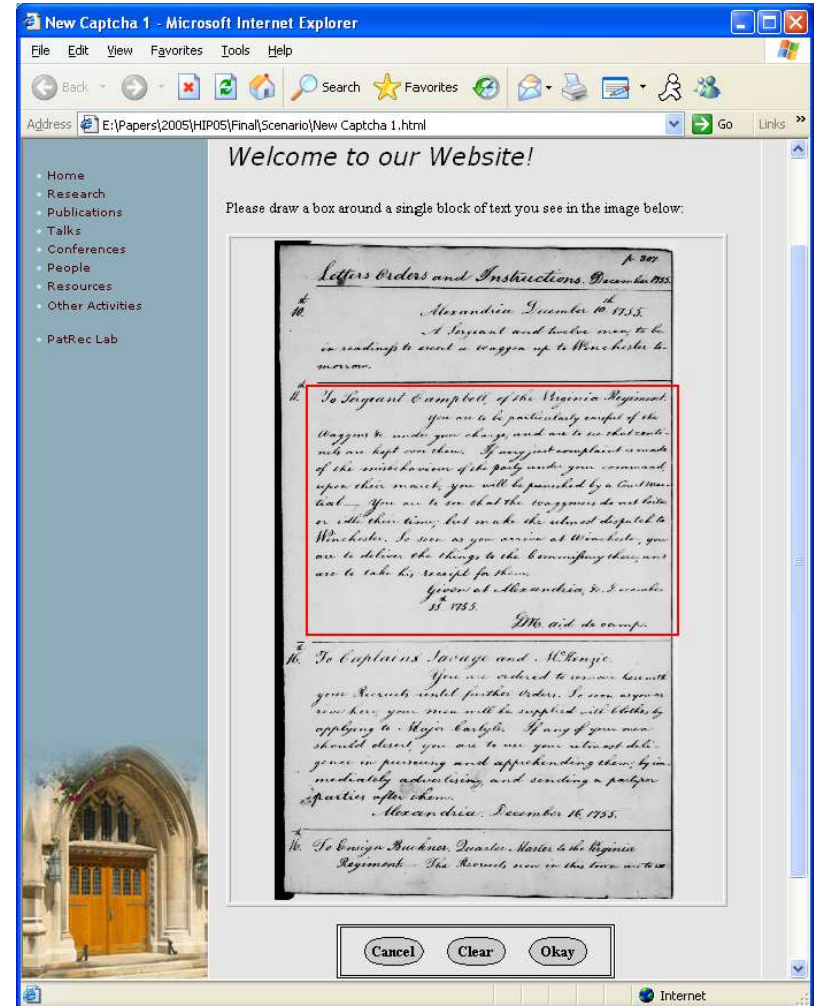
George Washington Papers at the Library of Congress  
<http://memory.loc.gov/ammem/gwhtml/gwhome.html>



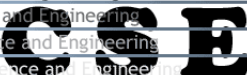


# Scenario 2

Please draw a box around a single block of text you see in the image.

