Leveraging the CAPTCHA Problem

Daniel Lopresti

Computer Science & Engineering
Lehigh University
Bethlehem, PA 18015, USA
lopresti@cse.lehigh.edu

Leveraging the CAPTCHA Problem

Daniel Lopresti

Computer Science & Engineering
Lehigh University
Bethlehem, PA 18015, USA
lopresti@cse.lehigh.edu

CAPTCHA's

Goal is to prevent automated attacks on networked 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.

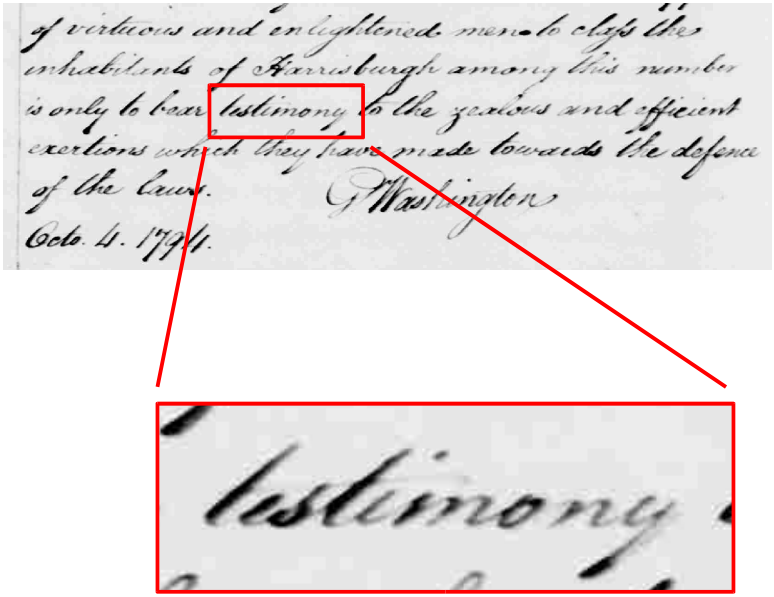
Points to Ponder

- 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?

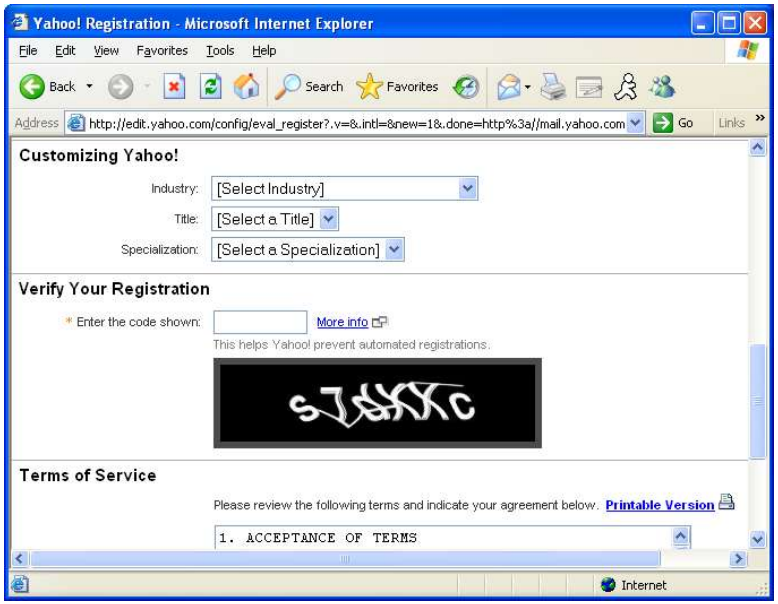
Cause for hope: variety of pattern recognition tasks in real world is almost endless. Note apparent disconnect, however, between natural tasks and synthetic versions we use for CAPTCHA's.

Natural vs. Synthetic #1



What word do you see in the box?

