Balancing Security and Usability in a Video CAPTCHA

Richard Zanibbi[1] and Kurt Kluever[2]

November 19, 2008

[1] Assistant Professor, Department of Computer Science
Rochester Institute of Technology, USA

[2] Google, New York

First Things First: Some Definitions

C ompletely

A utomated

P ublic *(data, alg's)*

T uring Test, to tell

C omputers and

H umans

A part

Secure Test

Machines fail frequently (few false positives)

Usable Test

People pass frequently (many true positives), comfortable task

CAPTCHA Tasks: AI and Pattern Recognition Problems

Natural Language Understanding

Filling in missing words in sentences, pronoun disambiguation

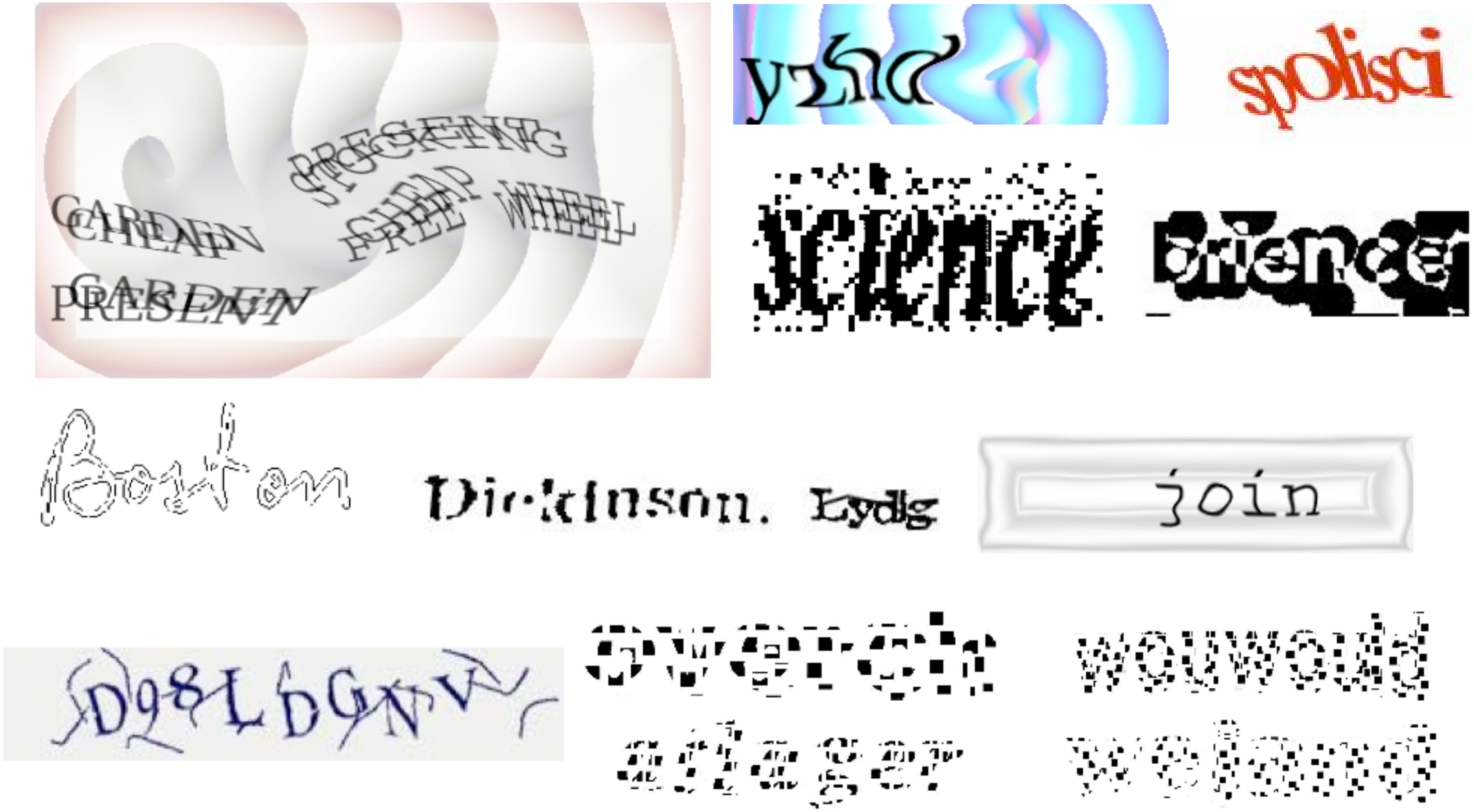
Audio-Based

Transcribe text in a (noisy) audio file

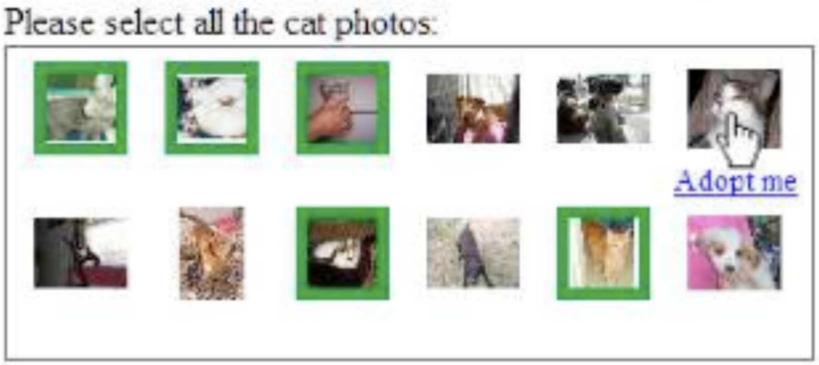
Image-Based

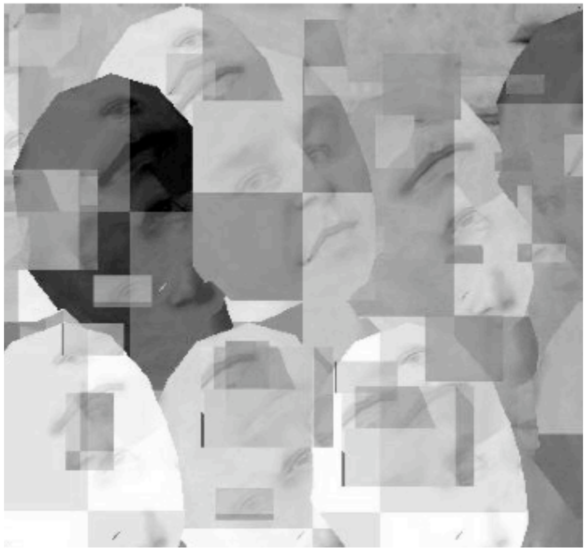
Distorted characters, image region/content labeling, etc.

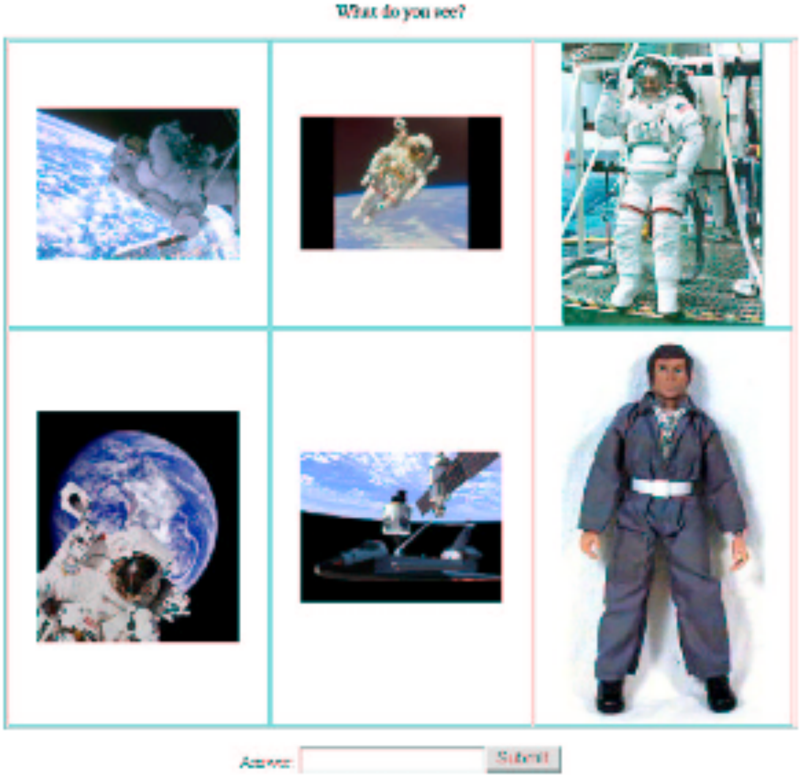
Distorted Text Tests

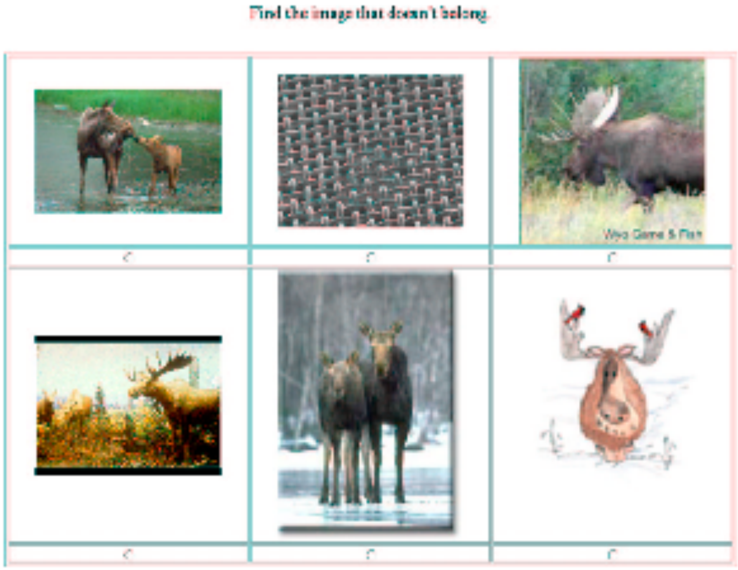


Other Image-Based Tests


Please select all the cat photos:




What do you see?


