### 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



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





#### 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?









### 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



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





### 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* 

Why is there an invisible Web?

Author's biased view

- 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







### 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.

<sup>&</sup>quot;Invisible Web: Finding Hidden Content," Diane Clark, 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 prevelant 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.

🕲 New York Mets : Sortable Player Stats - Netscape	
🖌 Eile Edit <u>V</u> iew <u>G</u> o <u>B</u> ookmarks Iools <u>Wi</u> ndow <u>H</u> elp	
Search	Note header
🖌 / 🖽 🔪 🖾 Mail 🔏 Home 🞧 Radio 🕅 Netscape 🔍 Search 🗦 Bookmarks	misalignment
🛃 🔊 New York Mets : Sortable Player Stats	0
Register - Log in MLB.com N POWERED BY SUIT	
Scoreboard Stats Schedule Fan Forum MLB Sites Multimedia News Roster Shea	
Elle Edit View Insert Format Table Iools Window Help	
PLAYER LOCATOR New York Mets Hitting Stats, 2005	
Enter Last Name Loon Ink Image Table Spell	
New York Mets up to 5 players 2. C Bettran NYM Player TEAM POS G AB R H 2B 3B HR RBI TB BB SO SB CS OBP	SLG AVG
National League	11 301 382 276
Major League	2 222 425 267
Select Team         6. M Cameron         NYM         C         D <thd< th=""> <thd< th=""> <thd< th=""> <thd< th=""></thd<></thd<></thd<></thd<>	5 .352 .425.207
O Select Position	6 .389 .514.308
B. D Mentkiewicz NYM         Image: Arrow of the second secon	2 .368 .526 .286
Overview Dis. to Walksul NYM 5. M Piazza NYM C 97 351 34 91 23 0 14 55 156 34 60 0	0 .326 .444 .259
O Pitching Stats □ 11. V Diaz NYM (	
	= <b>11</b> =







#### 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

📓 Me	ts2.sxw - OpenOffice.org 1.1.0
<u>F</u> ile <u>E</u>	<u>Edit View Insert Format T</u> ools <u>W</u> indow <u>H</u> elp
C:\Doo	cuments and SettingsVdal9\My Documents\Presenta 🗾 🖹 🍅 🔛 🔛 🔛 🖨 🙏 🗈 🛍 🗾 🕂 🔹 🔝
	New Theme B iii 3D Effects - dd2066
1.	
Text b	ody _ Courier New _ 9 _ 18 i U E E E = 12 12 12 E
	L · · · · · · · · · · · · · · · · · · ·
虐	$Player \rightarrow TEAM \rightarrow POS \rightarrow G \rightarrow AB \rightarrow R \rightarrow H \rightarrow 2B \rightarrow 3B \rightarrow HR \rightarrow RBI \rightarrow TB \rightarrow BB \rightarrow SO \rightarrow SB \rightarrow CS \rightarrow OBP \rightarrow SLG \rightarrow AVG1$
<b>O</b> <sup>k</sup>	$1. \cdot J \cdot \text{Reyes} \rightarrow \text{NYM} \rightarrow \text{SS} \rightarrow 124 \rightarrow 536 \rightarrow 75 \rightarrow 148 \rightarrow 18 \rightarrow 12 \rightarrow 5 \rightarrow 50 \rightarrow 205 \rightarrow 19 \rightarrow 63 \rightarrow 45 \rightarrow 11 \rightarrow .301 \rightarrow .382 \rightarrow .276 \text{T}$
	$2.\cdot C \cdot Beltran \rightarrow NYM \rightarrow OF \rightarrow 114 \rightarrow 438 \rightarrow 63 \rightarrow 117 \rightarrow 26 \rightarrow 2 \rightarrow 13 \rightarrow 59 \rightarrow 186 \rightarrow 43 \rightarrow 73 \rightarrow 14 \rightarrow 3 \rightarrow .332 \rightarrow .425 \rightarrow .267 \text{M}$
	$3. \cdot D \cdot \text{Wright} \rightarrow \text{NYM} \rightarrow 3B \rightarrow 123 \rightarrow 438 \rightarrow 78 \rightarrow 135 \rightarrow 34 \rightarrow 1 \rightarrow 18 \rightarrow 80 \rightarrow 225 \rightarrow 56 \rightarrow 84 \rightarrow 14 \rightarrow 6 \rightarrow .389 \rightarrow .514 \rightarrow .308 \text{M}$
AÞ	$4. \cdot C \cdot Floyd \rightarrow NYM \rightarrow OF \rightarrow 116 \rightarrow 426 \rightarrow 71 \rightarrow 122 \rightarrow 19 \rightarrow 1 \rightarrow 27 \rightarrow 79 \rightarrow 224 \rightarrow 48 \rightarrow 74 \rightarrow 10 \rightarrow 2 \rightarrow .368 \rightarrow .526 \rightarrow .286 \text{T} \rightarrow .28$
I.	5. $\cdot$ M·Piazza $\rightarrow$ NYM $\rightarrow$ C $\rightarrow$ 97 $\rightarrow$ 351 $\rightarrow$ 34 $\rightarrow$ 91 $\rightarrow$ 23 $\rightarrow$ 0 $\rightarrow$ 14 $\rightarrow$ 55 $\rightarrow$ 156 $\rightarrow$ 34 $\rightarrow$ 60 $\rightarrow$ 0 $\rightarrow$ 0 $\rightarrow$ .326 $\rightarrow$ .444 $\rightarrow$ .259 $\mu$
<u> 1</u>	
Page 1	/1 Default 100% INSRT STD HYP