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



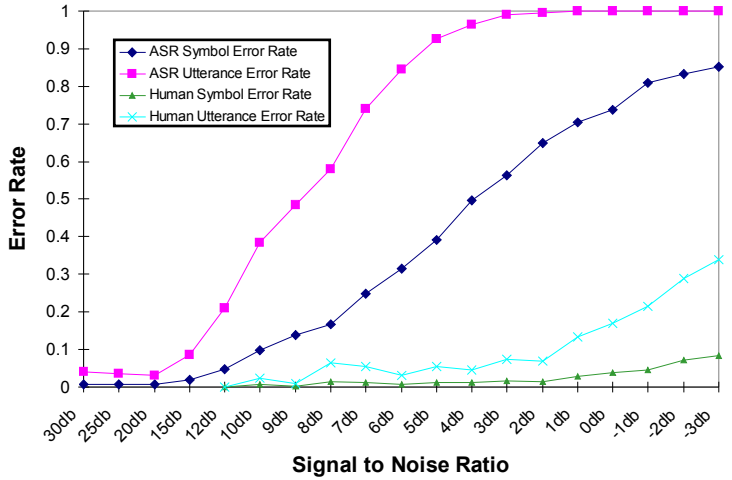
What "word" do you see in the box?

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



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.

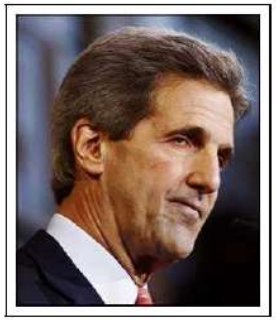
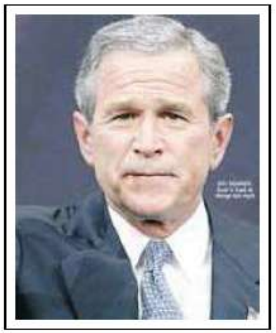
Where Do CAPTCHA's Grow?

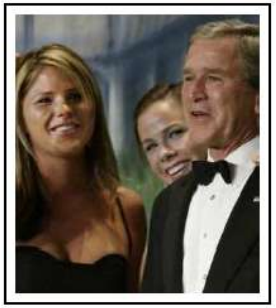
Describe the weather in this scene.

From WABC Central Park WebCam, http://abclocal.go.com/kabc/features/cams/082102_central_Park_cam.html

Where Do CAPTCHA's Grow?

Which photos show the same person?

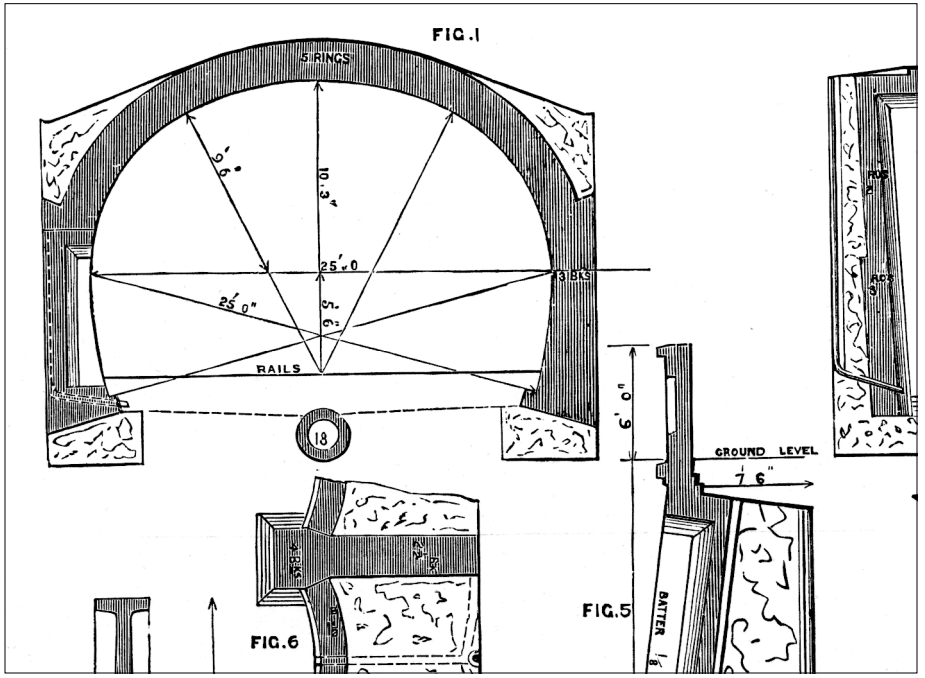

Where Do CAPTCHA's Grow?

How many cars do you see in this image?

