Old Dominion University
ODU Digital Commons

Computer Science Presentations Computer Science

4-11-2016

Combining Heritrix and PhantomJS for Better Crawling of Pages with Javascript

Justin F. Brunelle
Old Dominion University, jbrun008@odu.edu

Michele C. Weigle
Old Dominion University, mweigle@odu.edu

Michael L. Nelson
Old Dominion University, mnelson@odu.edu

Follow this and additional works at: https://digitalcommons.odu.edu/computerscience_presentations

Part of the Archival Science Commons

Recommended Citation
Brunelle, Justin F.; Weigle, Michele C.; and Nelson, Michael L., "Combining Heritrix and PhantomJS for Better Crawling of Pages with Javascript" (2016). *Computer Science Presentations*. 1.
https://digitalcommons.odu.edu/computerscience_presentations/1

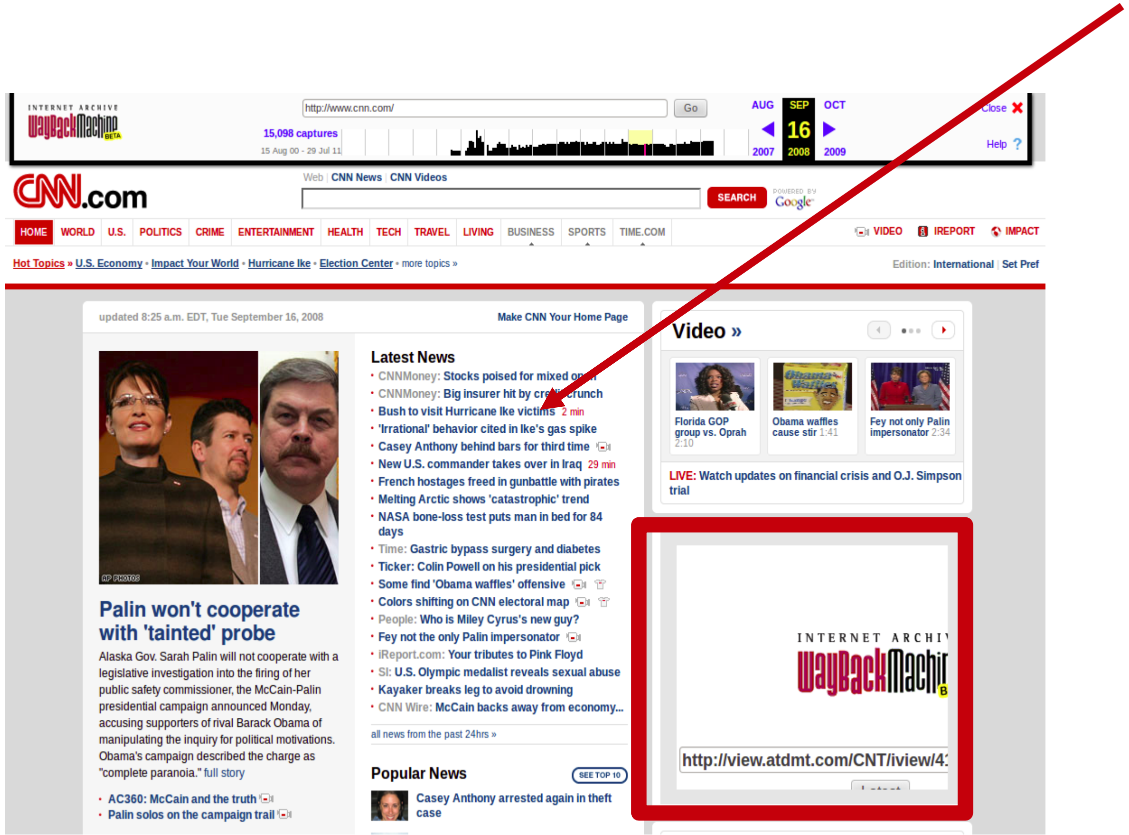
This Book is brought to you for free and open access by the Computer Science at ODU Digital Commons. It has been accepted for inclusion in Computer Science Presentations by an authorized administrator of ODU Digital Commons. For more information, please contact digitalcommons@odu.edu.

Combining Heritrix and PhantomJS for Better Crawling of Pages with Javascript

Justin F. Brunelle
Michele C. Weigle
Michael L. Nelson
Web Science and Digital Libraries Research Group
Old Dominion University
@WebSciDL

IIPC 2016
Reykjavik, Iceland, April 11, 2016

Javascript can create missing resources (bad)



2008

INTERNET ARCHIVE WayBackMachine BETA

http://www.cnn.com/ Go

15,098 captures

15 Aug 00 - 29 Jul 11

AUG SEP OCT

2007 16 2008 2009

Close ✖

Help ?

CNN.com

Web | CNN News | CNN Videos

SEARCH POWERED BY Google

HOME | WORLD | U.S. | POLITICS | CRIME | ENTERTAINMENT | HEALTH | TECH | TRAVEL | LIVING | BUSINESS | SPORTS | TIME.COM | VIDEO | IREPORT | IMPACT

Hot Topics » U.S. Economy • Impact Your World • Hurricane Ike • Election Center • more topics »

Edition: International | Set Pref

updated 8:25 a.m. EDT, Tue September 16, 2008

Make CNN Your Home Page

AP PHOTOS

Palin won't cooperate with 'tainted' probe

Alaska Gov. Sarah Palin will not cooperate with a legislative investigation into the firing of her public safety commissioner, the McCain-Palin presidential campaign announced Monday, accusing supporters of rival Barack Obama of manipulating the inquiry for political motivations. Obama's campaign described the charge as "complete paranoia." full story

- AC360: McCain and the truth
- Palin solos on the campaign trail

Latest News

- CNNMoney: Stocks poised for mixed open
- CNNMoney: Big insurer hit by credit crunch
- Bush to visit Hurricane Ike victims 2 min
- 'Irrational' behavior cited in Ike's gas spike
- Casey Anthony behind bars for third time
- New U.S. commander takes over in Iraq 29 min
- French hostages freed in gunbattle with pirates
- Melting Arctic shows 'catastrophic' trend
- NASA bone-loss test puts man in bed for 84 days
- Time: Gastric bypass surgery and diabetes
- Ticker: Colin Powell on his presidential pick
- Some find 'Obama waffles' offensive
- Colors shifting on CNN electoral map
- People: Who is Miley Cyrus's new guy?
- Fey not the only Palin impersonator
- iReport.com: Your tributes to Pink Floyd
- SI: U.S. Olympic medalist reveals sexual abuse
- Kayaker breaks leg to avoid drowning
- CNN Wire: McCain backs away from economy...

all news from the past 24hrs »

Popular News

SEE TOP 10

Casey Anthony arrested again in theft case

Video »