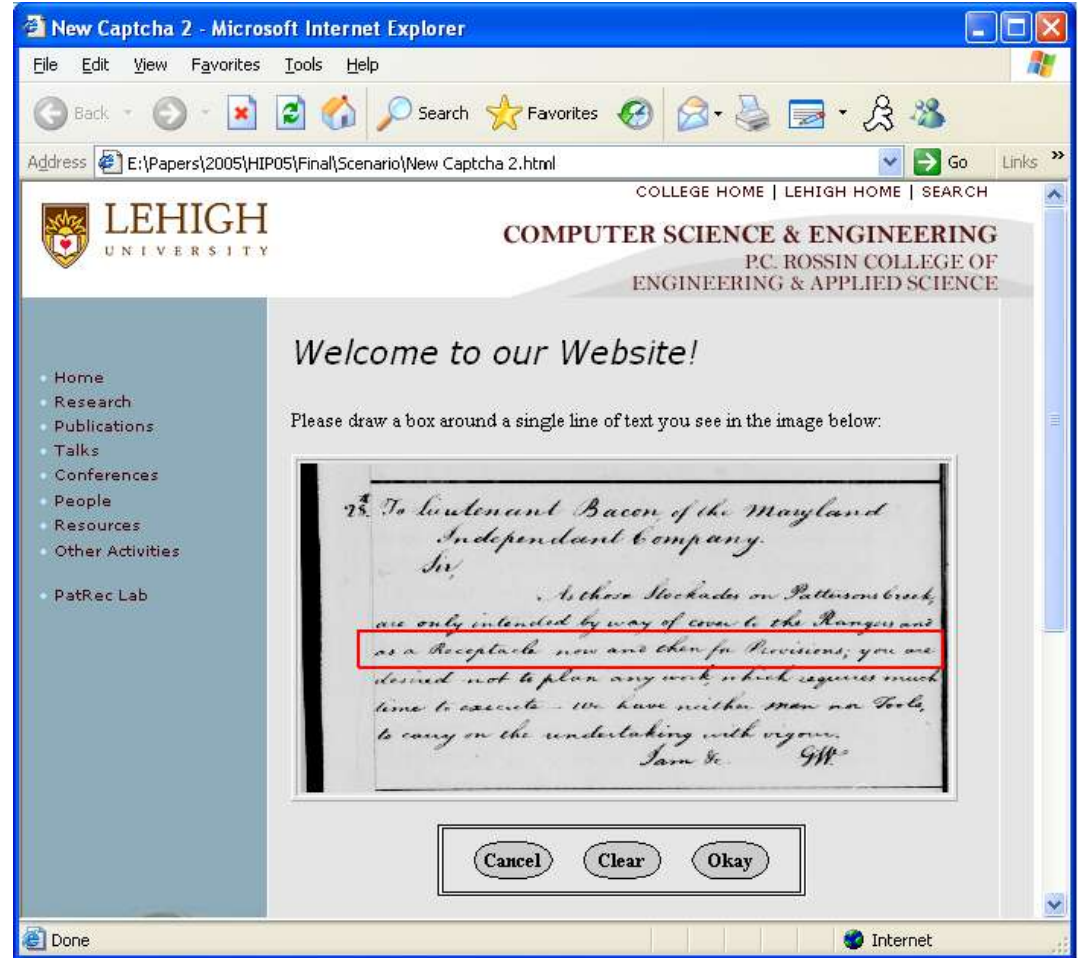


George Washington Papers at the Library of Congress  
<http://memory.loc.gov/ammem/gwhtml/gwhome.html>



# Scenario 3

Please draw a box around a single line of text you see in the image.