From WCPO Cincinnati Ohio Skycam, http://webcambiglook.com/cinn_skycam.html

Where Do CAPTCHA's Grow?

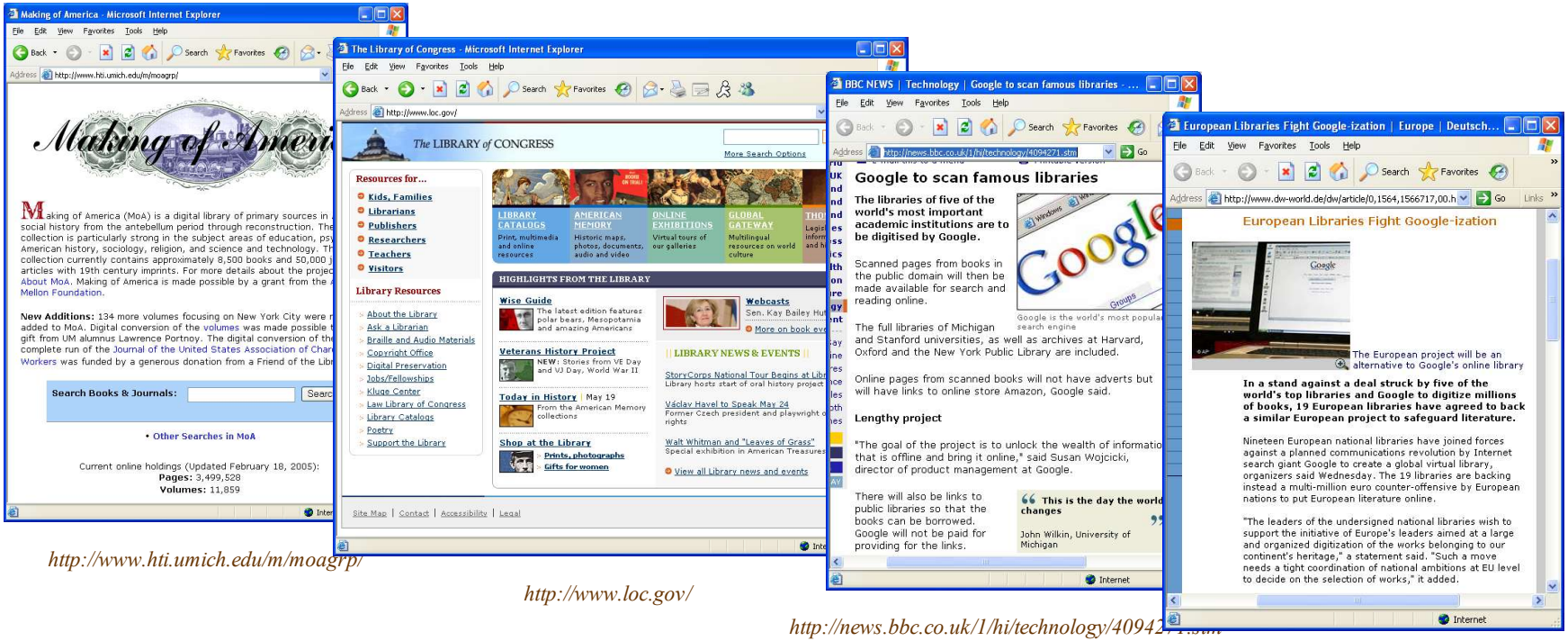
Draw a box around a text string in this image.



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

Where Do CAPTCHA's Grow?

One obvious answer: in digital libraries.