Florida GOP group vs. Oprah 2:10

Obama waffles cause stir 1:41

Fey not only Palin impersonator 2:34

LIVE: Watch updates on financial crisis and O.J. Simpson trial

INTERNET ARCHIV

WayBackMachi

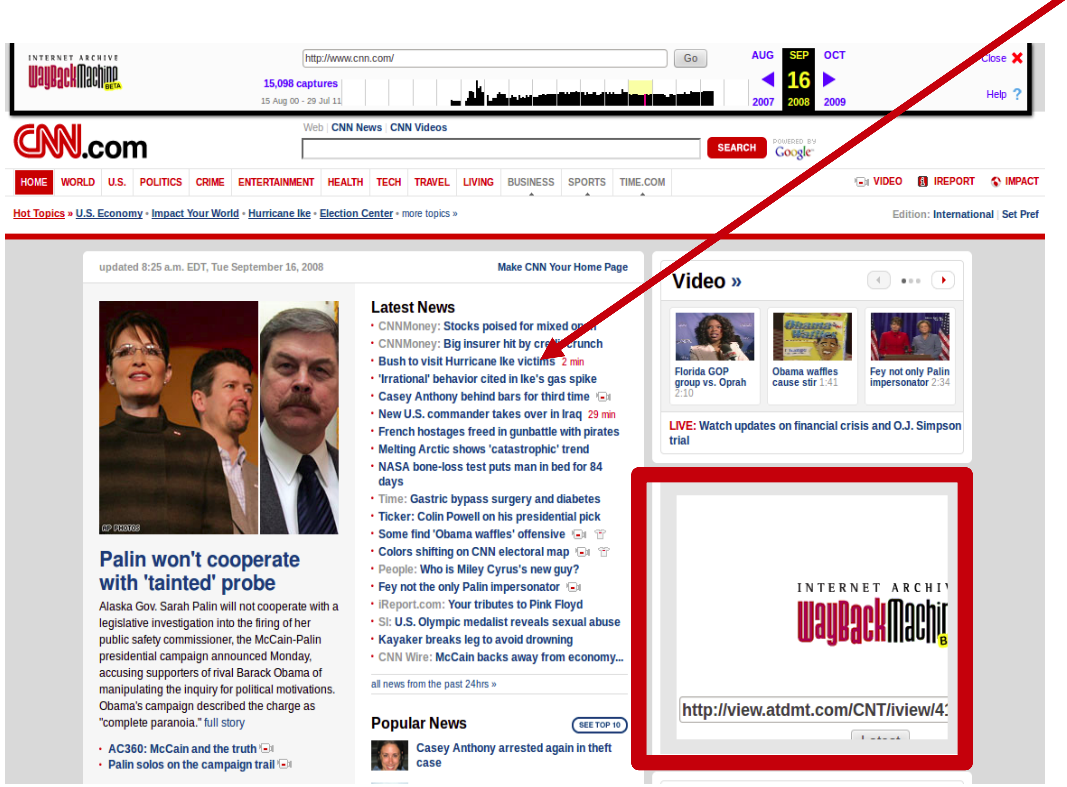
http://view.atdmt.com/CNT/iview/41

http://ws-dl.blogspot.com/2012/10/2012-10-10-zombies-in-archives.html

Javascript can create missing resources (bad) or Temporal violations (worse)

2008

2012



INTERNET ARCHIVE WayBackMachine BETA

http://www.cnn.com/ Go

15,098 captures

15 Aug 00 - 29 Jul 11

AUG SEP OCT

16

2007 2008 2009

Close ✖

Help ?

CNN.com

Web | CNN News | CNN Videos

SEARCH POWERED BY Google

HOME | WORLD | U.S. | POLITICS | CRIME | ENTERTAINMENT | HEALTH | TECH | TRAVEL | LIVING | BUSINESS | SPORTS | TIME.COM

VIDEO | IREPORT | IMPACT

Hot Topics » U.S. Economy • Impact Your World • Hurricane Ike • Election Center • more topics »

Edition: International | Set Pref

updated 8:25 a.m. EDT, Tue September 16, 2008

Make CNN Your Home Page

AP PHOTOS

Palin won't cooperate with 'tainted' probe

Alaska Gov. Sarah Palin will not cooperate with a legislative investigation into the firing of her public safety commissioner, the McCain-Palin presidential campaign announced Monday, accusing supporters of rival Barack Obama of manipulating the inquiry for political motivations. Obama's campaign described the charge as "complete paranoia." full story

- AC360: McCain and the truth
- Palin solos on the campaign trail

Latest News

- CNNMoney: Stocks poised for mixed open
- CNNMoney: Big insurer hit by credit crunch
- Bush to visit Hurricane Ike victims 2 min
- 'Irrational' behavior cited in Ike's gas spike
- Casey Anthony behind bars for third time
- New U.S. commander takes over in Iraq 29 min
- French hostages freed in gunbattle with pirates
- Melting Arctic shows 'catastrophic' trend
- NASA bone-loss test puts man in bed for 84 days
- Time: Gastric bypass surgery and diabetes
- Ticker: Colin Powell on his presidential pick
- Some find 'Obama waffles' offensive
- Colors shifting on CNN electoral map
- People: Who is Miley Cyrus's new guy?
- Fey not the only Palin impersonator
- iReport.com: Your tributes to Pink Floyd
- SI: U.S. Olympic medalist reveals sexual abuse
- Kayaker breaks leg to avoid drowning
- CNN Wire: McCain backs away from economy...

all news from the past 24hrs »

Popular News SEE TOP 10

Casey Anthony arrested again in theft case

Video »

Florida GOP group vs. Oprah 2:10

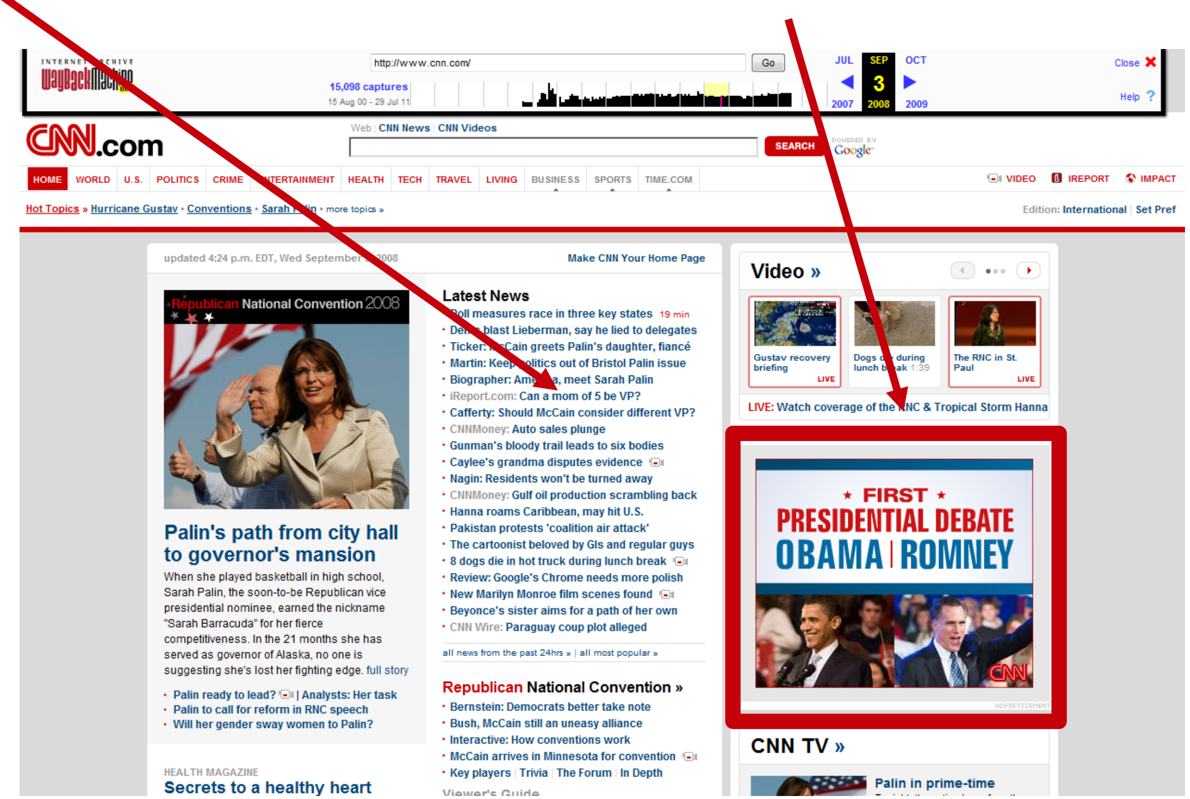
Obama waffles cause stir 1:41

Fey not only Palin impersonator 2:34

LIVE: Watch updates on financial crisis and O.J. Simpson trial

INTERNET ARCHIVE WayBackMachine

http://view.atdmt.com/CNT/iview/4



INTERNET ARCHIVE WayBackMachine

http://www.cnn.com/ Go

15,098 captures

15 Aug 00 - 29 Jul 11

JUL SEP OCT

3

2007 2008 2009

Close ✖

Help ?