Find the image that doesn't belong.

Motivation for New Tests

Distorted Text CAPTCHAs most prevalent

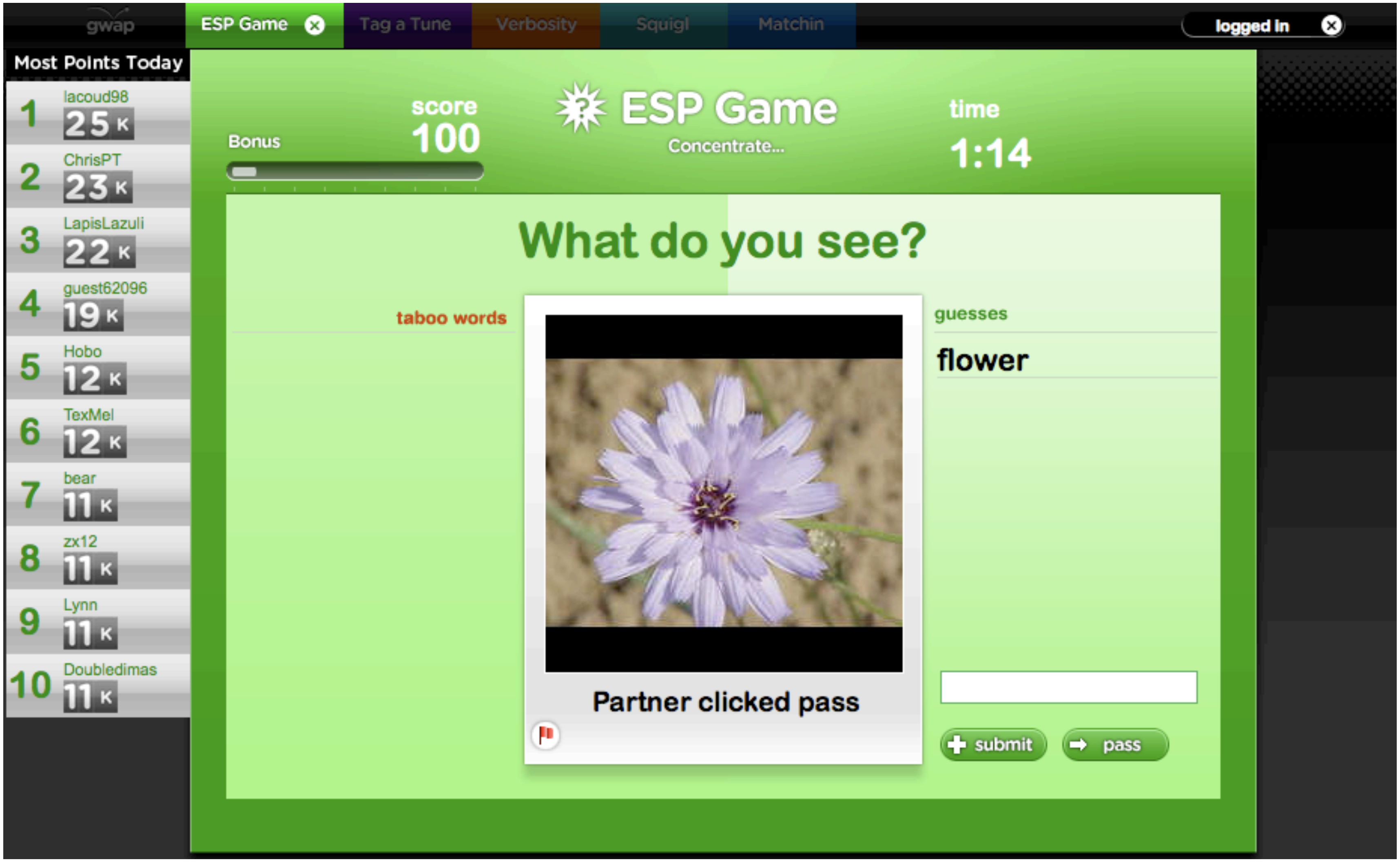
- Many people report finding these frustrating (significant distortion needed for security)
- Becoming vulnerable, e.g. Microsoft text CAPTCHA recently broken with a 60% pass rate (Yan & Ahmad, CCS 2008)

...a more secure but user-friendly task is needed

R·I·T

The ESP Game

(Von Ahn et al., CHI 2004) http://gwap.com



Most Points Today

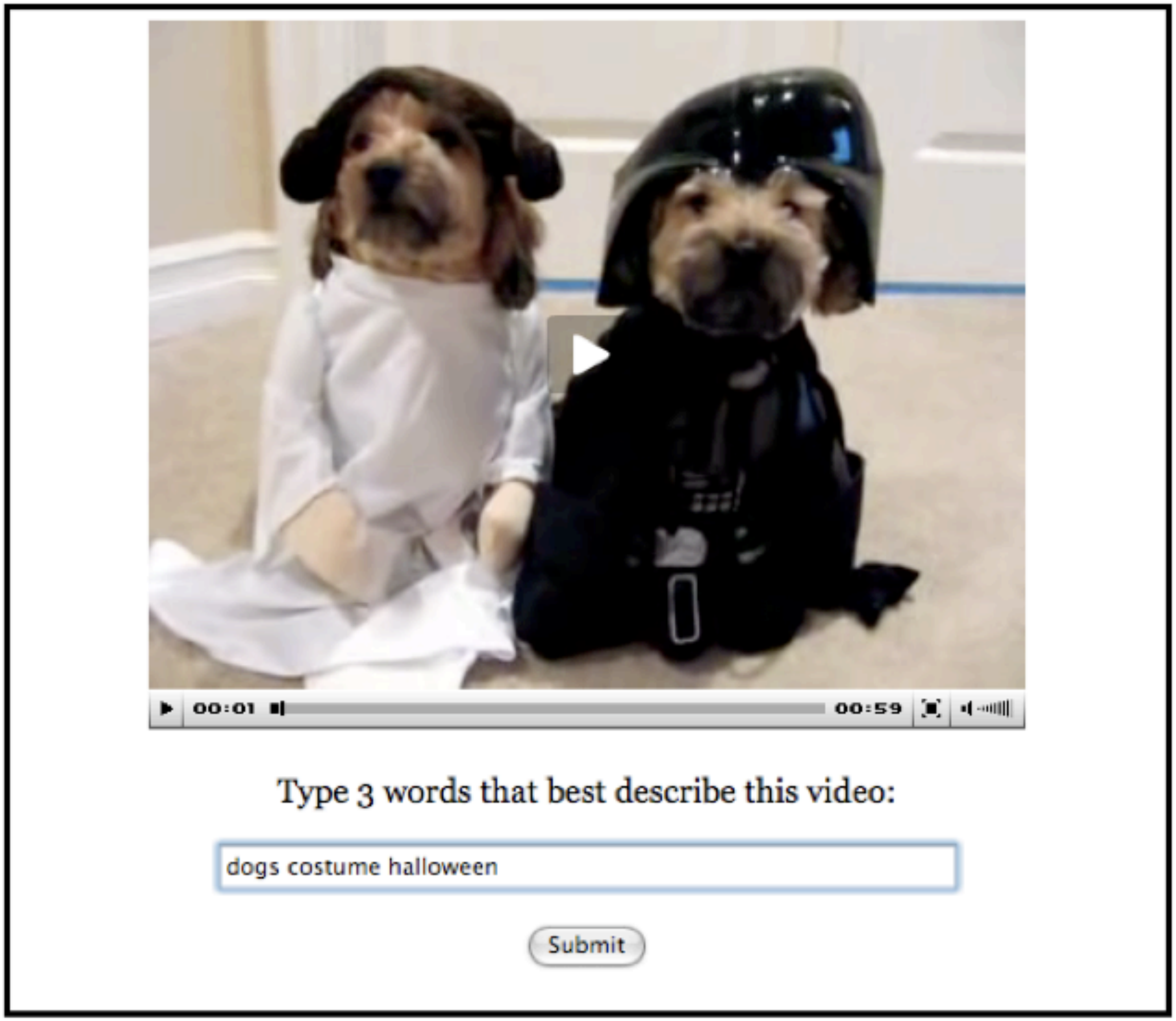
1	lacoud98	25 K
2	ChrisPT	23 K
3	LapisLazuli	22 K
4	guest62096	19 K
5	Hobo	12 K
6	TexMel	12 K
7	bear	11 K
8	zx12	11 K
9	Lynn	11 K
10	Doubledimas	11 K

What do you see?

flower

Partner clicked pass

A Video CAPTCHA



Type 3 words that best describe this video:

dogs costume halloween

Properties of our Video CAPTCHA

Almost Completely Automatic

May need to check appropriateness of video content

Public

Algorithms, data (e.g. YouTube) open

Security

Comparable to existing methods against submission of three most frequent tags. Additional attacks (e.g. CBIR) need study

Usability

Equal/better pass rates than for existing methods, small majority of users in study preferred task to "distorted text" tasks

Test Generation and Grading

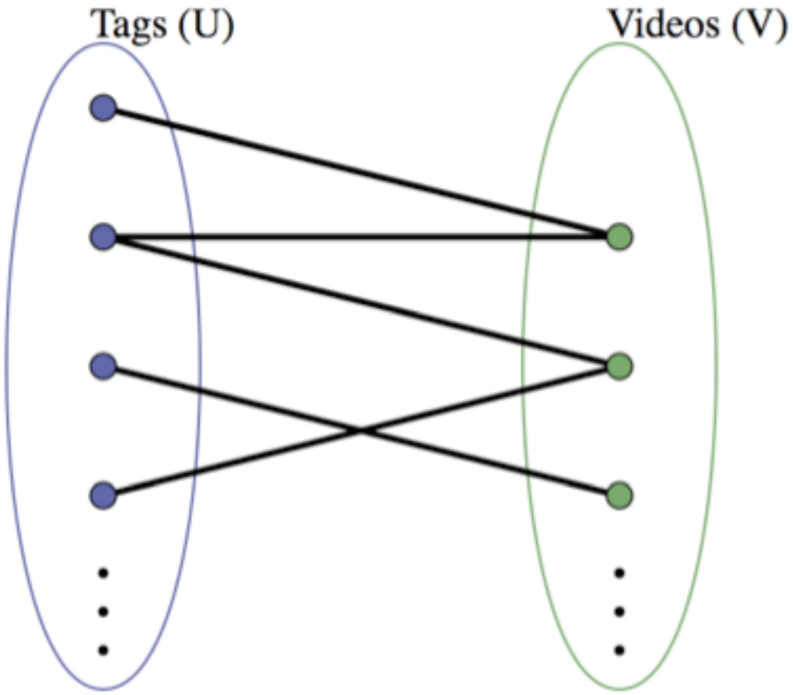
Public Video Data Set: YouTube.com

Data Set

- ~150 Million Videos (August 2008)
- Individuals upload videos with ‘tags’ in a120 character field