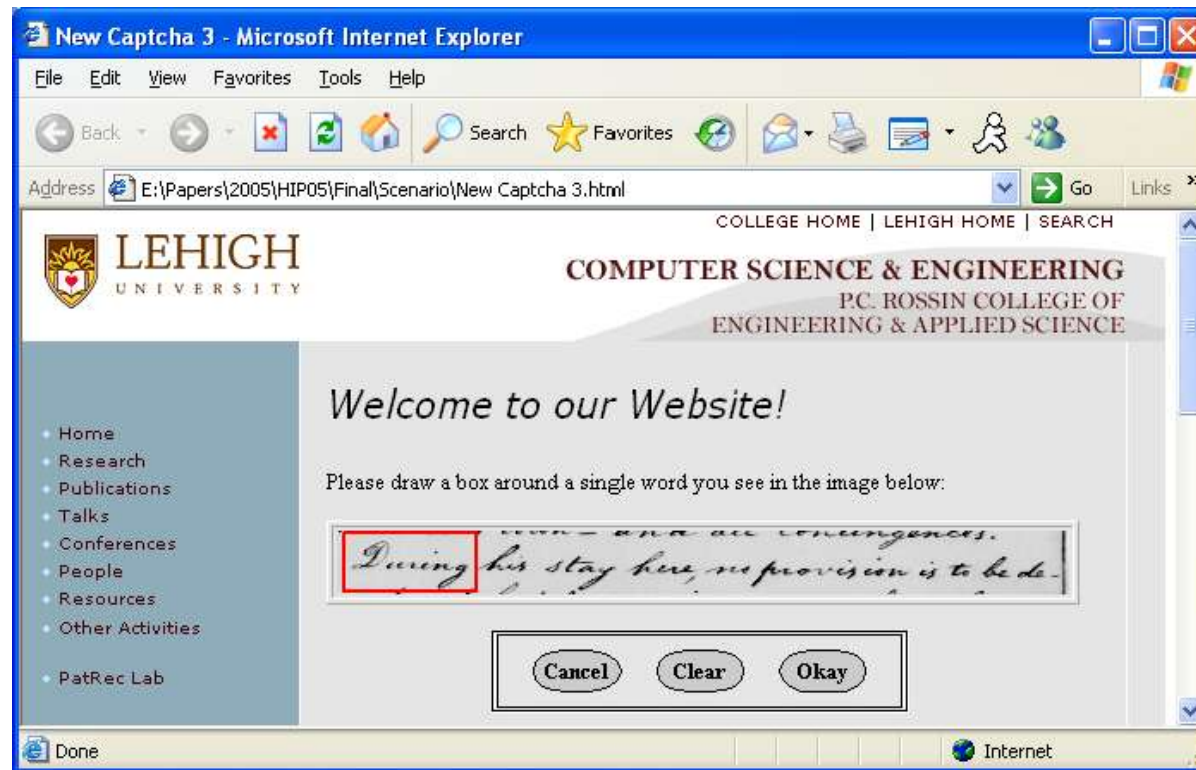


George Washington Papers at the Library of Congress  
<http://memory.loc.gov/ammem/gwhtml/gwhome.html>



# Scenario 4

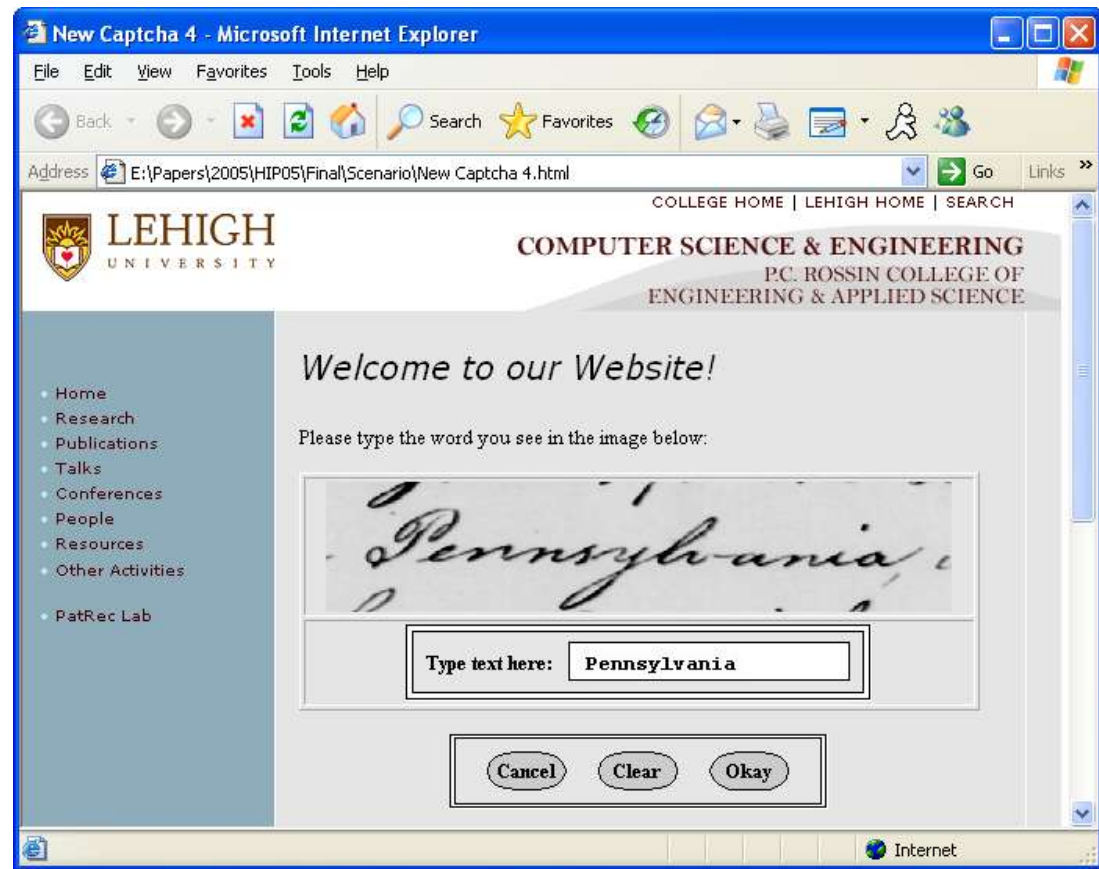
Please draw a box around a single word you see in the image.