CNN.com

Web | CNN News | CNN Videos

SEARCH POWERED BY Google

HOME | WORLD | U.S. | POLITICS | CRIME | ENTERTAINMENT | HEALTH | TECH | TRAVEL | LIVING | BUSINESS | SPORTS | TIME.COM

VIDEO | IREPORT | IMPACT

Hot Topics » Hurricane Gustav • Conventions • Sarah Palin • more topics »

Edition: International | Set Pref

updated 4:24 p.m. EDT, Wed September 3, 2008

Make CNN Your Home Page

Republican National Convention 2008

Palin's path from city hall to governor's mansion

When she played basketball in high school, Sarah Palin, the soon-to-be Republican vice presidential nominee, earned the nickname "Sarah Barracuda" for her fierce competitiveness. In the 21 months she has served as governor of Alaska, no one is suggesting she's lost her fighting edge. full story

- Palin ready to lead? | Analysts: Her task
- Palin to call for reform in RNC speech
- Will her gender sway women to Palin?

HEALTH MAGAZINE

Secrets to a healthy heart

Latest News

- Poll measures race in three key states 19 min
- Dems blast Lieberman, say he lied to delegates
- Ticker: McCain greets Palin's daughter, fiancé
- Martin: Keep politics out of Bristol Palin issue
- Biographer: America, meet Sarah Palin
- iReport.com: Can a mom of 5 be VP?
- Cafferty: Should McCain consider different VP?
- CNNMoney: Auto sales plunge
- Gunman's bloody trail leads to six bodies
- Caylee's grandma disputes evidence
- Nagin: Residents won't be turned away
- CNNMoney: Gulf oil production scrambling back
- Hanna roams Caribbean, may hit U.S.
- Pakistan protests 'coalition air attack'
- The cartoonist beloved by GIs and regular guys
- 8 dogs die in hot truck during lunch break
- Review: Google's Chrome needs more polish
- New Marilyn Monroe film scenes found
- Beyonce's sister aims for a path of her own
- CNN Wire: Paraguay coup plot alleged

all news from the past 24hrs » | all most popular »

Republican National Convention »

- Bernstein: Democrats better take note
- Bush, McCain still an uneasy alliance
- Interactive: How conventions work
- McCain arrives in Minnesota for convention
- Key players | Trivia | The Forum | In Depth

Viewer's Guide

Video »

Gustav recovery briefing LIVE

Dogs die during lunch break 1:39

The RNC in St. Paul LIVE

LIVE: Watch coverage of the RNC & Tropical Storm Hanna

★ FIRST ★ PRESIDENTIAL DEBATE OBAMA | ROMNEY CNN

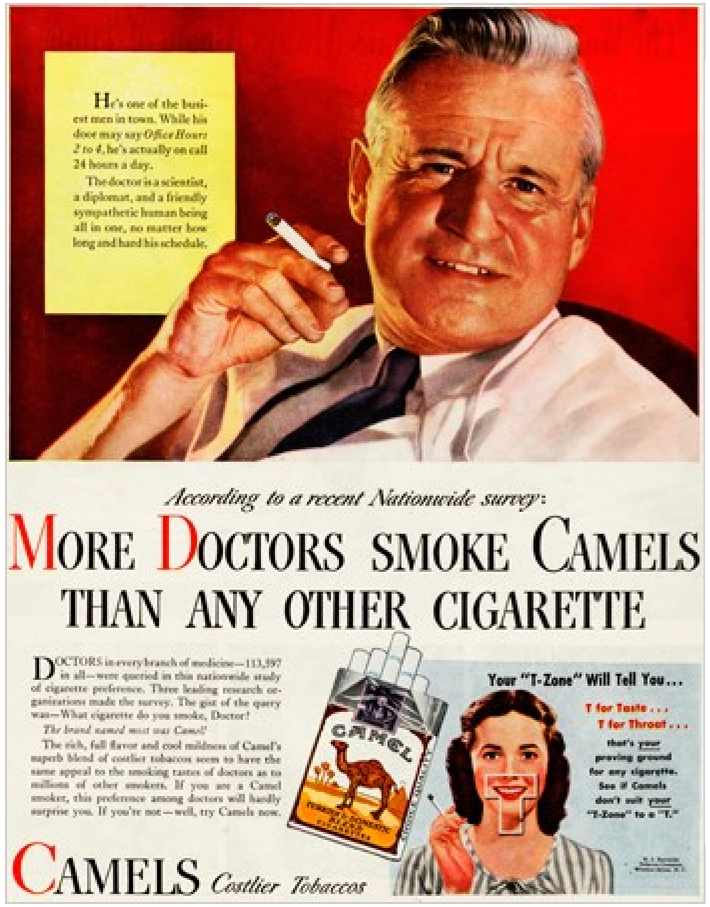
ADVERTISEMENT

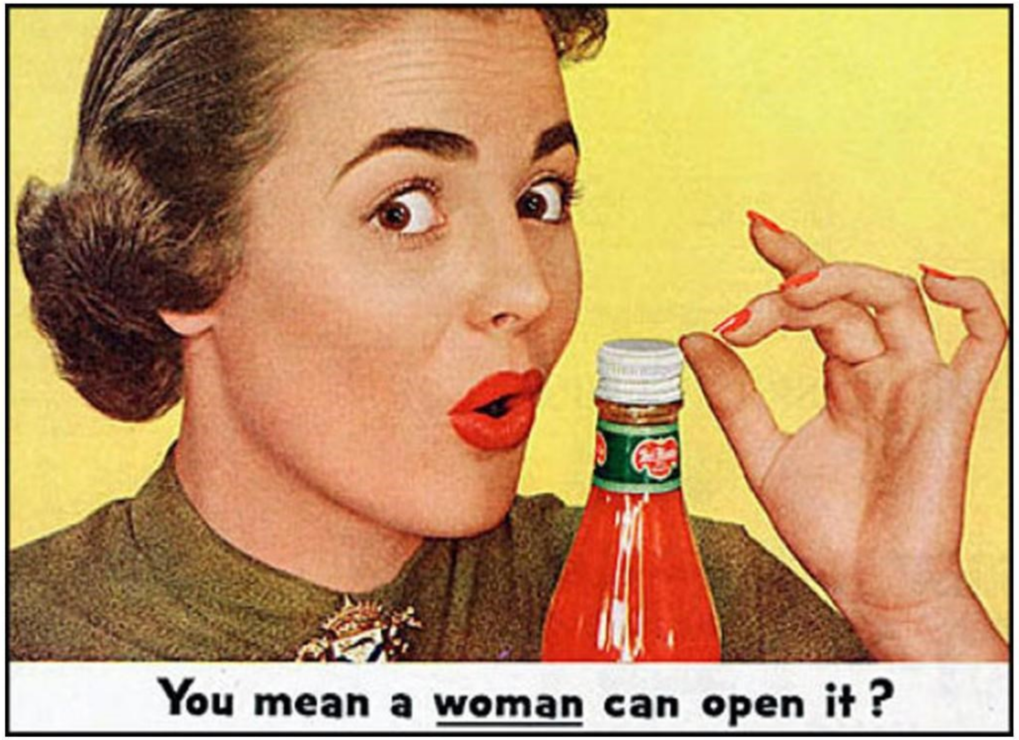
CNN TV »