Sampling YouTube

- Random generation of video id’s impractical
- Limits on number of accesses per day



Solution: Use dictionary word to ‘seed’ a random walk

Generating Tests

1. Select random dictionary word, query database
2. Random walk of [1,100] steps, return video reached
3. From ‘related videos’ add n additional tags (list sorted by cosine similarity of tags to test video)
4. Remove tags estimated to be more frequent than a threshold t
5. Normalize tags: Remove stop words (‘the,’ ‘a’ etc.), convert to lower case, remove punctuation

Comparing Tag Sets: Cosine Similarity Metric

Let A and B be binary vectors of the same length (represent all tags in A&B)

$$\textsc{Sim}(A, B) = \cos\theta = \frac{A \cdot B}{\|A\|\|B\|}$$

$$\cos\theta = \frac{|A_t \cap R_t|}{\sqrt{|A_t|}\sqrt{|R_t|}}$$

Tag Set	Occ. Vector	dog	puppy	funny	cat
A_t	A	1	1	1	0
R_t	B	1	1	0	1

R·I·T

Grading Tests

User Provides Three Non-Stop Words

Normalization: set tags to lower case, punctuation stripped

Pass if a 'valid' test tag is submitted

'Usability' Parameters

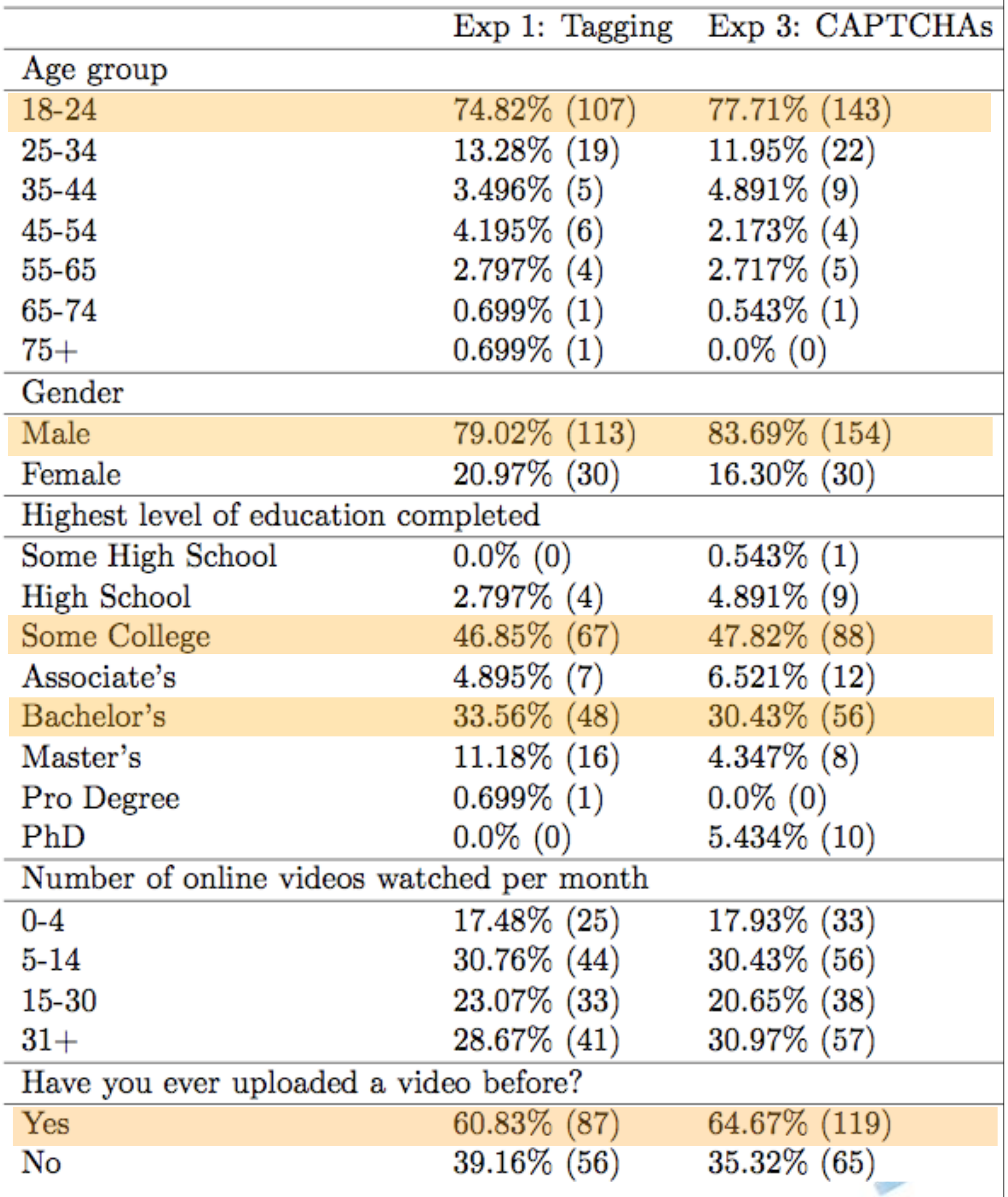
- Stemming: add word stems (Porter alg.; max +3 tags) e.g. running ⇒ run
- Edit distance: accept submitted tags within normalized similarity of 'valid' test tags (≥0.8; 1 edit for strings length 5-9)

$$\textsc{NormLevenshtein}(s_1, s_2) = 1 - \frac{\textsc{Levenshtein}(s_1, s_2)}{\textsc{Max}(|s_1|, |s_2|)}$$

Experiments

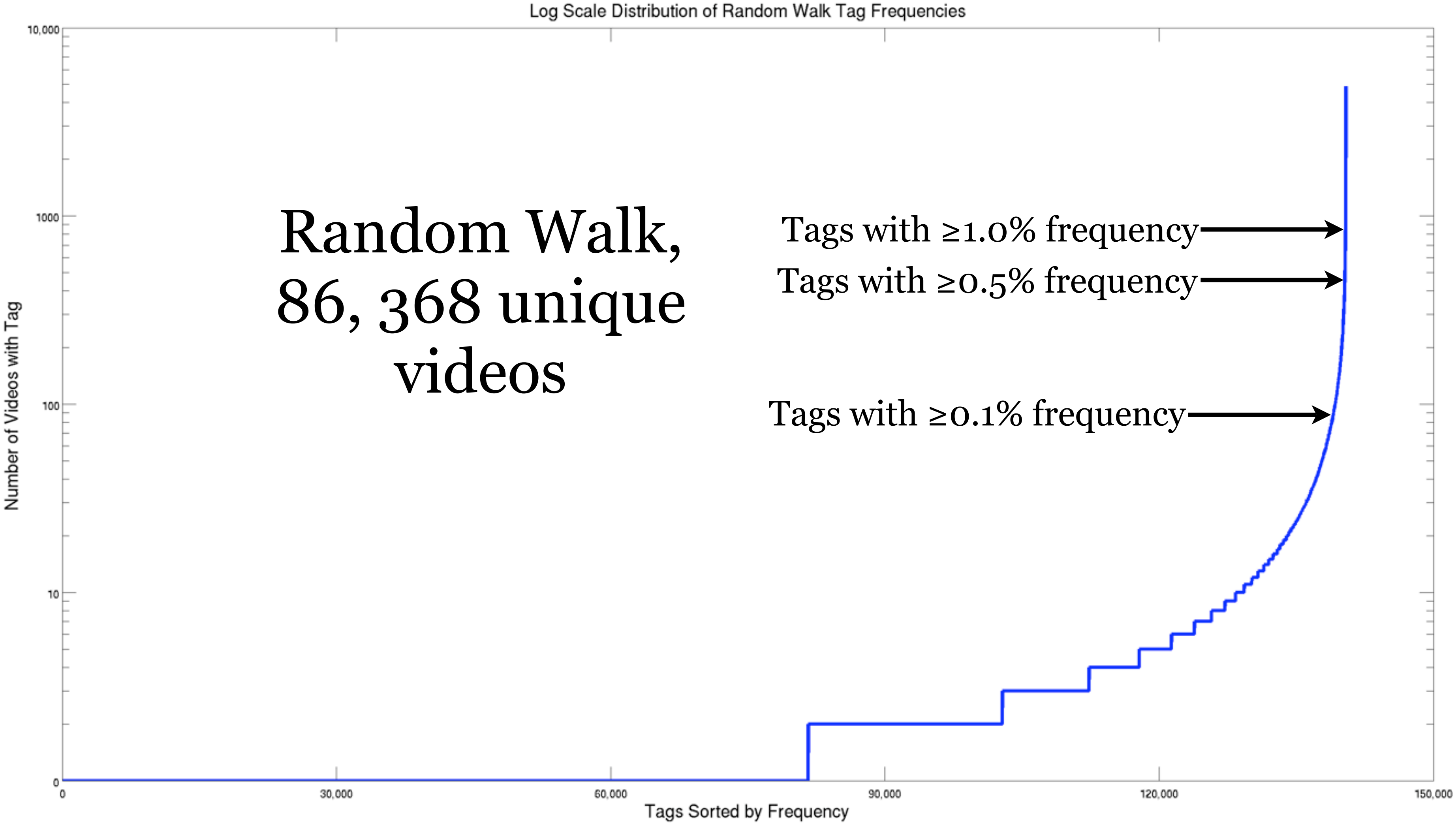
Three Experiments

1. Tagging (Design/Training)
    - 143 participants (online)
    - 20 videos, selected manually
2. Video CAPTCHA
    - 184 participants (online)
    - 20 videos, selected via random walk
3. Attack Simulation
    - 5146 videos, selected via random walk