George Washington Papers at the Library of Congress <http://memory.loc.gov/ammem/gwhtml/gwhome.html>

# Scenario 5

Please type the word you see in the image.



George Washington Papers at the Library of Congress <http://memory.loc.gov/ammem/gwhtml/gwhome.html>



# Leveraging CAPTCHA's

Allowing for differences in people's drawing skills (and a myriad of other important details), this ought to work.

Note that even if transcript for document is available online, most of the information we asked for typically isn't:

- text block segmentation,
- text line segmentation,
- word segmentation.

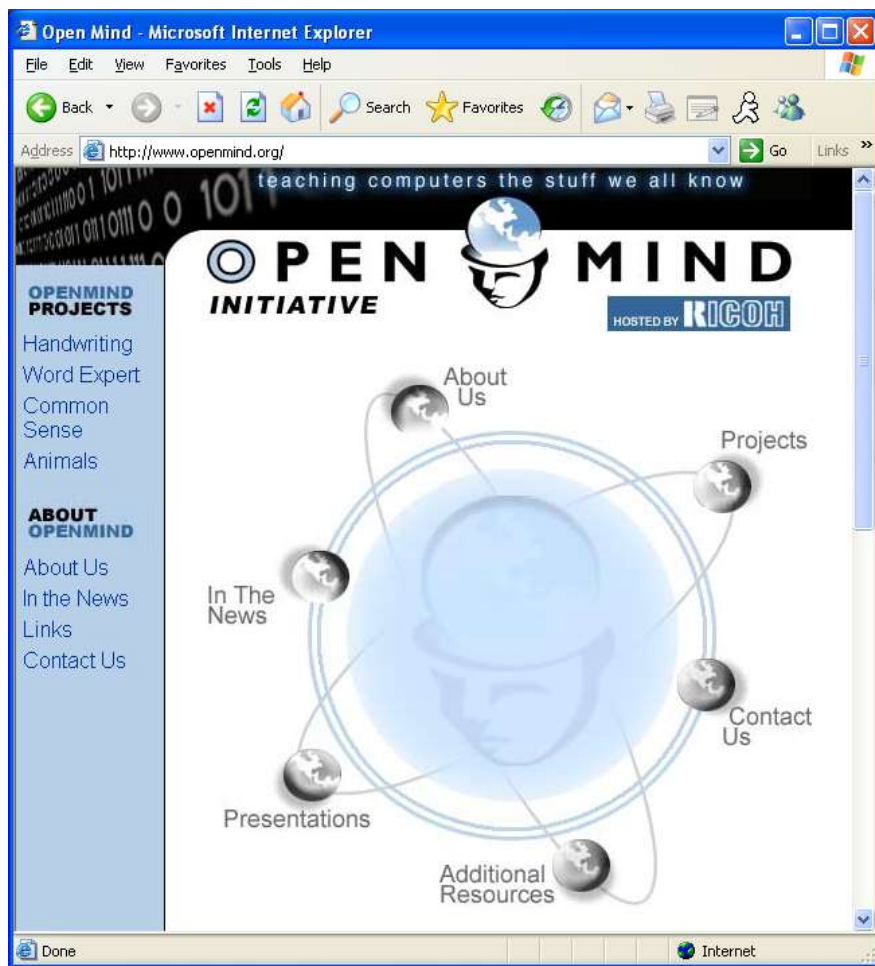
Why did we ask for it?

Because it's vitally important data (“ground-truth”) for building and evaluating document analysis systems.





# The Open Mind Initiative



“The Open Mind Initiative is a novel world-wide collaborative effort to develop 'intelligent' software. Open Mind collects information from people like you – non-expert 'netizens' – in order to teach computers the myriad things which we all know and which underlie our general intelligence but which we usually take for granted.”

<http://www.openmind.org/>

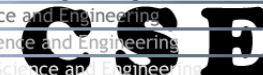
# The Open Mind Initiative

“After many decades of research, there are still very many tasks for which computers are far worse than humans: recognizing speech, reading printed or handwritten text, recognizing objects from their image, understanding scenes, making complex plans, summarizing a story, and so on ...”