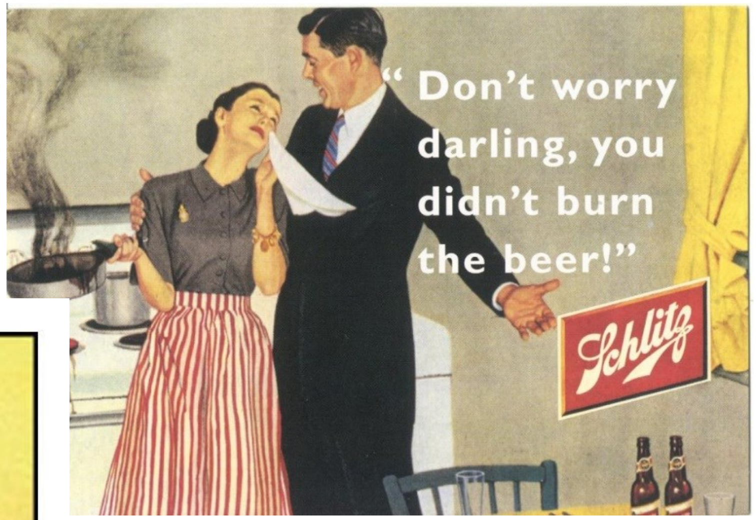
Palin in prime-time

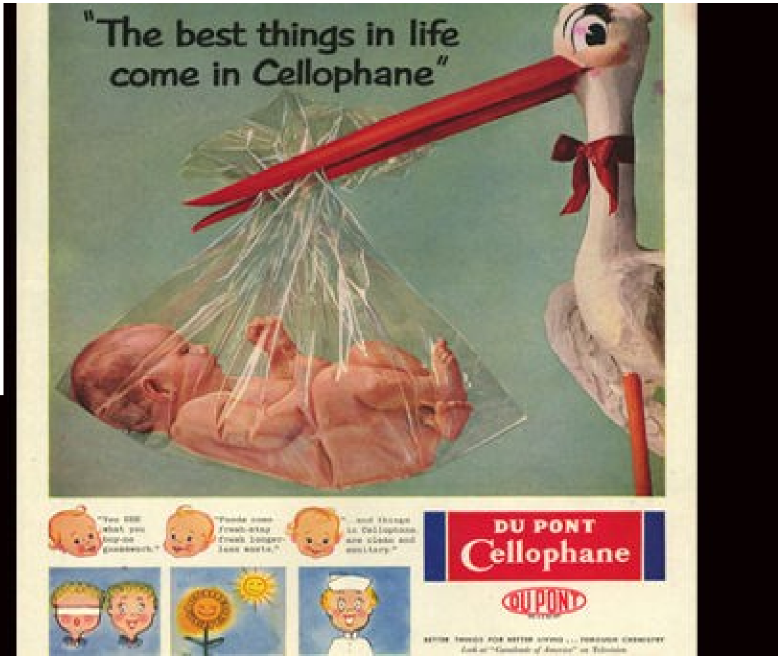
http://ws-dl.blogspot.com/2012/10/2012-10-10-zombies-in-archives.html

Old ads are interesting

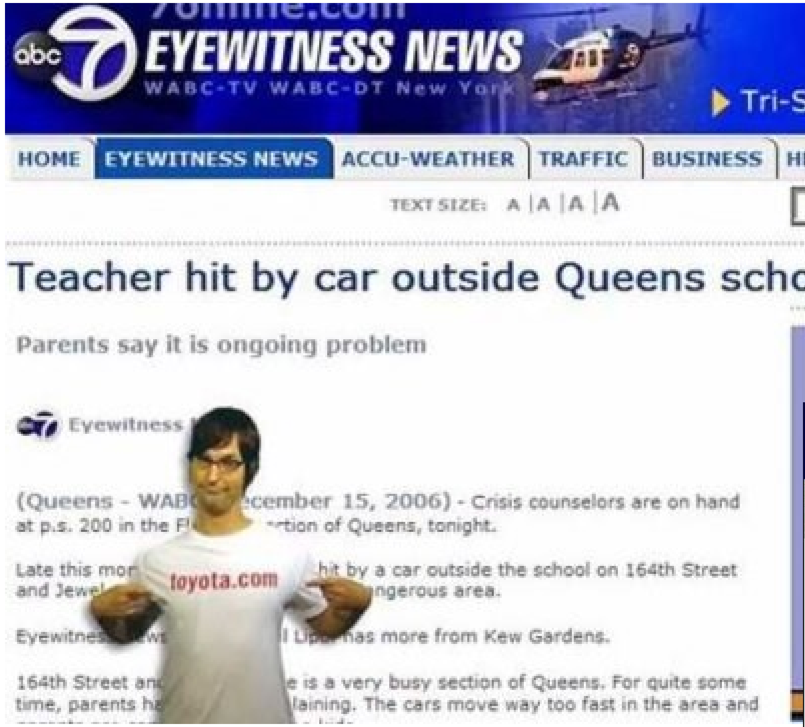

He's one of the busiest men in town. While his door may say *Office Hours 2 to 4*, he's actually on call 24 hours a day.
The doctor is a scientist, a diplomat, and a friendly sympathetic human being all in one, no matter how long and hard his schedule.
According to a recent Nationwide survey:
MORE DOCTORS SMOKE CAMELS
THAN ANY OTHER CIGARETTE
DOCTORS in every branch of medicine—113,597 in all—were queried in this nationwide study of cigarette preference. Three leading research organizations made the survey. The gist of the query was—What cigarette do you smoke, Doctor?
The brand named most was Camel!
The rich, full flavor and cool mildness of Camel's superb blend of costlier tobaccos seem to have the same appeal to the smoking tastes of doctors as to millions of other smokers. If you are a Camel smoker, this preference among doctors will hardly surprise you. If you're not—well, try Camels now.
CAMEL
Your "T-Zone" Will Tell You...
T for Taste...
T for Throat...
that's your proving ground for any cigarette. See if Camels don't suit your "T-Zone" to a "T."
CAMELS *Costlier Tobaccos*






"The best things in life come in Cellophane"
"You SEE what you buy-no guesswork."
"Foods come fresh-stay fresh longer-less waste."
"...and things in Cellophane are clean and sanitary."
DU PONT Cellophane
DU PONT
BETTER THINGS FOR BETTER LIVING ... THROUGH CHEMISTRY
Look at "Cavalcade of America" on Television

New ads are annoying…for now.



7online.com EYEWITNESS NEWS WABC-TV WABC-DT New York ▶ Tri-State News

HOME | EYEWITNESS NEWS | ACCU-WEATHER | TRAFFIC | BUSINESS | HEALTH | SPORT

TEXT SIZE: A | A | A | A

Teacher hit by car outside Queens school

Parents say it is ongoing problem

Eyewitness

(Queens - WABC, December 15, 2006) - Crisis counselors are on hand at p.s. 200 in the F… …rtion of Queens, tonight.