- 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?

<u>File E</u> dit <u>V</u> iew F <u>a</u> vorites	<u>T</u> ools <u>H</u> elp																		
🕝 Back 🔹 🕥 🕤 💌	2 🟠 🔎	Search	📌 Favo	rites 🎸	3 6	3-8		•	-	28									
Address 🔄 C:\Documents and S	ettings\dal9\My Do	ocuments\P	resentatio	ns\2005\\	VDA wo	orkshop\C	ata\bas	seball_	atistic	s.html								▼ ⇒	Go l
								2 G	<u>2</u> 1										
Player	TEAM	POS	G G	AB		R H		2B	38 F	IR R	BIT	В	BB	so	SB	CS	OBP	SLG	AVG
1. J Reyes	NYM	SS	124	1 536	5 7	5 1	48	18	12 5	) 5 	0 2	05	19	63	45	11	.301	.382	.27
2. C Beltran	NYM	OF	114	438	5 6	3 1	17	26	2 1	L3 5	9 1	.86	43	73	14	3	.332	. 425	.26
3. D Wright	NYM	3B	123	3 438	37	8 1:	35	34	1 1	L8 8	0 2	25	56	84	14	6	.389	.514	.30
4. C Floyd	NYM	OF	116	5 426	57	1 1:	22	19	1 2	27 7	9 2	24	48	74	10	2	.368	.526	.28
5. M Piazza	NYM	C	97	35:	13	4 9	1	23	0 1	L4 5	5 1	.56	34	60	0	0	.326	.444	.25
<							/////							-					
Cone	osoft Internet	Fxplorer					//**										5	My Compu	iter
Done     Done     Done     Done     Done	osoft Internet	Explorer					7,200										5	My Compu	
Done           Done           Done           Daseball statistics - Micr           Eile         Edit	osoft Internet Tools <u>H</u> elp	Explorer				4 182	7/10										5	My Compu	iter
Done baseball statistics - Micr Eile Edit View Favorites Back - O - I	osoft Internet Tools Help	Explorer Search S	Arror Favor	ites <i>@</i>	) 6	3- 8	. 🗩	3 • [	_ 1	3				ľ			3	My Compu	iter
Done Done Done Done Done Done Done Done	osoft Internet Iools <u>H</u> elp 2 () () ettings\dal9\My Do	Explorer Search	Array Favor	ites 🕑	DA wo	3 + 👌	ata\base	• [ eball_s	J 4	3.html								My Compu	iter
Done     Done	osoft Internet Tools Help Cols Models	Explorer Search	Favor resentatio	rites @	) 6 /DA wo	) + 🎝	ata\base	] ▼ [ eball_s	tatistics	3.html						0.0.0	2 I	My Compu	iter
Done     Done     Done     Done     Daseball statistics - Micr     File Edit View Favorites     Back - O - I I     Address C:\Documents and Se     Player     L Rever	osoft Internet Iools Help	Explorer Search S cuments\P POS	Favor resentatio	ites € ns\2005\₩ AB	DA wo	) + 🕹 rkshop\D: H	ata\base 2B	• eball_s	J a tatistics	3.html	TB	BB 10	SO 63	SB	CS	OBP	SL	My Compu	Go L