http://www.hti.umich.edu/m/moagrp/

http://www.loc.gov/

http://news.bbc.co.uk/1/hi/technology/4094271.stm

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

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

Something is Missing

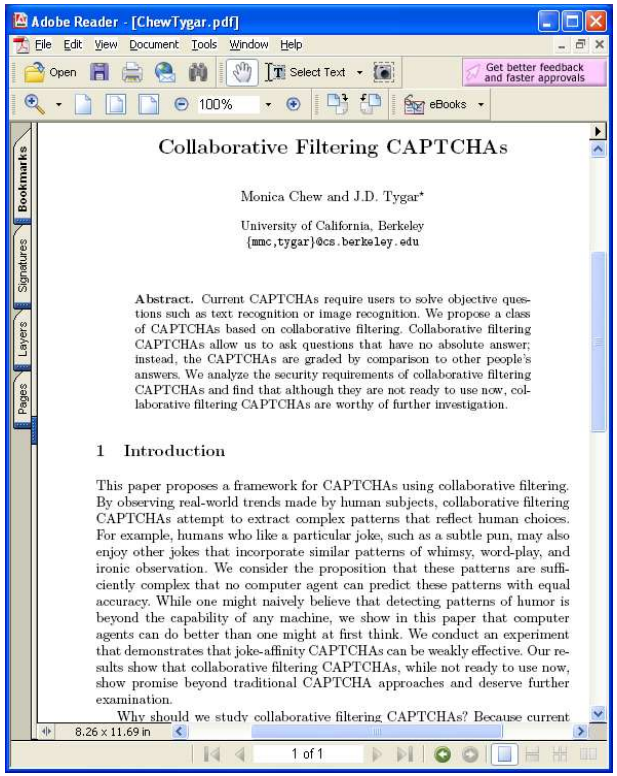
To grade response to a CAPTCHA challenge, we need to know "correct" answer (or, rather, how a human would respond).

- 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

See paper by Monica Chew and Doug Tygar in this workshop:

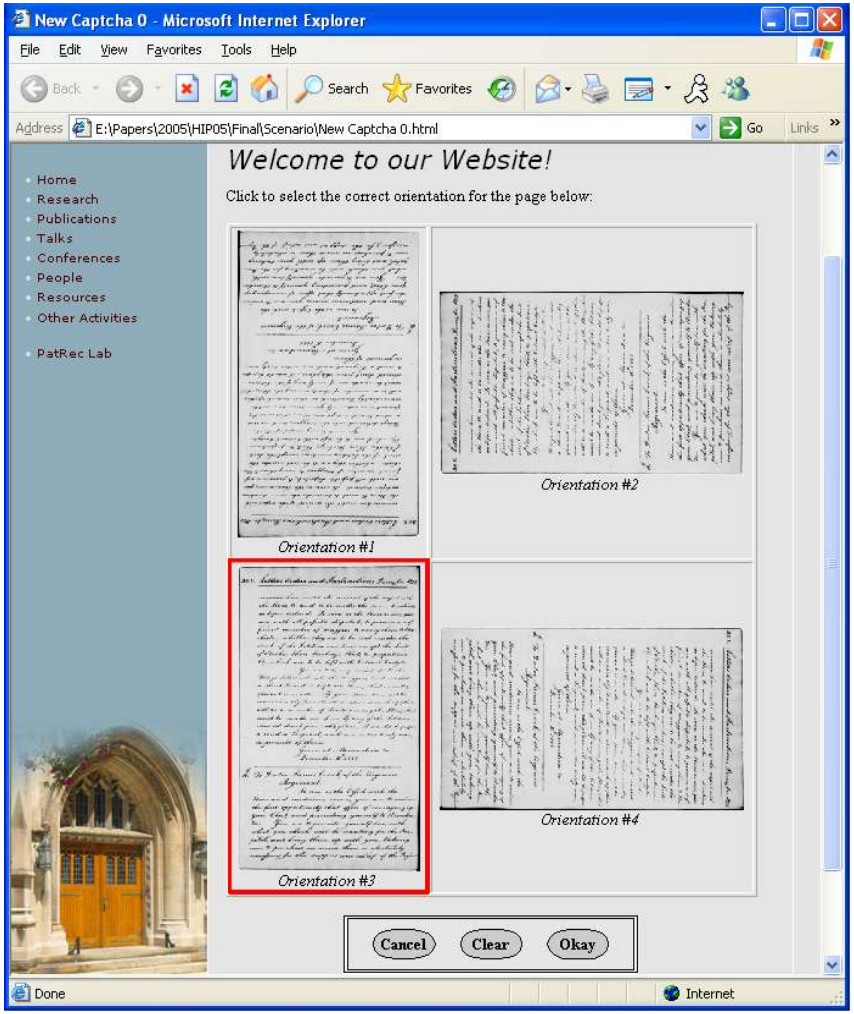


- Bootstrap from tests with known answers.
- Require users to solve more than one CAPTCHA.
- Collect responses to new candidate CAPTCHA's from proven humans to grow collection of available tests.