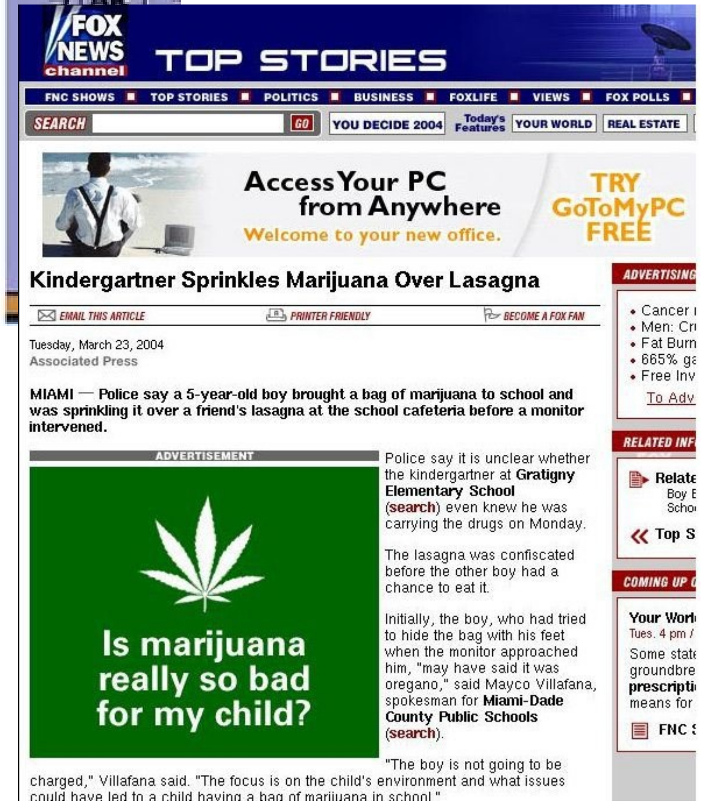
Late this mor… …hit by a car outside the school on 164th Street and Jewe… …ngerous area.

Eyewitne… …l Lipper has more from Kew Gardens.

164th Street an… …e is a very busy section of Queens. For quite some time, parents h… …laining. The cars move way too fast in the area and

Get a new GS6.
Save $300 on a new Tab S2.
How smartphone purchase req'd. Tablet requires new 2yr agmt.
verizon✓

Get $300 off a new Samsung Galaxy Tab S2
www.verizonwireless.com
When you purchase a Samsung GS6 or GS6 Edge



FOX NEWS channel TOP STORIES

FNC SHOWS | TOP STORIES | POLITICS | BUSINESS | FOXLIFE | VIEWS | FOX POLLS

SEARCH GO | YOU DECIDE 2004 | Today's Features | YOUR WORLD | REAL ESTATE

Kindergartner Sprinkles Marijuana Over Lasagna

EMAIL THIS ARTICLE | PRINTER FRIENDLY | BECOME A FOX FAN

Tuesday, March 23, 2004
Associated Press

MIAMI — Police say a 5-year-old boy brought a bag of marijuana to school and was sprinkling it over a friend's lasagna at the school cafeteria before a monitor intervened.

ADVERTISEMENT

Police say it is unclear whether the kindergartner at **Gratigny Elementary School** (search) even knew he was carrying the drugs on Monday.

The lasagna was confiscated before the other boy had a chance to eat it.

Initially, the boy, who had tried to hide the bag with his feet when the monitor approached him, "may have said it was oregano," said Mayco Villafana, spokesman for **Miami-Dade County Public Schools** (search).

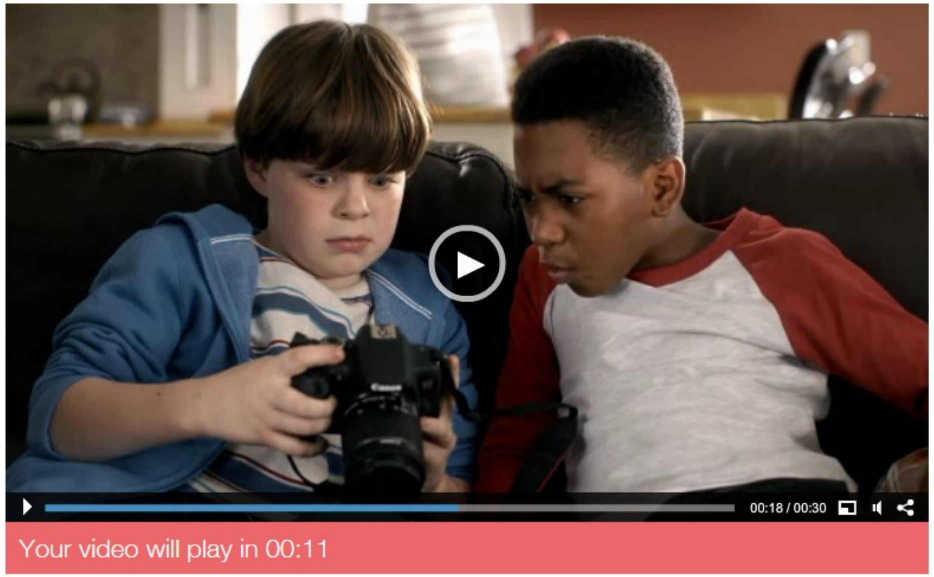
"The boy is not going to be charged," Villafana said. "The focus is on the child's environment and what issues could have led to a child having a bag of marijuana in school."

ADVERTISING
- Cancer
- Men: Cr
- Fat Burn
- 665% ga
- Free Inv

To Adv

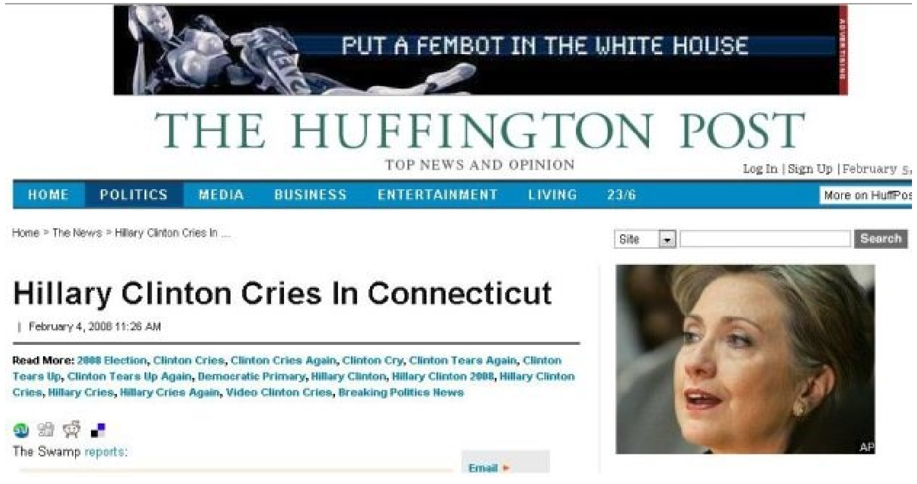
RELATED INF
Relate Boy E Scho
« Top S

COMING UP
Your Worl
Tues. 4 pm /
Some state groundbre **prescripti** means for
FNC



Your video will play in 00:11

"Why are your parents wrestling?"



THE HUFFINGTON POST
TOP NEWS AND OPINION

Log In | Sign Up | February 5,

HOME | POLITICS | MEDIA | BUSINESS | ENTERTAINMENT | LIVING | 23/6 | More on HuffPost

Home > The News > Hillary Clinton Cries In ...