Done Done Done Daseball statistics - Micr File Edit View Favorites Back - O - I Address C:\Documents and Se Player 1. J Reyes C Boltrar	osoft Internet Iools Help C C C C C C C C C C C C C C C C C C C	Explorer Search S cuments\P POS SS	Favor resentation G 124 114	ites ns\2005\// AB 536 428	/DA wo R 75	+	ata\base 2B 18 26	eball_s 3B 12	Latistics HR 5	3.html 850 50	TB 205	BB 19	SO 63 72	SB 45	CS 11	OBP . 30:	SL 1 .3	My Compu	Go L VG 276
Done Done Daseball statistics - Micr File Edit View Favorites Back - O - N Address C:\Documents and Se Player 1. J Reyes 2. C Beltran 2. D Wright	osoft Internet Iools Help C C C C C C C C C C C C C C C C C C C	Explorer Search S cuments\P POS SS OF 2P	← Favor resentation G 124 114	ites € ns\2005\₩ AB 536 438 429	/DA wo R 75 63 70	+ 2 rkshop\D H 148 117 125	2B 18 26	eball_s 3B 12 2	HR 5 13	3.html 881 50 59	TB 205 186 225	BB 19 43	SO 63 73	SB 45 14	CS 11 3	OBP .30: .332	SL 1 .3 2 .4	My Compu My Compu My Compute My Compute	Go L VG 276 267
Done Done	osoft Internet Iools Help C C V V attings\dal9\My Do TEAM NYM NYM NYM NYM	Explorer Search S cuments\P POS SS OF 3B OF	← Favor resentation G 124 114 123 116	ites € AB 536 438 438	/DA wo R 75 63 78 71	+ 2 rkshop\D H 148 117 135 122	2B 18 26 34	eball_s 3B 12 2 1	HR 5 13 18 27	3.html 881 50 59 80 79	TB 205 186 225	BB 19 43 56	so 63 73 84 74	SB 45 14 14	CS 11 3 6 2	OBP .30: .332 .382	SL 1 .3 2 .4 9 .5	My Compu G AV 82 .2 25 .2 14 .3	Go L VG 276 286
Done Done Done Daseball statistics - Micr Elle Edit View Favorites Back - O - I Address C:\Documents and Se Player 1. J Reyes 2. C Beltran 3. D Wright 4. C Floyd 5. M Diagram	osoft Internet Iools Help C C P ettings\dal9\My Do TEAM NYM NYM NYM NYM NYM NYM	Explorer Search S cuments\Pi POS SS OF 3B OF C	← Favor resentatio 124 114 123 116 87	/ites € ns\2005\W AB 536 438 426 251	/DA wo R 75 63 78 71 24	H 148 117 135 122	2B 18 26 34 19 22	eball_s 3B 12 2 1 1	HR 5 13 18 27	3.html 881 50 59 80 79 55	TB 205 186 225 224	BB 19 43 56 48 24	SO 63 73 84 74	SB 45 14 14 10 0	CS 11 3 6 2	OBP .301 .332 .389	SL 1 .3 2 .4 9 .5 3 .5	My Compu G AV 82 .2 25 .2 14 .3 26 .2 44	Go L VG 276 286 286
Done Done	osoft Internet Iools Help Cols H	Explorer Search S cuments\P POS SS OF 3B OF C	G 124 114 123 116 97	rites AB 536 438 428 426 351	R 75 63 78 71 34	H 148 117 135 122 91	2B 28 28 26 34 19 23	3B 12 2 1 1 0	HR 5 13 18 27 14	3.html 83.html 50 59 80 79 55	TB 205 186 225 224 156	BB 19 43 56 48 34	SO 63 73 84 74 60	SB 45 14 14 10 0	CS 11 3 6 2 0	OBP .30: .332 .389 .368 .326	SL 1 .3 2 .4 9 .5 3 .5 5 .4	My Compu G AV 82 .2 25 .2 14 .2 26 .2 44 .2	ter Go L VG 276 267 308 286 259