	Exp 1: Tagging	Exp 3: CAPTCHAs
Age group		
18-24	74.82% (107)	77.71% (143)
25-34	13.28% (19)	11.95% (22)
35-44	3.496% (5)	4.891% (9)
45-54	4.195% (6)	2.173% (4)
55-65	2.797% (4)	2.717% (5)
65-74	0.699% (1)	0.543% (1)
75+	0.699% (1)	0.0% (0)
Gender		
Male	79.02% (113)	83.69% (154)
Female	20.97% (30)	16.30% (30)
Highest level of education completed		
Some High School	0.0% (0)	0.543% (1)
High School	2.797% (4)	4.891% (9)
Some College	46.85% (67)	47.82% (88)
Associate's	4.895% (7)	6.521% (12)
Bachelor's	33.56% (48)	30.43% (56)
Master's	11.18% (16)	4.347% (8)
Pro Degree	0.699% (1)	0.0% (0)
PhD	0.0% (0)	5.434% (10)
Number of online videos watched per month		
0-4	17.48% (25)	17.93% (33)
5-14	30.76% (44)	30.43% (56)
15-30	23.07% (33)	20.65% (38)
31+	28.67% (41)	30.97% (57)
Have you ever uploaded a video before?		
Yes	60.83% (87)	64.67% (119)
No	39.16% (56)	35.32% (65)

Tag Frequency Distribution

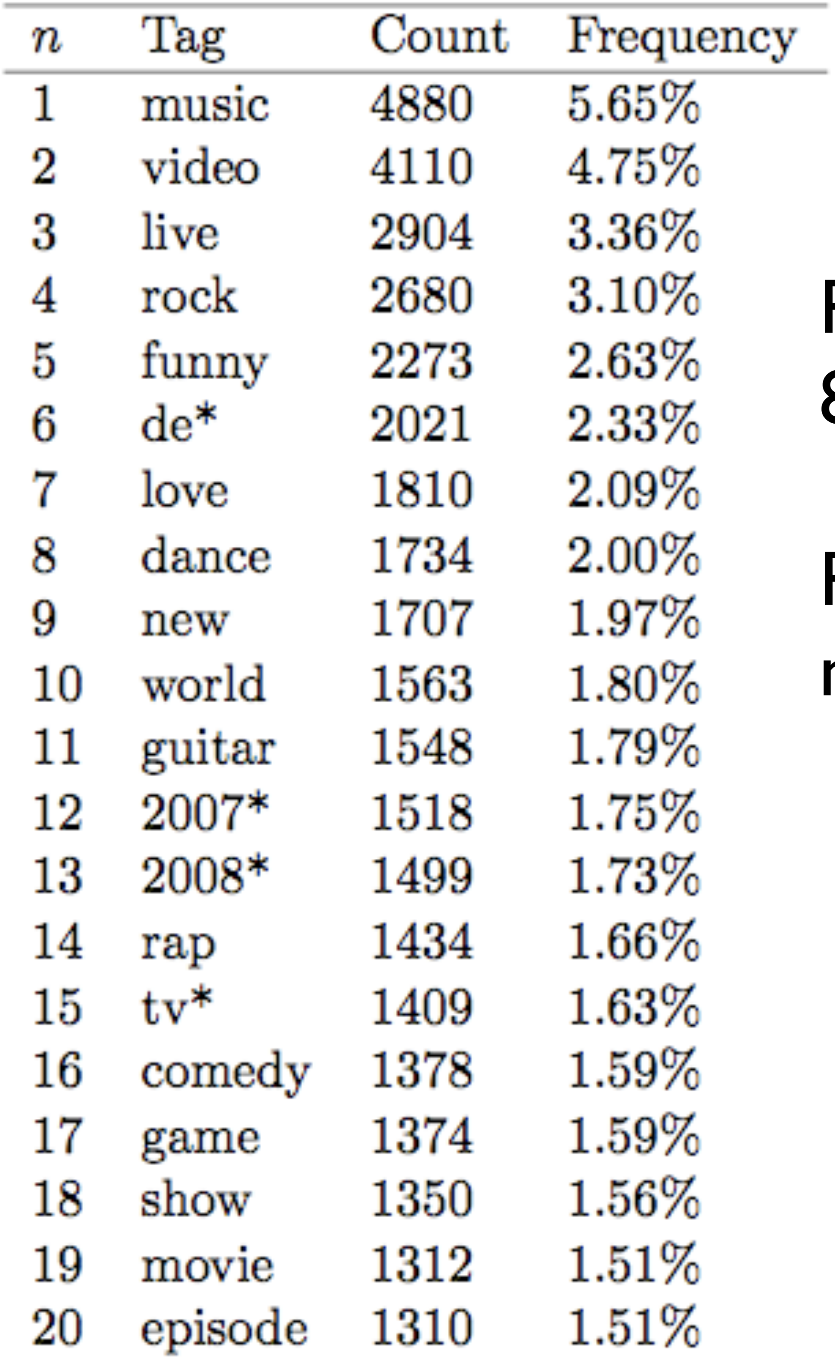


Random Walk, 86, 368 unique videos

Tags with ≥1.0% frequency

Tags with ≥0.5% frequency

Tags with ≥0.1% frequency



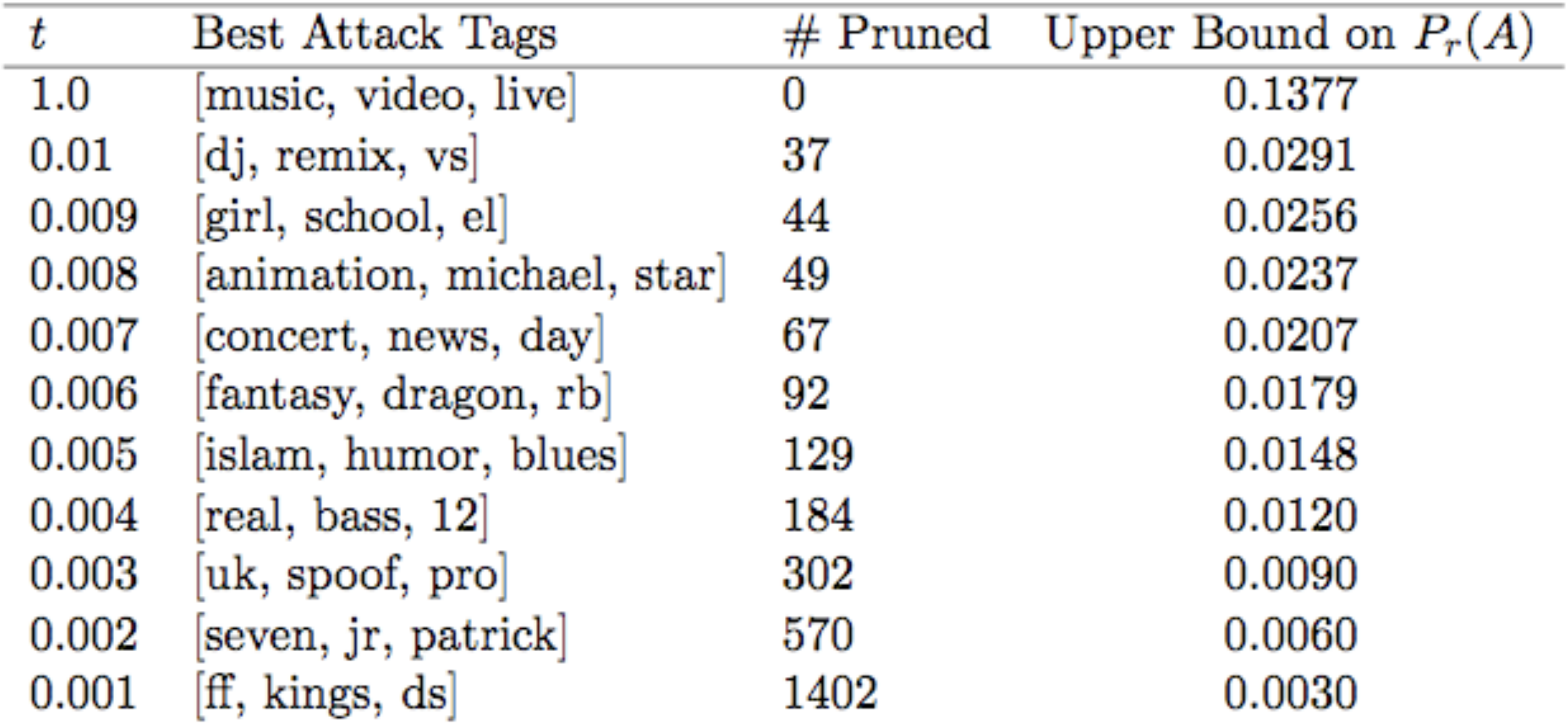
n	Tag	Count	Frequency
1	music	4880	5.65%
2	video	4110	4.75%
3	live	2904	3.36%
4	rock	2680	3.10%
5	funny	2273	2.63%
6	de*	2021	2.33%
7	love	1810	2.09%
8	dance	1734	2.00%
9	new	1707	1.97%
10	world	1563	1.80%
11	guitar	1548	1.79%
12	2007*	1518	1.75%
13	2008*	1499	1.73%
14	rap	1434	1.66%
15	tv*	1409	1.63%
16	comedy	1378	1.59%
17	game	1374	1.59%
18	show	1350	1.56%
19	movie	1312	1.51%
20	episode	1310	1.51%

Random Walk reaching 86,368 Unique Videos

Random walk revealed tags not in our dictionary (*)