Site ▾ | Search

Hillary Clinton Cries In Connecticut

| February 4, 2008 11:26 AM

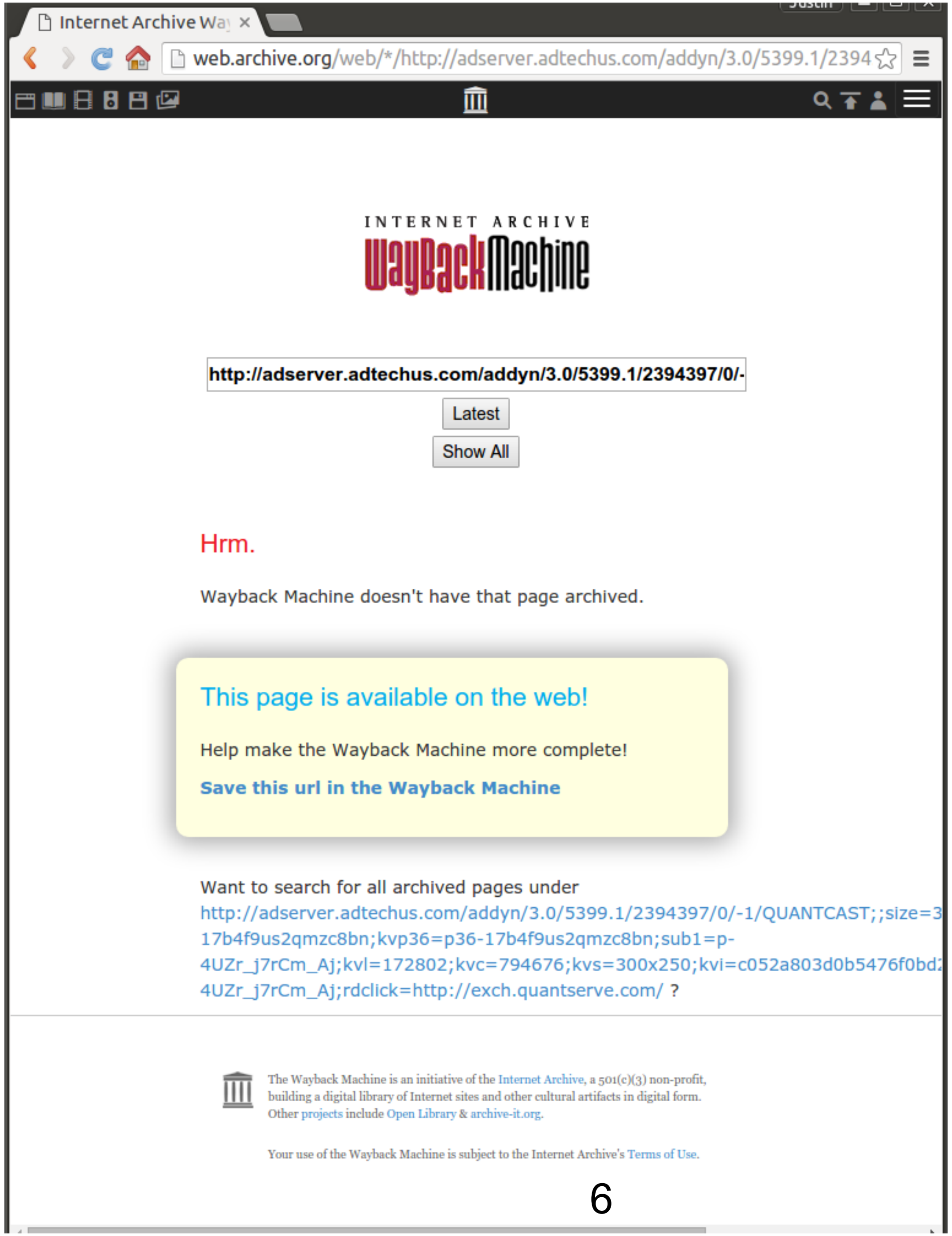
Read More: 2008 Election, Clinton Cries, Clinton Cries Again, Clinton Cry, Clinton Tears Again, Clinton Tears Up, Clinton Tears Up Again, Democratic Primary, Hillary Clinton, Hillary Clinton 2008, Hillary Clinton Cries, Hillary Cries, Hillary Cries Again, Video Clinton Cries, Breaking Politics News

The Swamp reports:

Email ▸

Today’s ads are missing from the archives

http://adserver.adtechus.com/addyn/3.0/5399.1/2394397/0/-1/QUANTCAST;;size=300x250;target=_blank;alias=p36-17b4f9us2qmzc8bn;kvp36=p36-17b4f9us2qmzc8bn;sub1=p-4UZr_j7rCm_Aj;kvl=172802;kvc=794676;kvs=300x250;kvi=c052a803d0b5476f0bd2f2043ef237e27cd48019;kva=p-4UZr_j7rCm_Aj;rdclick=http://exch.quantserve.com/r?a=p-4UZr_j7rCm_Aj;labels=_qc.clk,_click.adserver.rtb,_click.rand.85854;rtbip=192.184.64.144;rtbdata2=EAQaFUhSQmxvY2tfMjAxNlRheFNlYXNvbiCZiRcogsYKMLTAMDoSaHR0cDovL3d3dy5jbm4uY29tWihUUEhwYlUzM3ZqeFU5LTA1SGZEMk1SXzE0anBVcGU0d0dxTG10STFUdUs2IECAAb_JicoFoAEBqAGhy7YCugEoVFBIcGJVMzN2anhVOS0wNUhmRDJNUl8xNGpwVXBlNHdHcUxtdEkxVMAB3ed3yAGUp7GUqSraAShjMDUyYTgwM2QwYjU0NzZmMGJkMmYyMDQzZWYyMzdlMjdjZDQ4MDE5QHvEWs-6AFkmAK2wQqoAgWoAgawAgi6AgTAuECQwAICyAIA0ALe9baMj4Cos-oB


Internet Archive Wa
web.archive.org/web/*/http://adserver.adtechus.com/addyn/3.0/5399.1/2394
INTERNET ARCHIVE
WaybackMachine
http://adserver.adtechus.com/addyn/3.0/5399.1/2394397/0/-
Latest
Show All

Hrm.

Wayback Machine doesn't have that page archived.

This page is available on the web!

Help make the Wayback Machine more complete!

Save this url in the Wayback Machine

Want to search for all archived pages under http://adserver.adtechus.com/addyn/3.0/5399.1/2394397/0/-1/QUANTCAST;;size=3
17b4f9us2qmzc8bn;kvp36=p36-17b4f9us2qmzc8bn;sub1=p-
4UZr_j7rCm_Aj;kvl=172802;kvc=794676;kvs=300x250;kvi=c052a803d0b5476f0bd
4UZr_j7rCm_Aj;rdclick=http://exch.quantserve.com/ ?

The Wayback Machine is an initiative of the Internet Archive, a 501(c)(3) non-profit, building a digital library of Internet sites and other cultural artifacts in digital form. Other projects include Open Library & archive-it.org.