Human Interactive Proofs: Second International Workshop, Springer LNCS Volume 3517, May 2005, pp. 66-81.

Scenario 1

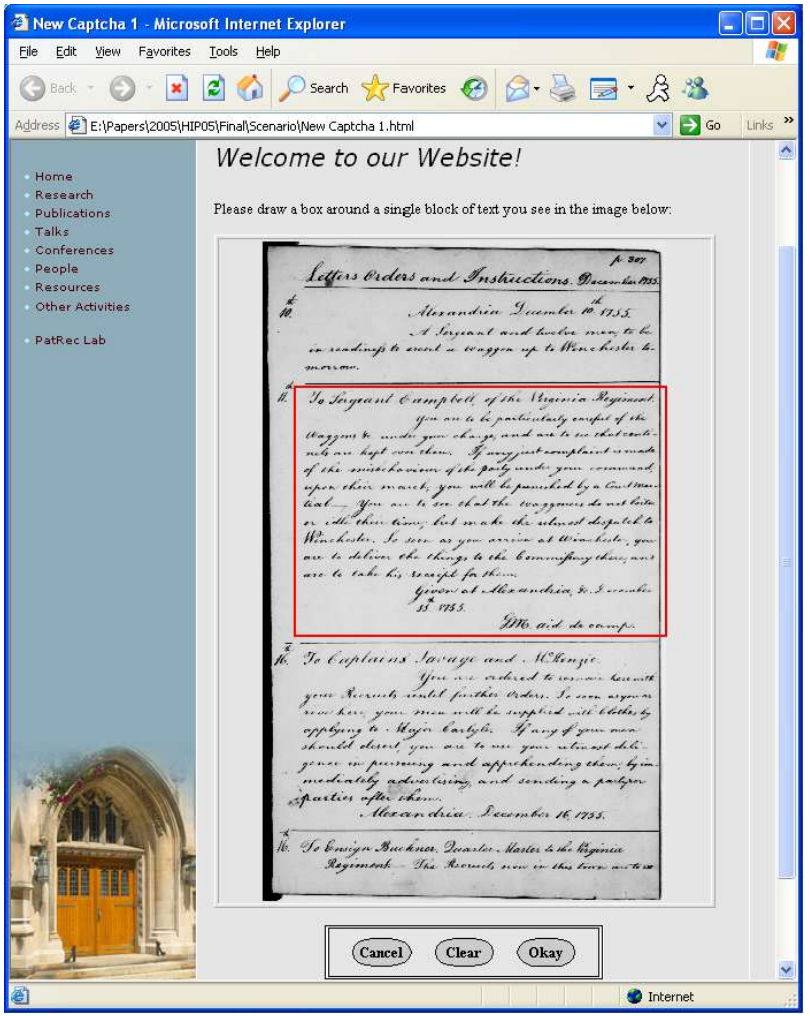
Click to select the correct orientation for this page.