Computer Scier

#### Examine HTML Tags

	Window H	<u>H</u> elp				_										
		3	Q			÷									4	
Body Text V A -a +a E	3 / U	:= ]=		image E E	- Table 5	Jen										
FONT																
FONT SPAN Player	Т	EAM	POS	G Al	3 R	Н	2B 3B	B HR	RBI	ΤВ	BB SO	SB	CS O	BP S	LG AV	G
SPAN 1. J Reyes	NYM	SS	124	536 1	75 148	18 1	L2 5	50	205	19 6	3 45	11	.301	.382	.276	=
SPAN 2. C Beltran	NYM	OF	114	438 6	53 117	26 2	2 13	59	186	43 7	3 14	3	.332	.425	.267	
SPAN 3. D Wright	NYM	ЗВ	123	438 7	78 135	34 1	L 18	80	225	56 8	4 14	6	.389	.514	.308	
SPAN 4. C Floyd	NYM	OF	116	426 7	71 122	19 1	L 27	79	224	48 7	4 10	2	.368	.526	.286	
SPAN 5. M Piazza	NYM	С	97	351 3	34 91	23 0	) 14	55	156	34 6	0 0	0	.326	.444	.259	
Normal 🖉 baseball statistics [file:/	./baseball_	_statistics	3.html] -	Composer												
Normal Baseball statistics [file:/ Elle Edit View Insert Formal New Onen	./baseball_ t Table Io	_statistics iols <u>Windor</u>	s <mark>3.html]</mark> - w <u>H</u> elp		Image	Table	Spell									
Normal	./baseball_ t Table <u>T</u> o 200 4 5344 Pub -a + <b>a</b>	statistics iols <u>Windor</u> olish Brow	s3.html] - w <u>H</u> elp }} wse Prin ⊈ ∷≣	Composer	) (J Image FI E :	Table	Spell									. (
Normal	./baseball_ t Table Io 2010 Pub -a +a	_statistics nols <u>W</u> indov Dish Brow B /	<mark>s3.html] -</mark> w <u>H</u> elp }} wse Prin U ∷≣	Composer ि स्राधन रेड रड	Image	Table	Spell									(
Normal	./baseball_ t Table Io 20 Pub -a +a	statistics ols Windo blish Brow B /	B.html] - w Help se Prin U ∷≣	Composer	Image	Table	Spell	<b>P</b> HR	RBI	ртв		so P	SB P	CS POB	P OSLG	
Normal	./baseball t Table Io Pub -a +a Pub DNYM	statistics windo wind	:3.html] - w <u>H</u> elp wse Prin U ∷≣ P 124	Composer 	Image Image FI E : R O H 75 O 144	Table = 3 1 0 2B 3 0 18	Spell		RBI	C TB	<b>0</b> BB <b>0</b> 0 19 <b>0</b>	SO <b>Q</b>	SB 0	CS <mark>P</mark> OB	P P SLG	• AVG
Normal	./baseball_ t Table To Date To -a +a Put -a +a	statistics window wi	:3.html] - w <u>H</u> elp wse Prin U :≡ P 124 J P 114 J	Composer	■ Image ■ ■ ■ = = = = = = = = = = = = = = = = =	Table = 3 0 2B 3 0 18 7 0 26	Spell Spell	<b>9</b> HR <b>9</b> 5 <b>9</b> 13	● RBI ● 50 ● 59	<ul> <li>TB</li> <li>205</li> <li>186</li> </ul>	<b>0</b> BB <b>0</b> 0 19 <b>0</b> 0 19 <b>0</b>	50 <b>0</b> 63 <b>0</b> 73 <b>0</b>	SB 0 45 0 14 0	CS <b>P</b> OB 11 <b>P</b> .31 3 <b>P</b> .31	P P SLG 01 P .38 32 P .42	<b>9 AVG</b> 2 <b>9</b> .27 5 <b>9</b> .26
Normal	./baseball_ t Table To Put -a +a NYM NYM	statistics windo wind	•3.html] - w Help vse Prin U ∷≣ 0 G 0 124 0 123	Composer	Image Image FI E : R 0 H 75 0 144 63 0 11' 78 0 13	E 28 3 0 18 7 0 26 3 0 18 7 0 26 5 0 34	Spell	♀ HR ♀ 5 ♀ 13 ♀ 18	<ul> <li>RBI</li> <li>50</li> <li>59</li> <li>80</li> </ul>	<ul> <li>■ TB</li> <li>■ 205</li> <li>■ 186</li> <li>■ 225</li> </ul>	<b>0</b> BB <b>0</b> 0 19 <b>0</b> 0 43 <b>0</b> 0 56 <b>0</b>	50 63 73 84	38 <mark>8</mark> 45 8 14 8 14 8	CS <b>Q</b> OB 11 <b>Q</b> .3 3 <b>Q</b> .3 5 <b>Q</b> .3	P <mark>9</mark> SLG 01 0 .38 32 9 .42	<b>9 AVG</b> 2 <b>9</b> .27 5 <b>9</b> .26 4 <b>9</b> .30