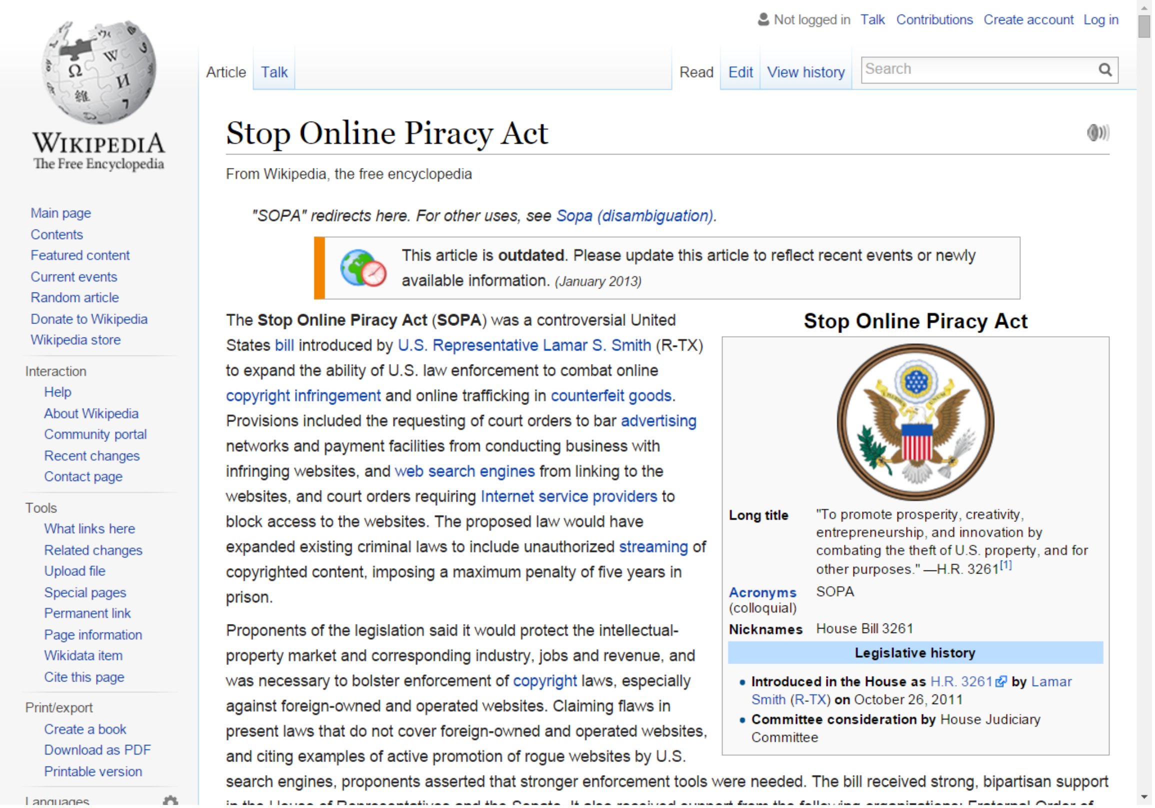
Your use of the Wayback Machine is subject to the Internet Archive's Terms of Use.

JavaScript is hard to replay

What happens when an event is completely lost?

http://ws-dl.blogspot.com/2013/11/2013-11-28-replaying-sopa-protest.html

SOPA: Historically significant, archivally difficult



Stop Online Piracy Act

From Wikipedia, the free encyclopedia

"SOPA" redirects here. For other uses, see Sopa (disambiguation).

This article is **outdated**. Please update this article to reflect recent events or newly available information. *(January 2013)*

Stop Online Piracy Act

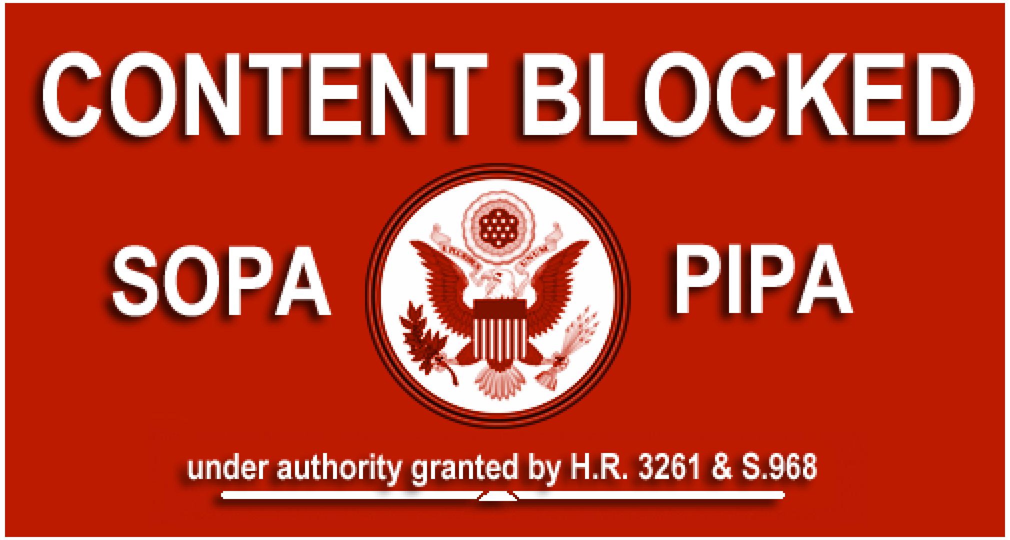
Long title	"To promote prosperity, creativity, entrepreneurship, and innovation by combating the theft of U.S. property, and for other purposes." —H.R. 3261[1]
Acronyms (colloquial)	SOPA
Nicknames	House Bill 3261

Legislative history

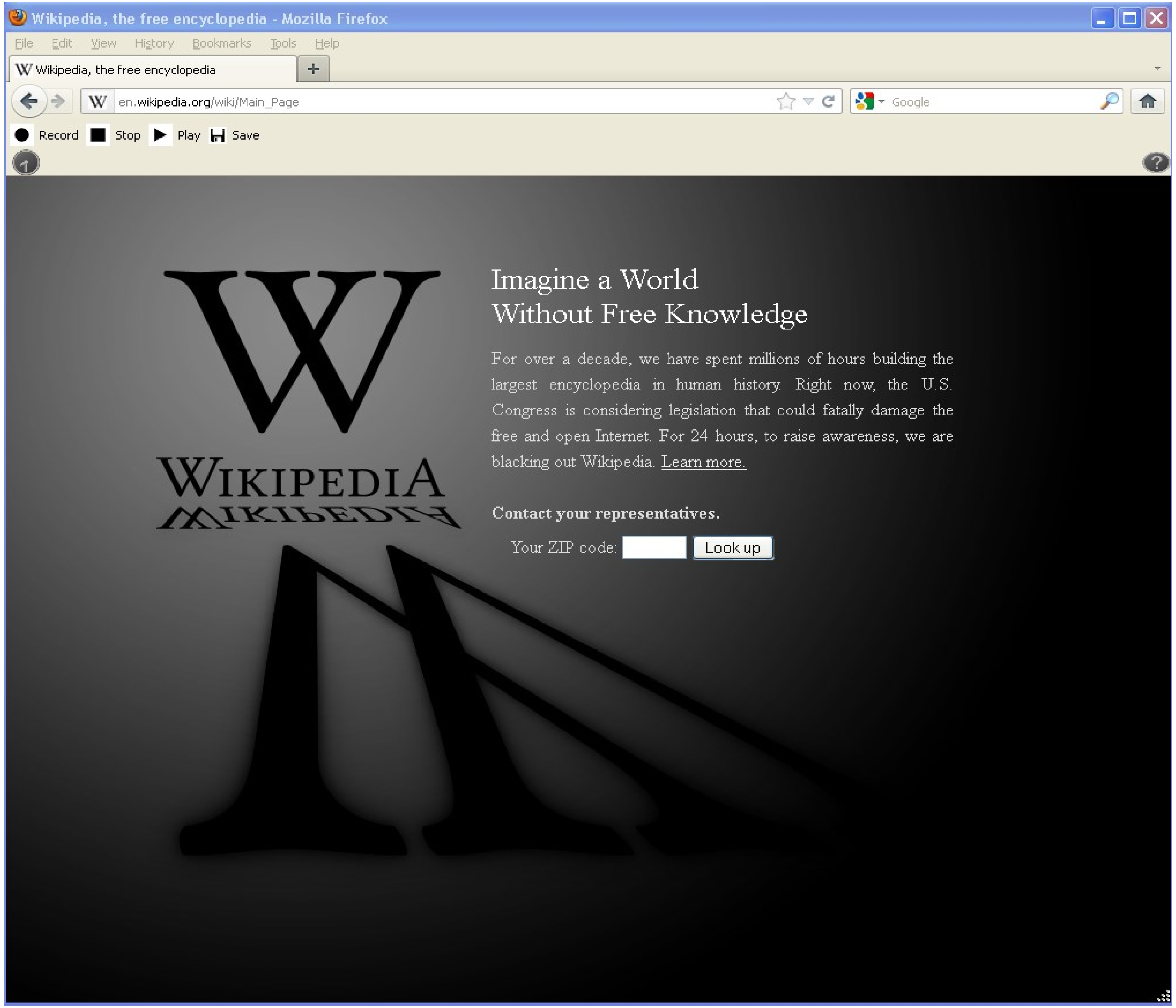
- **Introduced in the House as** H.R. 3261 **by** Lamar Smith (R-TX) **on** October 26, 2011
- **Committee consideration by** House Judiciary Committee