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

Scenario 2

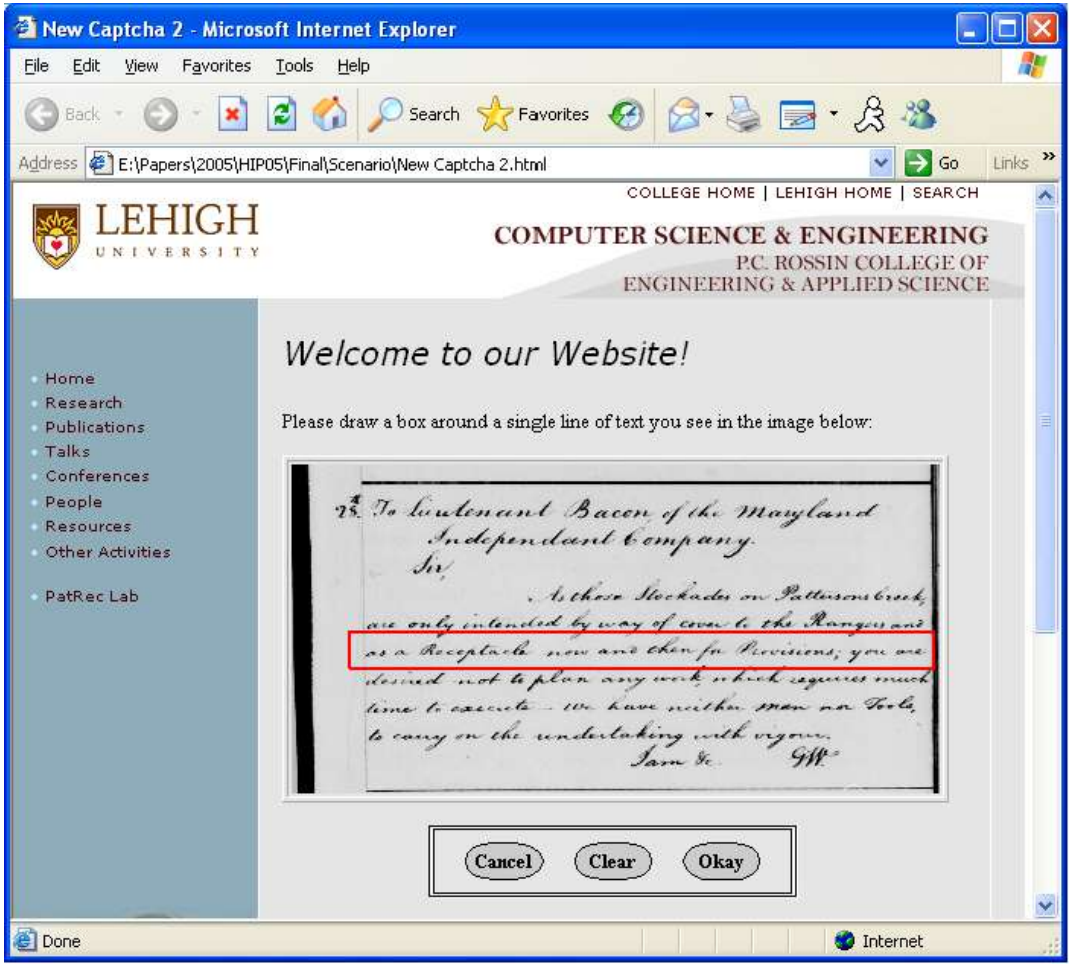
Please draw a box around a single block of text you see in the image.