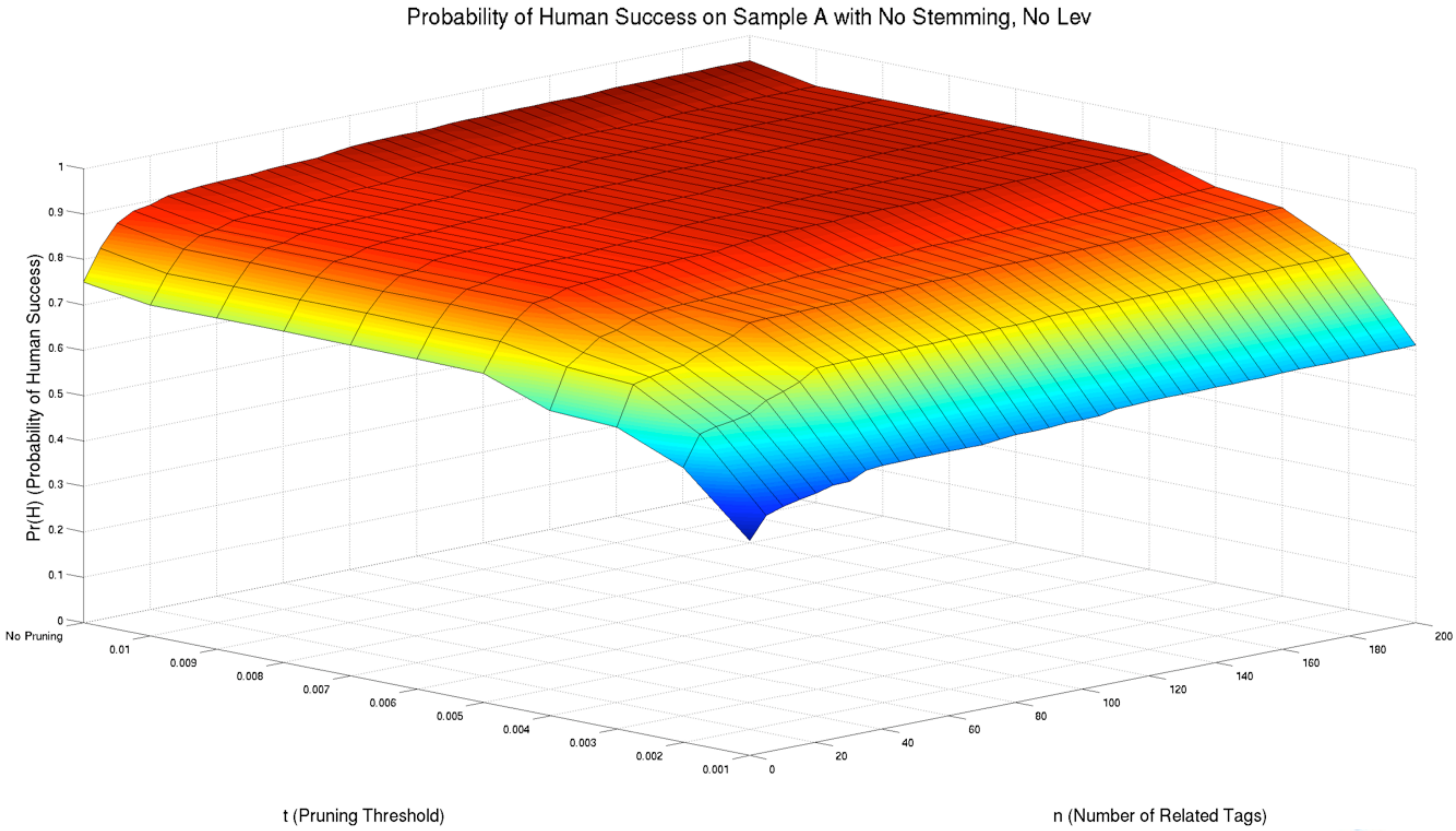
Frequency-Based Attacks

Most Frequent Tags Below Threshold t:

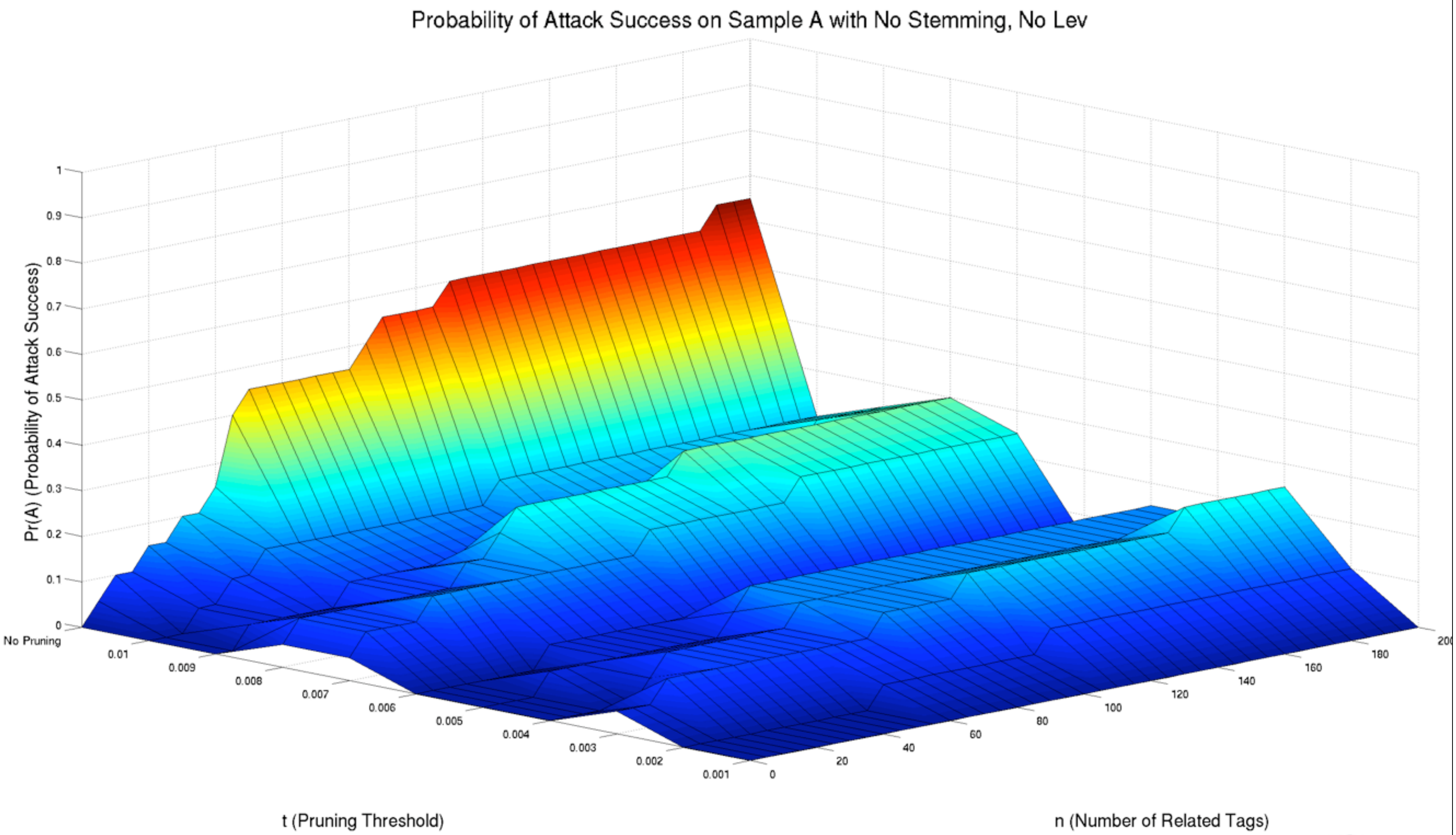


t	Best Attack Tags	# Pruned	Upper Bound on $P_r(A)$
1.0	[music, video, live]	0	0.1377
0.01	[dj, remix, vs]	37	0.0291
0.009	[girl, school, el]	44	0.0256
0.008	[animation, michael, star]	49	0.0237
0.007	[concert, news, day]	67	0.0207
0.006	[fantasy, dragon, rb]	92	0.0179
0.005	[islam, humor, blues]	129	0.0148
0.004	[real, bass, 12]	184	0.0120
0.003	[uk, spoof, pro]	302	0.0090
0.002	[seven, jr, patrick]	570	0.0060
0.001	[ff, kings, ds]	1402	0.0030