Normal	./baseball_ t Table Io	statistics windo slish Brow B / POS SS OF SS OF OF OF OF	3. html] - w Help se Prin U ∷ 0 G 124 0 124 0 124	Composer	R P H 75 P 144 63 P 137 78 P 137 71 P 127	Table Table 2 28 3 9 18 7 9 26 5 9 34 2 9 19	Spell Sp	<b>○</b> HR <b>○</b> 5 <b>○</b> 13 <b>○</b> 18 <b>○</b> 27	<ul> <li>RBI</li> <li>50</li> <li>59</li> <li>80</li> <li>79</li> </ul>	9 TB 9 205 9 186 9 225 9 224	<b>2</b> BB <b>2</b> <b>2</b> 19 <b>2</b> <b>2</b> 43 <b>2</b> <b>2</b> 56 <b>2</b> <b>3</b> 56 <b>2</b> <b>4</b> 8 <b>3</b>	SO 0 63 73 84 74	38 <b>9</b> 45 <b>9</b> 14 <b>9</b> 14 <b>9</b> 14 <b>9</b>	C5 90B 11 9.33 3 9.33 5 9.33 5 9.33	P <mark>9</mark> SLG D1 9.438 32 9.42 59 9.51 68 9.52	<b>A</b> VG 2 <b>9</b> .27 5 <b>9</b> .26 4 <b>9</b> .30 6 <b>9</b> .28
Normal	./baseball_ t Table Io C TEAM Pub -a +a Pub -a +a Pub NYM PNYM PNYM PNYM	statistics swindov sish Brow POS SS OF OF OF OF OF	<b>3. html]</b> - w <u>H</u> elp w <u>H</u> elp wse Prin U ∷ 0 G 0 124 0 114 0 116 0 116 0 977	Composer ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓	R P H 75 P 14 63 P 11 78 P 12 71 P 12 34 P 91	Table Table 2 2B 3 0 18 7 0 26 5 0 34 2 0 19 9 23	Spell Spell 0 12 0 2 0 1 0 1 0 1 0 1 0 1 0 1	PHR 5 13 13 218 27 914	<ul> <li>RBI</li> <li>50</li> <li>59</li> <li>80</li> <li>79</li> <li>55</li> </ul>	<b>9</b> TB <b>9</b> 205 <b>9</b> 186 <b>9</b> 225 <b>9</b> 224 <b>9</b> 156	2     BB     2       2     19     2       2     43     2       2     56     2       2     43     2       3     4     2	50 63 73 84 74 60	SB (1) 45 (2) 14 (2) 14 (2) 14 (2) 10 (2) 0 (2)	C5 00B 11 0.3 3 0.3	P O SLG 01 O .38 32 O .42 89 O .51 68 O .52 26 O .44	<b>9</b> AVG 2 <b>9</b> .27 5 <b>9</b> .26 4 .30 6 <b>9</b> .28 4 <b>9</b> .25
Normal Body Text Player Player C Beltran C C Floyd ES. M Piazza	./baseball t Table Io C Pub -a +a P TEAM P NYM NYM NYM NYM NYM NYM NYM NYM	statistics swindov sish Brow B / P POS SS P OF P 3B P 0F P C	3. html] - w Help see Prin U ∷ 0 G 0 124 0 114 0 116 0 116 0 977	Composer ↓ unk ↓ = ₹ ↓ = ↓	R P H 75 0 144 63 0 11 78 0 12 71 0 12 34 0 91	Image: Constraint of the second sec	Spell Spell 2 2 2 1 2 1 2 1 2 1 2 0 1 2 0 1	<ul> <li>● HR</li> <li>● 5</li> <li>● 13</li> <li>● 18</li> <li>● 27</li> <li>● 14</li> </ul>	<b>2</b> RBI <b>3</b> 50 <b>9</b> 59 <b>9</b> 79 <b>9</b> 55	<ul> <li>□ TB</li> <li>□ 205</li> <li>□ 225</li> <li>□ 224</li> <li>□ 224</li> <li>□ 156</li> </ul>	0 BB 0 0 19 0 0 43 0 0 43 0 0 48 0 0 34 0	50 9 63 9 84 9 74 9 60 9	SB         Q           45         Q           14         Q           14         Q           14         Q           10         Q           0         Q	CS <b>P</b> OB 11 <b>P</b> .31 3 <b>P</b> .31 3 <b>P</b> .32 3 <b>P</b> .32 2 <b>P</b> .33 2 <b>P</b> .33 3 <b>P</b> .33	P P SLG 01 P .38 32 P .51 89 P .51 26 P .52 26 P .44	<b>O</b> AVG 2 <b>O</b> .27 5 <b>O</b> .26 4 <b>O</b> .28 4 <b>O</b> .25