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

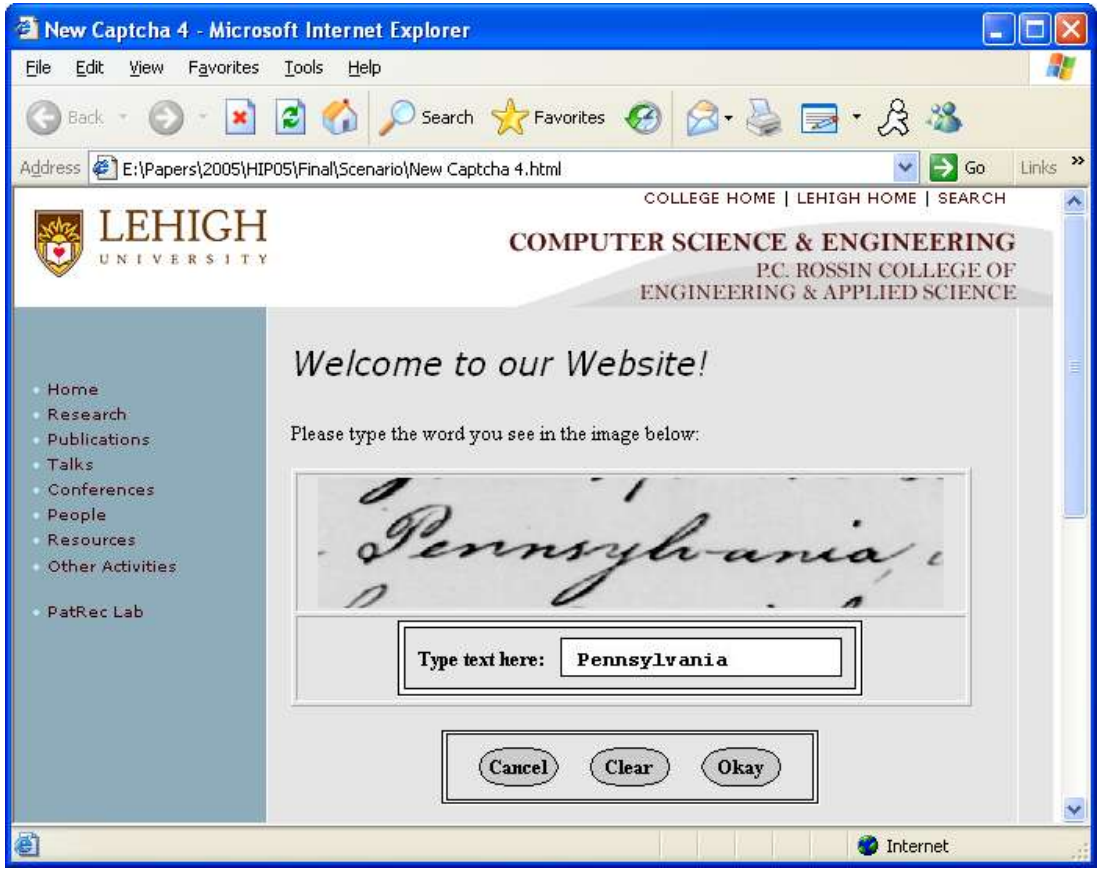
Scenario 4

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.

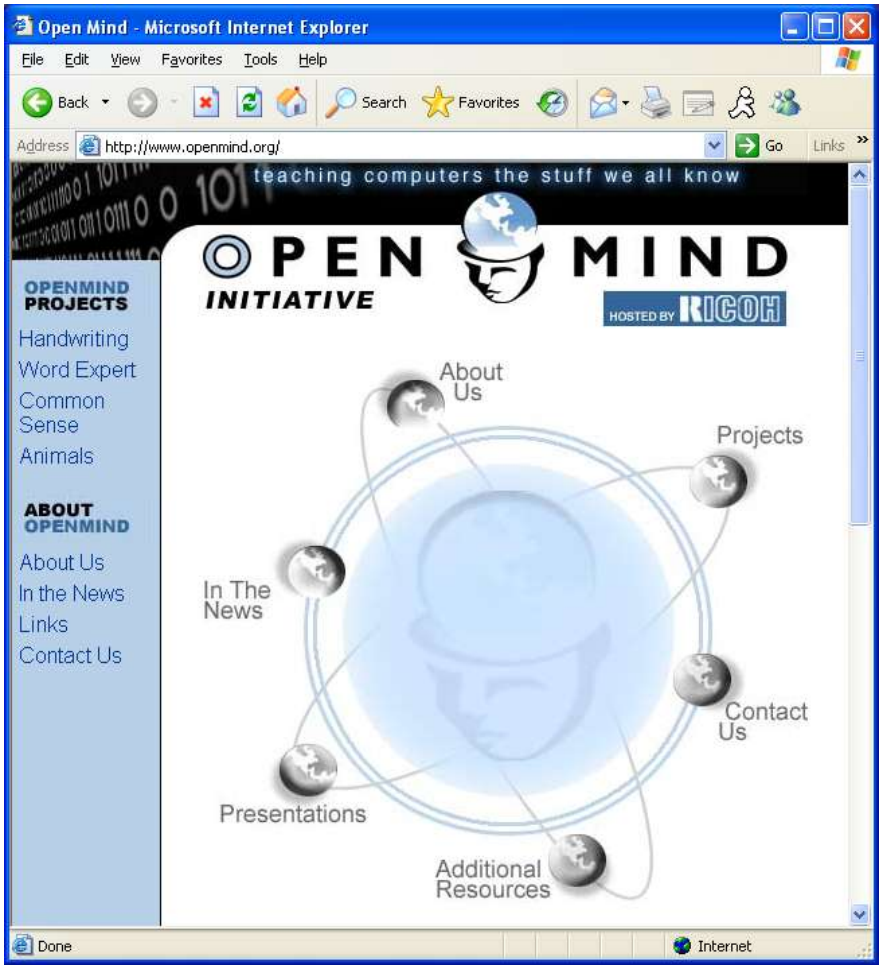
Note that even if transcript for document is available online, most of the information we asked for typically isn't:

- 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.

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 such CAPTCHA's suggests web service model (see paper by Tim Converse in this workshop).
- 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?

A Challenge

Building systems like this and performing large-scale CAPTCHA studies requires serious time and money.

Some company (or companies) with substantial resources should team with university researchers to study whether idea is feasible.

Risk / reward ratio seems highly favorable.