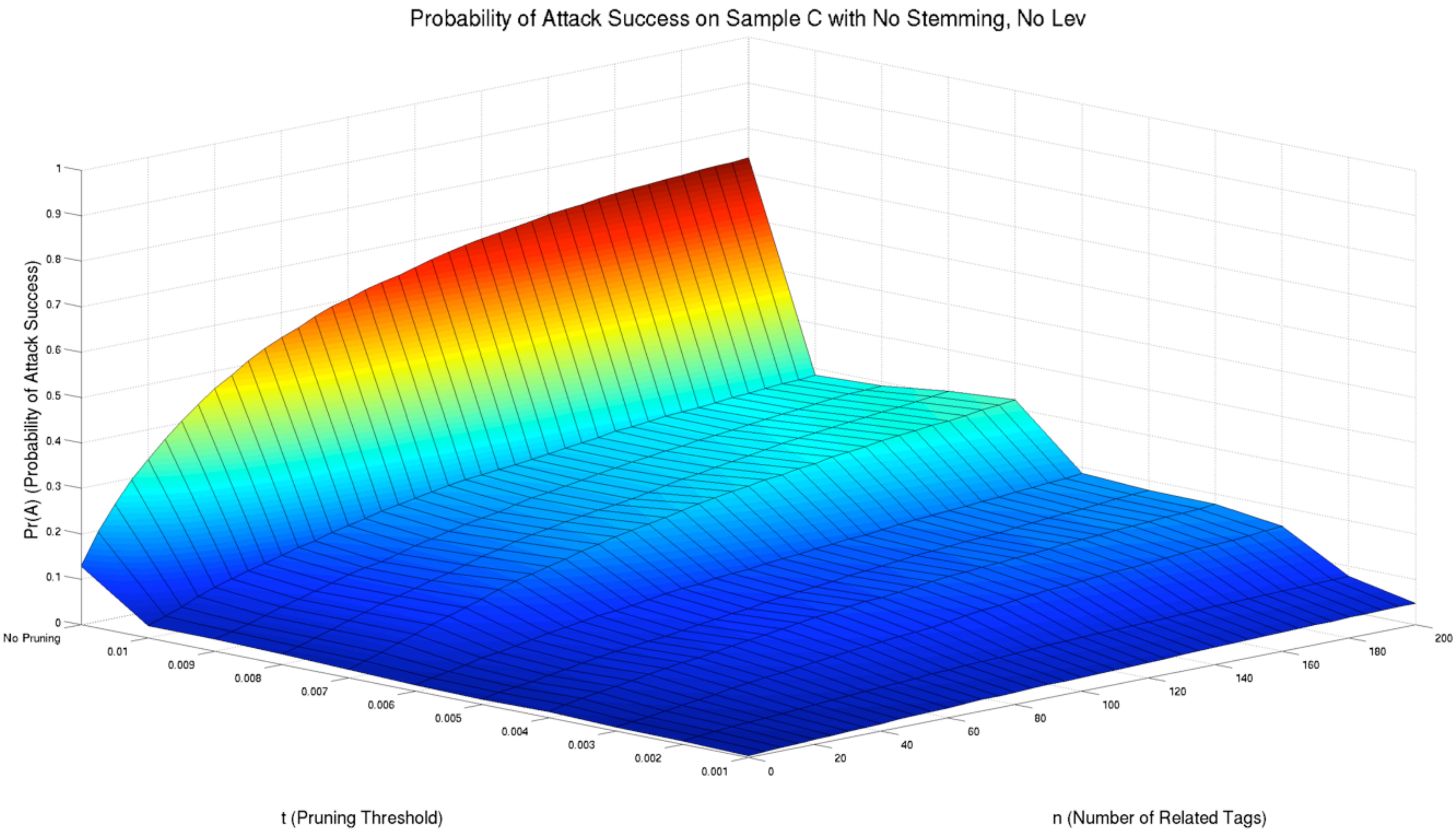
Human Rates Exp 1: 20 Videos, Manual Selection