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



Not a Table

Table





#### More Table Problems

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



# Work on HTML table detection by Wang and Hu:

	L	Т	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.

Genuine Table

"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 <u>is</u> a table if and only if it <u>looks like</u> a table.



Counter-arguments:

"Parsing is easier and usually works." "What about the Semantic Web?"









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/Ov erview.html







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

Ele Edit View Go Bookmarks Tools Window Help         Image: Search Structure       Image: Search Structure         Image: Search Structure       Search Structure         <	🕲 All Things Web: Understanding HTML Validation - Netscape
<ul> <li>Image: Search in the search ine</li></ul>	🖌 Eile Edit View <u>G</u> o Bookmarks Iools <u>Wi</u> ndow <u>H</u> elp
<ul> <li>Mail &amp; AIM A Hone Radio M Netscape Search Bookmarks Lehigh Univers</li> <li>All Things Web: Understanding HTML Valida</li> <li>Compatibility &amp; Accessibility</li> <li>Design Fundamentals</li> <li>Technical Tips</li> <li>HTML Watch +</li> <li>Webmaster's Comer</li> <li>About ATW</li> <li>About ATW</li> </ul>	Search Search
	🔺 / 🖽, 🛛 🖼 Mail 🙏 AIM 🐔 Home 🎧 Radio 🔤 Netscape 🔍 Search 🛛 Bookmarks 🛇 Lehigh Univers
Compatibility & Accessibility       Rules, Rules, Rules         Design Fundamentals       Most everyone knows that the "L" in HTML stands for Language. In one sense, computer languages are no different from spoken languages. Every language (including HTML) has certain known "rules" that determine what passes for sensible communication. One commonly understood HTML rule, for instance, is that every <h2> tag must be "closed" by a corresponding </h2> tag.         HTML Watch ←       All too many Web authors, however, use "This page looks just fine on my browser" as a substitute for "This page follows the basic rules of HTML." And thus is born the bratty Web page. Joe's browser is kind enough to overlook Joe's flagram violations of the rules undertying HTML, so he figures the page is fine, and he posts it on the Website. But when Jane's less tolerant browser hits the page, it chokes and dies. Many contemporary Web authors propose a simple solution: "No problem! Just switch browsers!!" (Does the phrase "Borrow a clue" mean anything to you?)	🔄 🛷 All Things Web: Understanding HTML Valida
Design Fundamentals       Most everyone knows that the "L" in HTML stands for Language. In one sense, computer languages are no different from spoken languages. Every language (including HTML) has certain known "rules" that determine what passes for sensible communication. One commonly understood HTML rule, for instance, is that every <h2> tag must be "closed" by a corresponding </h2> tag.         HTML Watch ←       All too many Web authors, however, use "This page looks just fine on my browser" as a substitute for "This page follows the basic rules of HTML." And thus is born the bratty Web page. Joe's browser is kind enough to overlook Joe's flagrant violations of the rules underlying HTML, so he figures the page is fine, and he posts it on the Website. But when Jane's less tolerant browser hits the page, it chokes and dies.         Many contemporary Web authors propose a simple solution: "No problem! Just switch browsers!!" (Does the phrase "Borrow a clue" mean anything to you?)	Compatibility & Rules, Rules, Rules
Webmaster's Corner       All too many Web authors, however, use "This page looks just fine on my browser" as a substitute for "This page follows the basic rules of HTML. " And thus is born the bratty Web page. Joe's browser is kind enough to overlook Joe's flagrant violations of the rules underlying HTML, so he figures the page is fine, and he posts it on the Website. But when Jane's less tolerant browser hits the page, it chokes and dies.         Many contemporary Web authors propose a simple solution: "No problem! Just switch browsers!!" (Does the phrase "Borrow a clue" mean anything to you?)	Design Fundamentals       Most everyone knows that the "L" in HTML stands for Language. In one sense, computer languages are no different from spoken languages. Every language (including HTML) has certain known "rules" that determine what passes for sensible communication. One commonly understood HTML rule, for instance, is that every <h2> tag must be "closed" by a corresponding </h2> tag.
Many contemporary Web authors propose a simple solution: "No problem! Just switch browsers!!" (Does the phrase "Borrow a clue" mean anything to you?)	Webmaster's Corner         About ATW         About ATW    All too many Web authors, however, use "This page looks just fine on my browser" as a substitute for "This page follows the basic rules of HTML. " And thus is born the bratty Web page. Joe's browser is kind enough to overlook Joe's flagrant violations of the rules underlying HTML, so he figures the page is fine, and he posts it on the Website. But when Jane's less tolerant browser hits the page, it chokes and dies.
	Many contemporary Web authors propose a simple solution: "No problem! Just switch browsers!!" (Does the phrase "Borrow a clue" mean anything to you?)

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







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.

![](_page_20_Picture_4.jpeg)

![](_page_20_Picture_6.jpeg)

![](_page_20_Picture_7.jpeg)

### 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!

![](_page_21_Picture_4.jpeg)

![](_page_21_Picture_6.jpeg)

![](_page_21_Picture_7.jpeg)

#### 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.

![](_page_22_Picture_3.jpeg)

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.

![](_page_22_Picture_9.jpeg)

![](_page_22_Picture_11.jpeg)

![](_page_22_Picture_12.jpeg)