“There is a growing realization that we now need information contained in very large data sets.”

“Open Mind relies on collecting, and exploiting large sets of data, such as the identities of millions of handwritten characters and spoken words, the names of objects in photographs, common sense about the world, and much, much more ...”

<http://www.openmind.org/>





# Leveraging CAPTCHA's

- Open Mind Initiative now appears to be moribund.
- Evidently appeal of labeling training and testing data does not rise to same level as participating in Open Source projects.

But economic force behind CAPTCHA's provides perfect incentive.

Major mutual benefits:

- CAPTCHA's get large source of natural pattern recognition tasks.
- Data labeled by users serves dual purpose. To pattern recognition community, could be difference in solving critical open problems.
- Attempts to break CAPTCHA's actually have a positive benefit.



# CAPTCHA's as a Web Service

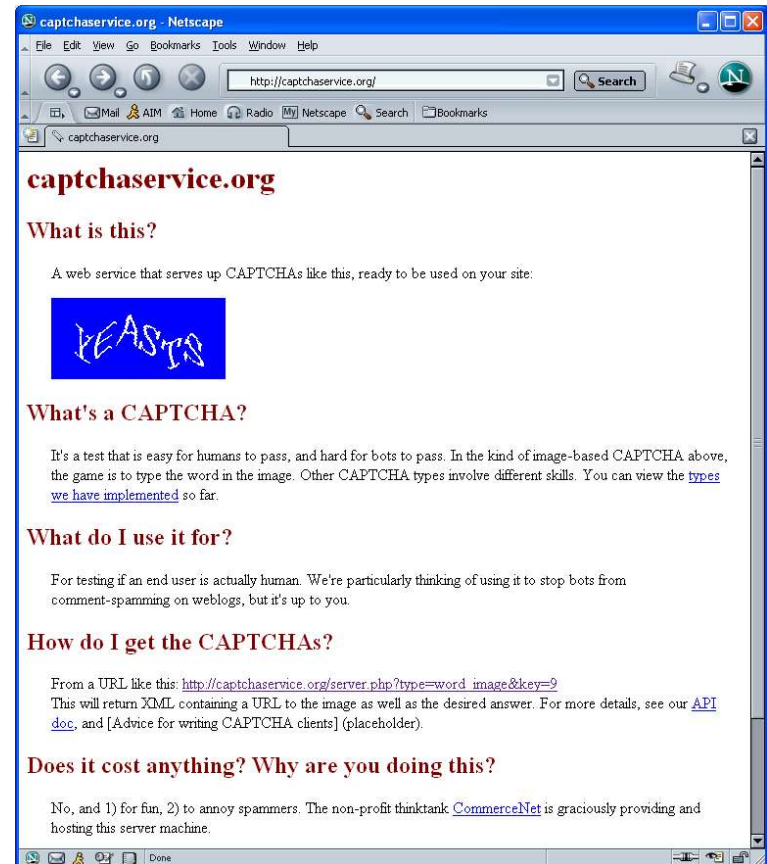
One mechanism for making this work financially and technically would be to create a Web service for CAPTCHA challenges:

- Client requests new challenge.
- Service generates challenge and delivers to client. Answer is also delivered, if known.
- Responses from users are saved for training.