Attack Rates Exp 1: 20 Videos, Manual Selection

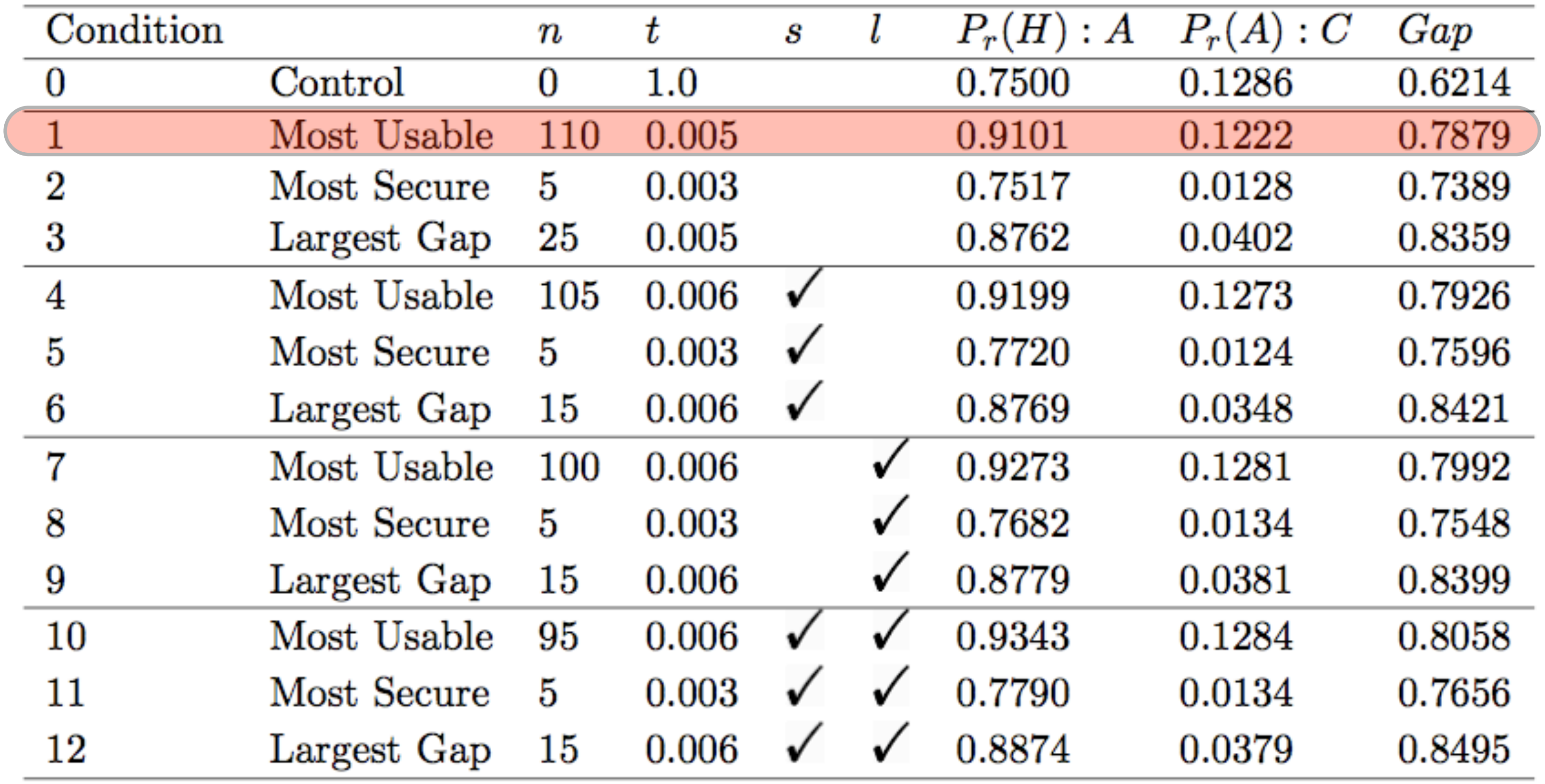


Attack Simulation: 5146 Videos, Random Walk



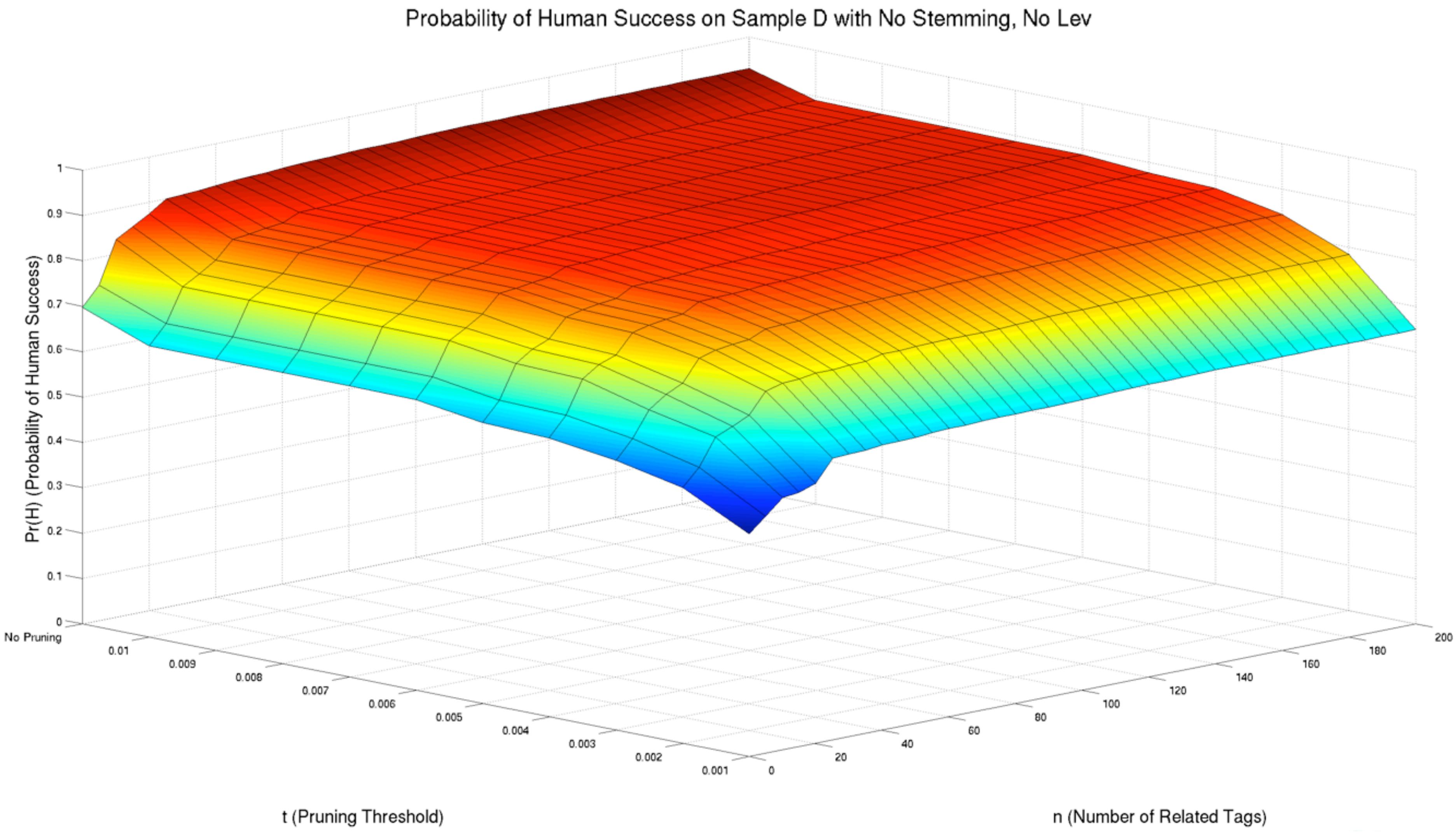
R·I·T

Experiment 1 (Tagging): Summary of Results

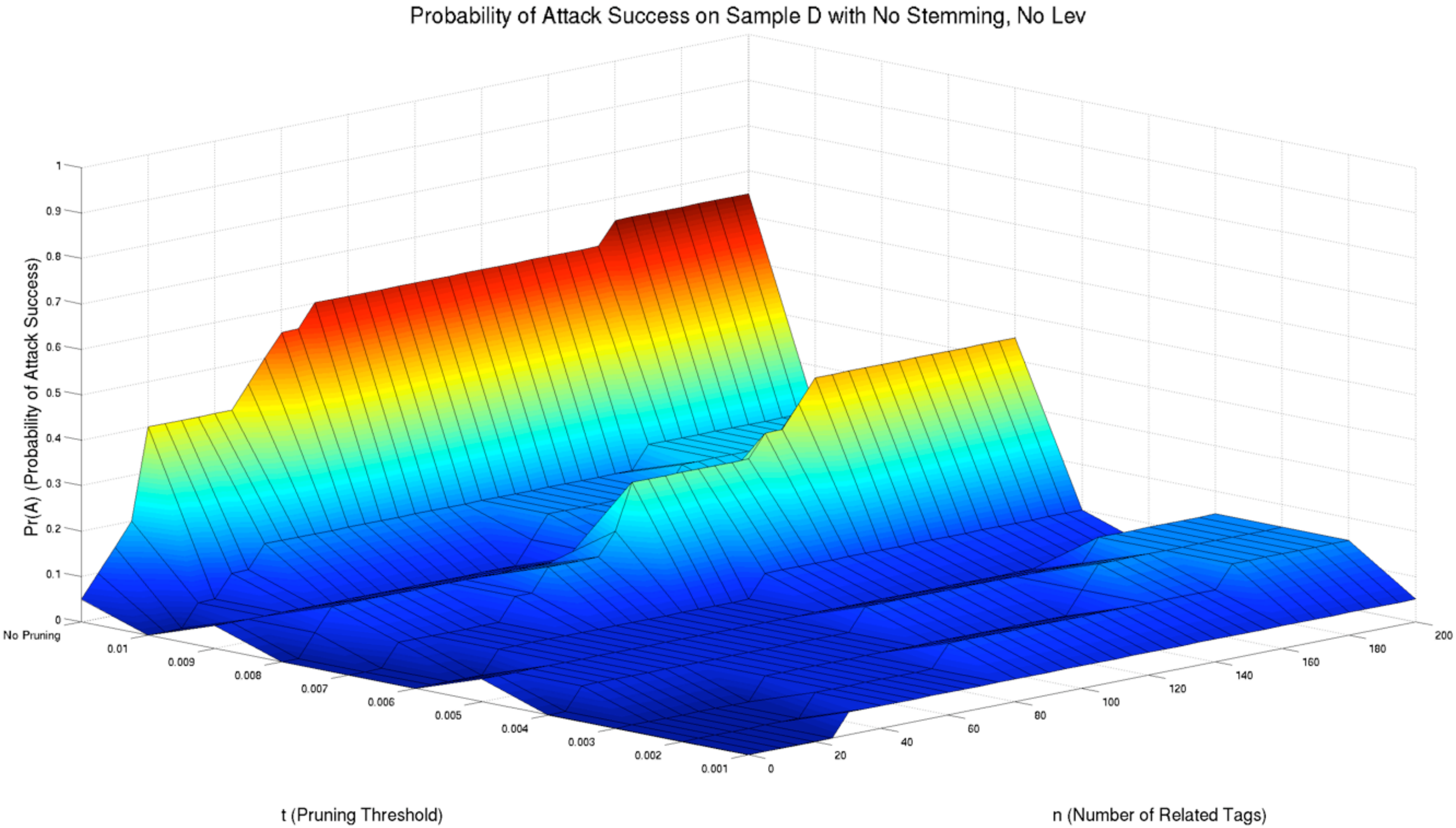


Condition		n	t	s	l	$P_r(H):A$	$P_r(A):C$	Gap
0	Control	0	1.0			0.7500	0.1286	0.6214
1	Most Usable	110	0.005			0.9101	0.1222	0.7879
2	Most Secure	5	0.003			0.7517	0.0128	0.7389
3	Largest Gap	25	0.005			0.8762	0.0402	0.8359
4	Most Usable	105	0.006	✓		0.9199	0.1273	0.7926
5	Most Secure	5	0.003	✓		0.7720	0.0124	0.7596
6	Largest Gap	15	0.006	✓		0.8769	0.0348	0.8421
7	Most Usable	100	0.006		✓	0.9273	0.1281	0.7992
8	Most Secure	5	0.003		✓	0.7682	0.0134	0.7548
9	Largest Gap	15	0.006		✓	0.8779	0.0381	0.8399
10	Most Usable	95	0.006	✓	✓	0.9343	0.1284	0.8058
11	Most Secure	5	0.003	✓	✓	0.7790	0.0134	0.7656
12	Largest Gap	15	0.006	✓	✓	0.8874	0.0379	0.8495

Human Rates Exp 2: 20 Videos, Random Walk

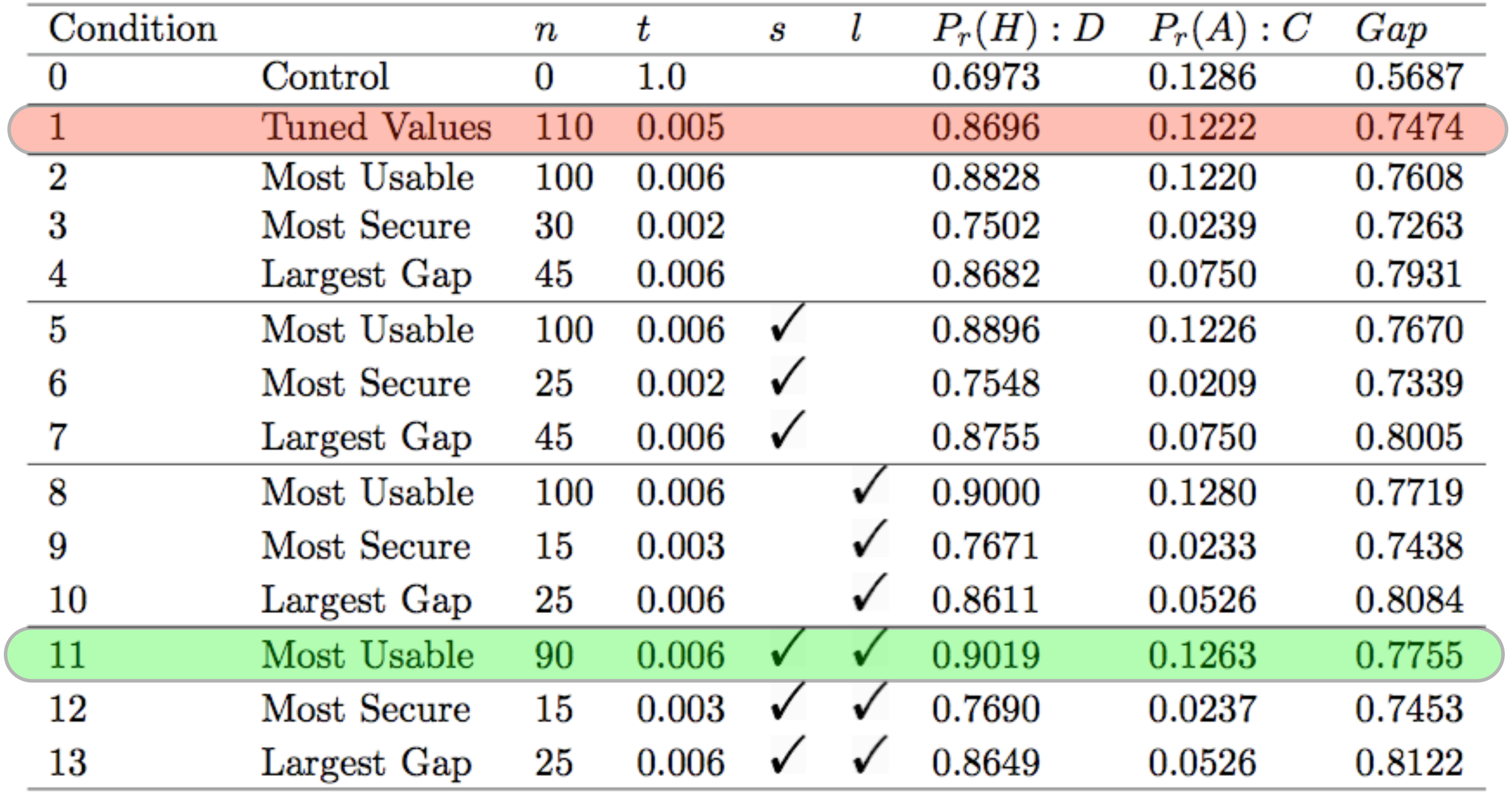


Attack Rates Exp 2: 20 Videos, Random Walk



R·I·T

Video CAPTCHA (Exp 2) and Attack Simulation Results



Condition		n	t	s	l	$P_r(H):D$	$P_r(A):C$	Gap
0	Control	0	1.0			0.6973	0.1286	0.5687
1	Tuned Values	110	0.005			0.8696	0.1222	0.7474
2	Most Usable	100	0.006			0.8828	0.1220	0.7608
3	Most Secure	30	0.002			0.7502	0.0239	0.7263
4	Largest Gap	45	0.006			0.8682	0.0750	0.7931
5	Most Usable	100	0.006	✓		0.8896	0.1226	0.7670
6	Most Secure	25	0.002	✓		0.7548	0.0209	0.7339
7	Largest Gap	45	0.006	✓		0.8755	0.0750	0.8005
8	Most Usable	100	0.006		✓	0.9000	0.1280	0.7719
9	Most Secure	15	0.003		✓	0.7671	0.0233	0.7438
10	Largest Gap	25	0.006		✓	0.8611	0.0526	0.8084
11	Most Usable	90	0.006	✓	✓	0.9019	0.1263	0.7755
12	Most Secure	15	0.003	✓	✓	0.7690	0.0237	0.7453
13	Largest Gap	25	0.006	✓	✓	0.8649	0.0526	0.8122

Completion Times and User Preferences

Completion times (in seconds)

- Tagging Exp: median = 20.6 seconds (μ = 29.7, σ = 34.7)
- CAPTCHA Exp: median = 17.1 seconds (μ = 22.0, σ = 23.6)

Which task is faster?