#### Protecting Web Services

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

Netscape:Welcom	e to Yahoo!
Sack Forward Reload Home Search Netscape Images	s Print Security Shop Stop
.ocation : 🥠 [http://edit.yahoo.com/config/eval_register?.int]=us≠	ew=1&.done=&.src=ym&.partner=&.p=&) 🍘 What's Rela
	0
rch Listing: 🔲 List my new Yahoo! Mail address for free	
Listing includes real name, city, state, and country.	
cial offers, promotions, and research surveys from selected Yahoo! pa	artners through Yahoo! Delivers.
e info on the Harris Poll Online! Take surveys to earn points redeema s	able for free gifts even win \$10,000 in the Harris Poll
Entertainment 🔲 Business 🔲 Shopping	
Home & Family Computers & Technology Sports & O Health Personal Finance Travel	Jutdoors
Music 🔲 Small Business 🔲 Sweepstak	kes & Free Stuff
Enter the word as it is shown in the box below.	Word Verification
STREETS STREETS STREETS STREETS STREETS STREETS STREETS	This step helps Yahoo! prevent automated registrations
	registrations.
	If you can not see this image <u>click here</u> .
and the state of the	
110 WO	
110 WS	
uwwa	

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

![](_page_23_Picture_4.jpeg)

![](_page_23_Picture_6.jpeg)

![](_page_23_Picture_7.jpeg)

### 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.

![](_page_24_Figure_2.jpeg)

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

![](_page_24_Picture_4.jpeg)

![](_page_24_Picture_6.jpeg)

![](_page_24_Picture_7.jpeg)

### Lehigh HIP Workshop, May 2005

#### Web page

![](_page_25_Picture_2.jpeg)

![](_page_25_Picture_3.jpeg)

![](_page_25_Picture_4.jpeg)

#### http://www.cse.lehigh.edu/prr/HIP05/index.html

![](_page_25_Picture_6.jpeg)

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

![](_page_25_Picture_8.jpeg)

![](_page_25_Picture_9.jpeg)

#### 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.

![](_page_26_Picture_10.jpeg)

![](_page_26_Picture_12.jpeg)

![](_page_26_Picture_13.jpeg)

#### 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.

![](_page_27_Picture_5.jpeg)

![](_page_27_Picture_7.jpeg)

![](_page_27_Picture_8.jpeg)

#### Natural vs. Synthetic #1

of virtuous and enlightened men-to closs thes inhabitants of Harrisburgh among this number is only to bear testimony to the zealows and efficient exertions which they have made towards the defence of the laws. Mastington
Golo. 4. 1794.
lestimony .
What word do you see in the box?
George Washington Papers at the Library of Congress

What "word" do you see in the box?

Please review the following terms and indicate your agreement below. Printable Version 📇

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

Yahoo! Registration - Microsoft Internet Explorer

Title:

Industry: [Select Industry] [Select a Title] 💙

Specialization: [Select a Specialization]

🔇 Back 🔹 🕥 - 💽 😰 🏠 🔎 Search 👷 Favorites 🤣 🔗 🌺 🤜 🎘 🦓

More info This helps Yahool prevent automated registrations.

1. ACCEPTANCE OF TERMS

sJAKKC

Address 🧃 http://edit.yahoo.com/config/eval\_register?.v=&.intl=&new=1&.done=http%3a//mail.yahoo.com 🗸 🋐 Go 🛛 Links 🍟

File Edit View Favorites Tools Help

**Customizing Yahoo!** 

Verify Your Registration \* Enter the code shown:

**Terms of Service** 

![](_page_28_Picture_4.jpeg)

http://memory.loc.gov/ammem/gwhtml/gwhome.html

![](_page_28_Picture_6.jpeg)

![](_page_28_Picture_7.jpeg)

Internet

3

#### Natural vs. Synthetic #2

![](_page_29_Picture_1.jpeg)

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

![](_page_29_Figure_3.jpeg)

#### 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

![](_page_29_Picture_6.jpeg)

![](_page_29_Picture_8.jpeg)

Computer Science and Engineering Computer Science and Engineering Computer Science and Engineering Computer Science and Engineering

#### 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.

![](_page_30_Picture_7.jpeg)

![](_page_30_Picture_9.jpeg)

![](_page_30_Picture_10.jpeg)

### 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.

![](_page_31_Picture_5.jpeg)

![](_page_31_Picture_7.jpeg)

![](_page_31_Picture_8.jpeg)

### Describe the weather in this scene.

![](_page_32_Picture_2.jpeg)

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

![](_page_32_Picture_4.jpeg)

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

![](_page_32_Picture_6.jpeg)

![](_page_32_Picture_7.jpeg)

## Which photos show the same person?

![](_page_33_Picture_2.jpeg)

![](_page_33_Picture_3.jpeg)

![](_page_33_Picture_4.jpeg)

![](_page_33_Picture_5.jpeg)

![](_page_33_Picture_6.jpeg)

![](_page_33_Picture_7.jpeg)

![](_page_33_Picture_8.jpeg)

![](_page_33_Picture_10.jpeg)

![](_page_33_Picture_11.jpeg)

### How many cars do you see in this image?

![](_page_34_Picture_2.jpeg)

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

![](_page_34_Picture_4.jpeg)

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

![](_page_34_Picture_6.jpeg)

![](_page_34_Picture_7.jpeg)

Draw a box around a text string in this image.