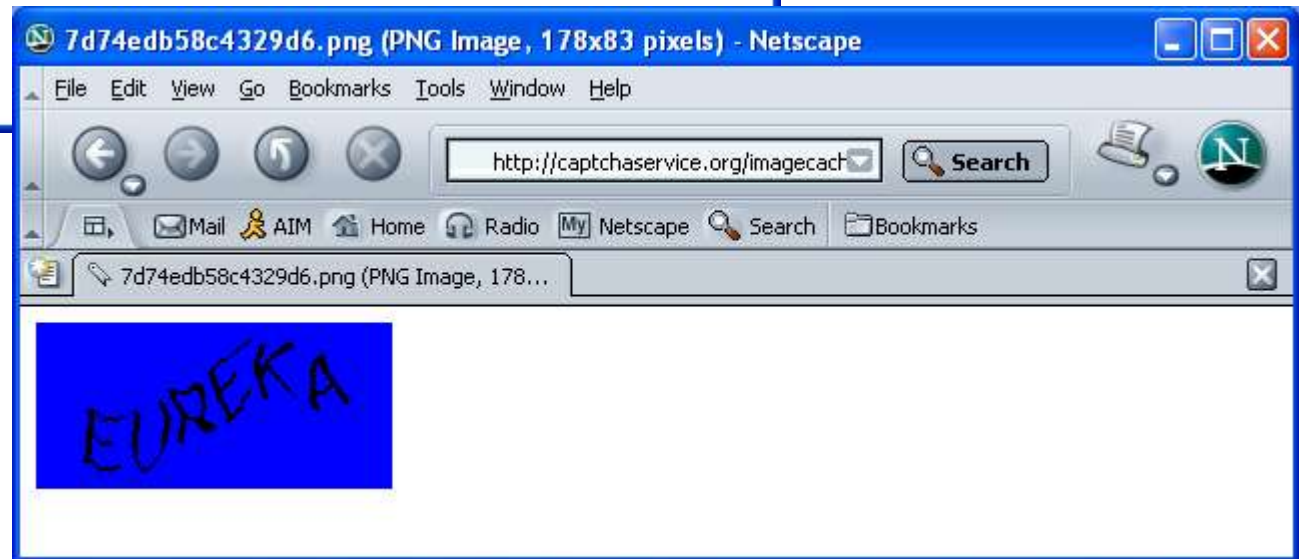
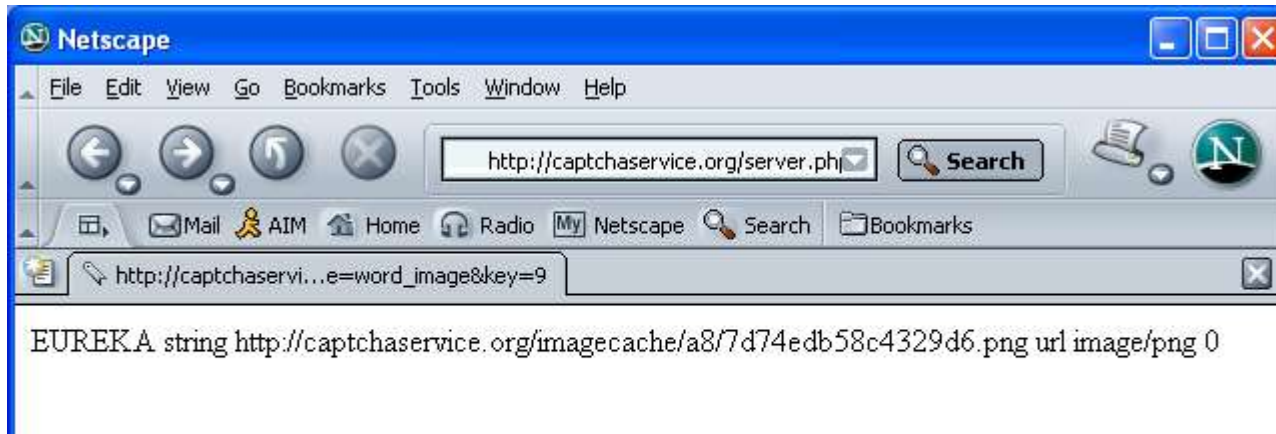
“CAPTCHA Generation as a Web Service,” Tim Converse, Proceedings of the Second International Workshop on Human Interactive Proofs, May 2005, pp. 82-96.

<http://captchaservice.org/>

Slightly different model:



# CAPTCHA's as a Web Service



<http://captchaservice.org/>

# Open Questions

- How will collaborative filtering work in such a framework?
- Are there attack modes which would allow an adversary to overwhelm system with false answers? (Bad for security and for data collection.)
- Approach requires multiple challenges: how to sequence them?
- Expertise required to field CAPTCHA's suggests server model.
- Ground-truth data only valuable once it gets released, but that makes it useless for future challenges. How to balance this?
- Idea only makes sense if it meets security needs. Does it?



# Conclusions

Web documents are inherently 2-D. While a 1-D parse is good enough for many applications, we will eventually need to bring document image analysis techniques to bear for further progress.

At the same time we're trying to make computers smart enough to understand the totality of the Web, we're depending on the fact that they don't in our design and use of CAPTCHA's.

Everywhere you look there are synergies – it's a great time to be a document analysis researcher.

THANK YOU!