- 16%: neither 64%: text 20%: video (Tagging Experiment)
- 13%: neither 60%: text 27%: video (CAPTCHA Experiment)

Which task is more enjoyable?

- 23%: no pref 15%: text 62%: video (Tagging Experiment)
- 22%: no pref 20%: text 58%: video (CAPTCHA Experiment)

Comparison with Other Methods

CAPTCHA Name	Type	$P_r(H)$	$P_r(A)$
Microsoft [3]	Text-based	0.90 [3]	0.60 [23]
Baffletext [4]	Text-based	0.89 [4]	0.25 [4]
Handwritten [19]	Text-based	0.76 [19]	0.13 [19]
ASIRRA [6]	Image-based	0.99 [6]	0.10 [8]
Video [13]	Video	0.90 [13]	0.13 [13]

[13] K. Kluever and R. Zanibbi. (2008) Video CAPTCHAs: Usability vs. Security. Proc. IEEE Western New York Image Processing Workshop, Rochester, NY (USA) (extended abstract).

Conclusion

Summary

- First attempt at using video for CAPTCHAs
- Meets CAPTCHA criteria; semi-automated
- Usability & security comparable to existing techniques
- Small majority of participants report preferring video to text CAPTCHAs (altern.?)

To do....

Other attacks

e.g. CBIR; adapting task for these

Accessibility

Effect of audio/video only?

Localization

Use different dictionaries to ‘seed’ random walks, different video databases

Other domains

Tag generation mechanism is not video-specific

Document and Pattern Recognition Lab, RIT

Primary Aims

Improve theories and tools for constructing recognition systems (e.g. Rec. Strategy Lang.)

Document recognition applications (online and offline)

DPRL: Members

Master's Students

Ling Ouyang *(OCR for math symbols)*

Ramesh Muraleedharan *(CAPTCHAs)*

Amit Pillay *(Combining structural pattern recognizers/RSL)*

Li Yu *(Content-based image retrieval for math)*

Collaborators

Matthew Casey

Research Assistants

Adam Risi, Ben Hughes

Thank You.

Acknowledgements

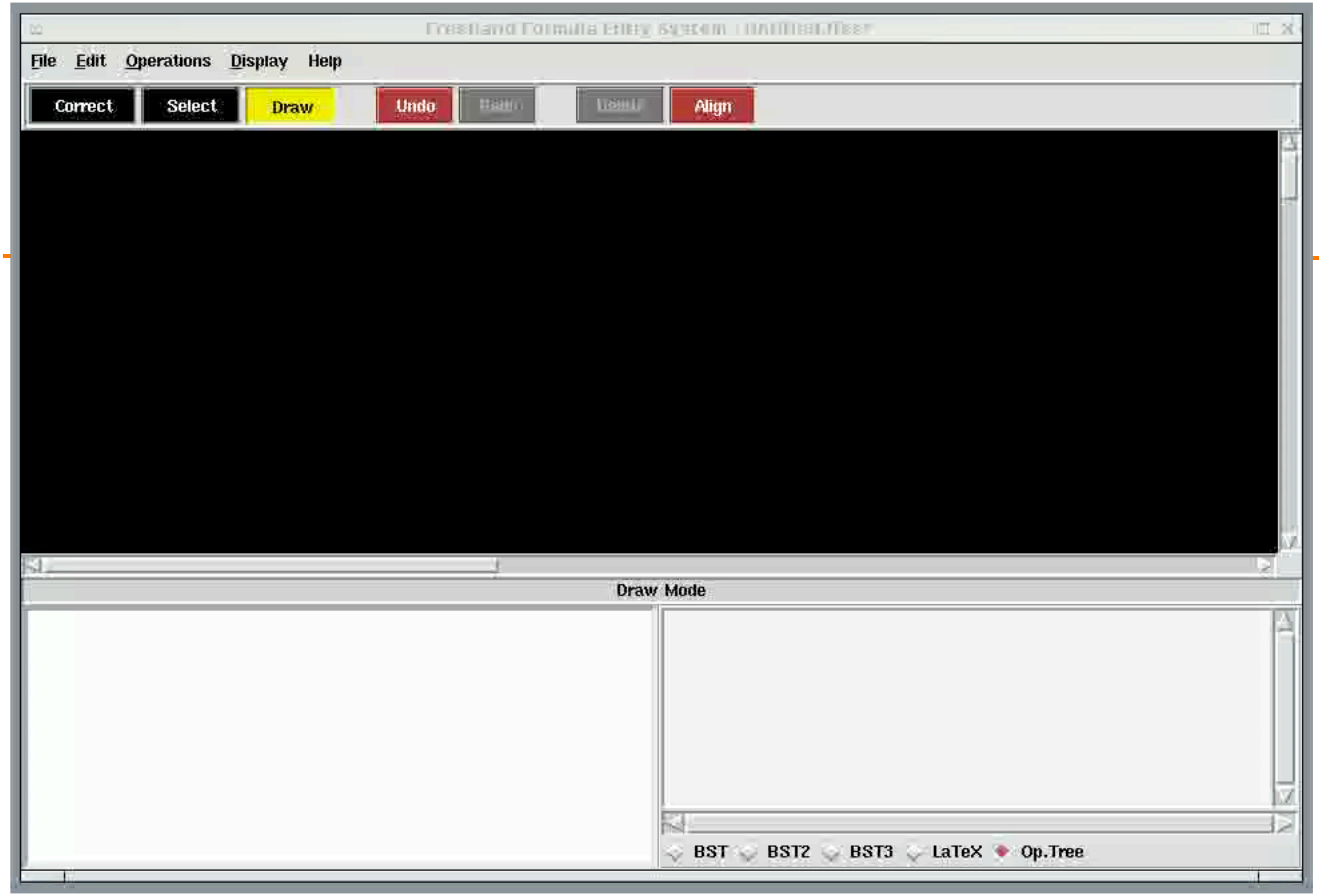
Xerox corporation (UAC grant)

- Bill Stumbo, XWRC

Matthew Casey, U. Surrey Dept. CS

Online Demonstration:

http://sudbury.cs.rit.edu/

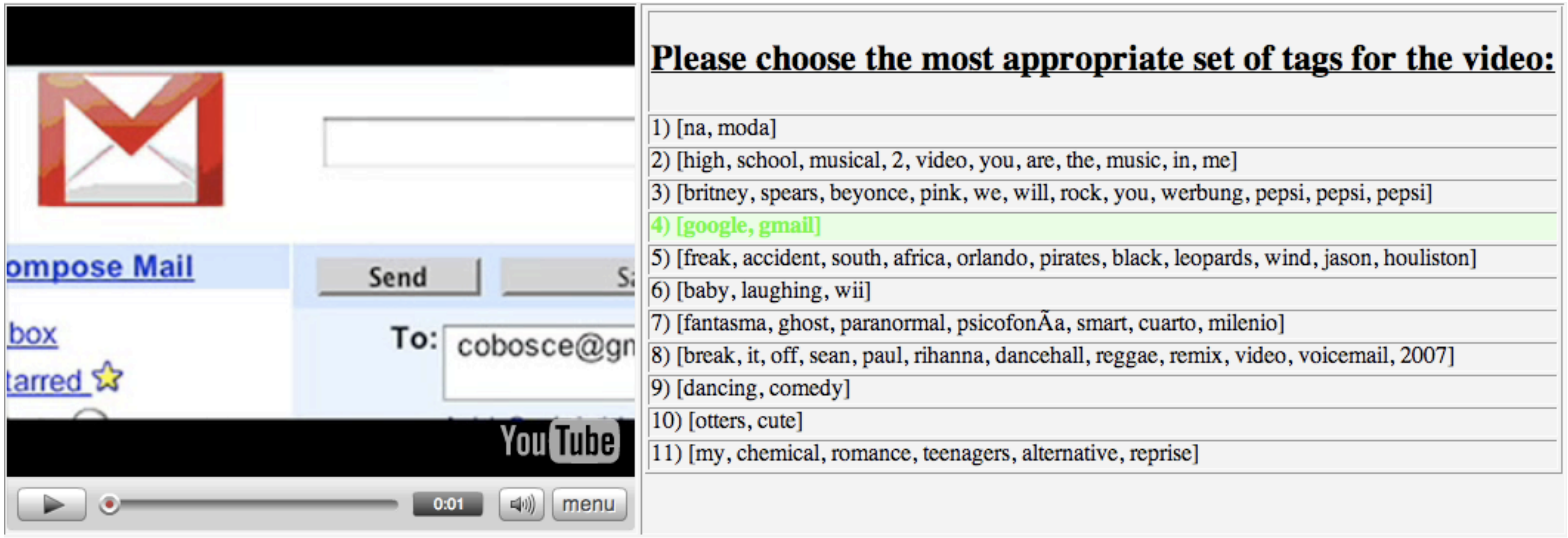

Freehand Formula Entry System
File
Edit
Operations
Display
Help
Correct
Select
Draw
Undo
Align
Draw Mode
○ BST
○ BST2
○ BST3
○ LaTeX
● Op.Tree

Video CAPTCHA Design

Ask a specific question about the video

- “What color shirt was the man wearing?”

Ask which set of tags best matches



Please choose the most appropriate set of tags for the video:

1) [na, moda]
2) [high, school, musical, 2, video, you, are, the, music, in, me]
3) [britney, spears, beyonce, pink, we, will, rock, you, werbung, pepsi, pepsi, pepsi]
4) [google, gmail]
5) [freak, accident, south, africa, orlando, pirates, black, leopards, wind, jason, houliston]
6) [baby, laughing, wii]
7) [fantasma, ghost, paranormal, psicofonÃa, smart, cuarto, milenio]
8) [break, it, off, sean, paul, rihanna, dancehall, reggae, remix, video, voicemail, 2007]
9) [dancing, comedy]
10) [otters, cute]
11) [my, chemical, romance, teenagers, alternative, reprise]

Ask for tags about the video

- “man shirt green”