![](_page_35_Figure_2.jpeg)

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

![](_page_35_Picture_4.jpeg)

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

![](_page_35_Picture_6.jpeg)

![](_page_35_Picture_7.jpeg)

#### One obvious answer: in digital libraries.

![](_page_36_Picture_2.jpeg)

http://www.dw-world.de/dw/article/0,1564,1566717,00.html

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

![](_page_36_Picture_5.jpeg)

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

![](_page_36_Picture_7.jpeg)

Computer Science and Engineering Computer Science and Engineering Computer Science and Engineering Computer Science and Engineering

### 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?

![](_page_37_Picture_6.jpeg)

![](_page_37_Picture_8.jpeg)

![](_page_37_Picture_9.jpeg)

#### **Collaborative Filtering**

![](_page_38_Picture_1.jpeg)

- 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.

![](_page_38_Picture_6.jpeg)

![](_page_38_Picture_8.jpeg)

![](_page_38_Picture_9.jpeg)

#### Click to select the correct orientation for this page.

![](_page_39_Picture_2.jpeg)

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

![](_page_39_Picture_4.jpeg)

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

![](_page_39_Picture_6.jpeg)

#### Computer Science and Engineering Computer Science and E

![](_page_39_Picture_8.jpeg)

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

![](_page_40_Picture_2.jpeg)

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

![](_page_40_Picture_4.jpeg)

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

![](_page_40_Picture_6.jpeg)

#### Computer Science and Engineering

![](_page_40_Picture_8.jpeg)

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

![](_page_41_Picture_2.jpeg)

![](_page_41_Picture_3.jpeg)

![](_page_41_Picture_5.jpeg)

Computer Science and Engineering Computer Science and Engineering Computer Science and Engineering

http://memory.loc.gov/ammem/gwhtml/gwhome.html

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

![](_page_42_Picture_2.jpeg)

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

![](_page_42_Picture_4.jpeg)

![](_page_42_Picture_6.jpeg)

![](_page_42_Picture_7.jpeg)

# Please type the word you see in the image.

![](_page_43_Picture_2.jpeg)

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

![](_page_43_Picture_4.jpeg)

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

![](_page_43_Picture_6.jpeg)

![](_page_43_Picture_7.jpeg)

### 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.

![](_page_44_Picture_8.jpeg)

![](_page_44_Picture_10.jpeg)

![](_page_44_Picture_11.jpeg)

### The Open Mind Initiative

![](_page_45_Picture_1.jpeg)

"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/

![](_page_45_Picture_4.jpeg)

![](_page_45_Picture_6.jpeg)

![](_page_45_Picture_7.jpeg)

### 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/

![](_page_46_Picture_5.jpeg)

![](_page_46_Picture_7.jpeg)

![](_page_46_Picture_8.jpeg)

### 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.

![](_page_47_Picture_8.jpeg)

![](_page_47_Picture_10.jpeg)

![](_page_47_Picture_11.jpeg)

### 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/

#### R

![](_page_48_Picture_8.jpeg)

#### Slightly different model:

![](_page_48_Picture_10.jpeg)

![](_page_48_Picture_11.jpeg)

#### CAPTCHA's as a Web Service

🕲 Netscape	
🔺 Eile Edit View <u>G</u> o Bookmarks <u>T</u> ools <u>W</u> indo	w <u>H</u> elp
	/captchaservice.org/server.php
🚽 🖽 🖂 Mail 🚴 AIM 🐔 Home 🎧 Radio	My Netscape 🔍 Search 🖾 Bookmarks
🕙 🛇 http://captchaservie=word_image&key=9	
EUREKA string http://captchaservice.org/u	magecache/a8/7d74edb58c4329d6.png url image/png 0
	🔊 7d74edb58c4329d6.png (PNG Image, 178x83 pixels) - Netscape 📃 🗖 📘
	Eile Edit View Go Bookmarks Iools Window Help
	Content of the service.org/imagecack Search
	🖌 🖂 🖂 Mail 🚴 AIM 🐔 Home 🎧 Radio 🔤 Netscape 🔍 Search 🛛 Bookmarks
	🕘 💊 7d74edb58c4329d6.png (PNG Image, 178
	EUREKA
http://captchaservice.org/	
	Computer Science and Engineering

![](_page_49_Picture_2.jpeg)

![](_page_49_Picture_4.jpeg)

Computer Science an Computer Science

### 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?

![](_page_50_Picture_7.jpeg)

![](_page_50_Picture_9.jpeg)

![](_page_50_Picture_10.jpeg)

#### 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!

![](_page_51_Picture_5.jpeg)

![](_page_51_Picture_7.jpeg)

![](_page_51_Picture_8.jpeg)