The **Stop Online Piracy Act** (**SOPA**) was a controversial United States bill introduced by U.S. Representative Lamar S. Smith (R-TX) to expand the ability of U.S. law enforcement to combat online copyright infringement and online trafficking in counterfeit goods. Provisions included the requesting of court orders to bar advertising networks and payment facilities from conducting business with infringing websites, and web search engines from linking to the websites, and court orders requiring Internet service providers to block access to the websites. The proposed law would have expanded existing criminal laws to include unauthorized streaming of copyrighted content, imposing a maximum penalty of five years in prison.

Proponents of the legislation said it would protect the intellectual-property market and corresponding industry, jobs and revenue, and was necessary to bolster enforcement of copyright laws, especially against foreign-owned and operated websites. Claiming flaws in present laws that do not cover foreign-owned and operated websites, and citing examples of active promotion of rogue websites by U.S. search engines, proponents asserted that stronger enforcement tools were needed. The bill received strong, bipartisan support in the House of Representatives and the Senate. It also received support from the following organizations: Fraternal Order of



https://en.wikipedia.org/wiki/Stop_Online_Piracy_Act
https://en.wikipedia.org/wiki/Protests_against_SOPA_and_PIPA


Wikipedia, the free encyclopedia - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Main_Page
Google
Record Stop Play Save
WIKIPEDIA

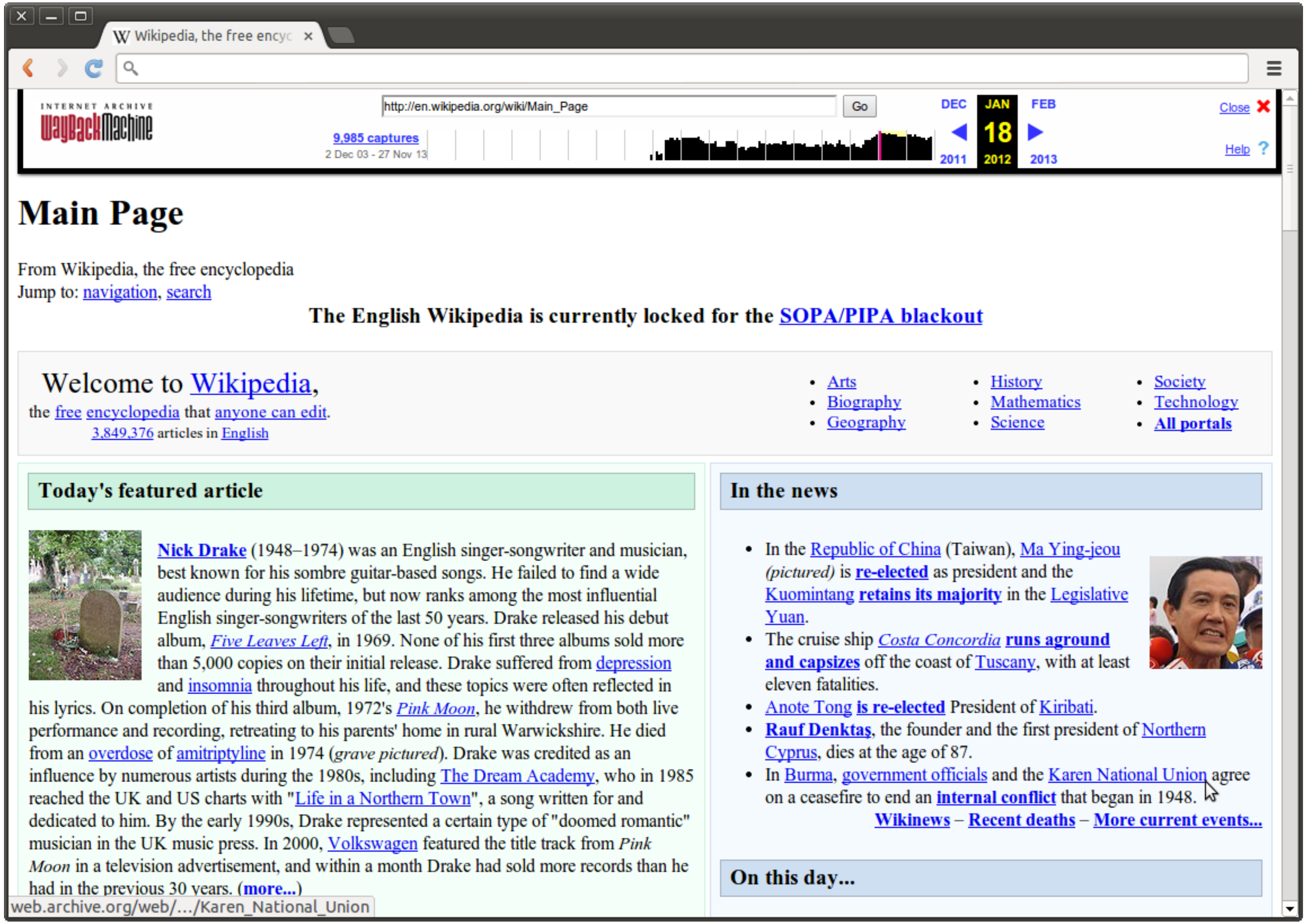
Imagine a World
Without Free Knowledge

For over a decade, we have spent millions of hours building the largest encyclopedia in human history. Right now, the U.S. Congress is considering legislation that could fatally damage the free and open Internet. For 24 hours, to raise awareness, we are blacking out Wikipedia. Learn more.

Contact your representatives.

Your ZIP code: Look up

http://en.wikipedia.org/wiki/Main_Page January 18th, 2012


Wikipedia, the free encyc
INTERNET ARCHIVE
WaybackMachine
http://en.wikipedia.org/wiki/Main_Page
Go
9,985 captures
2 Dec 03 - 27 Nov 13
DEC
JAN
FEB
18
2011
2012
2013
Close
Help
web.archive.org/web/.../Karen_National_Union

Main Page

From Wikipedia, the free encyclopedia
Jump to: navigation, search

The English Wikipedia is currently locked for the SOPA/PIPA blackout

Welcome to Wikipedia,
the free encyclopedia that anyone can edit.
3,849,376 articles in English

- Arts
- Biography
- Geography
- History
- Mathematics
- Science
- Society
- Technology
- **All portals**

Today's featured article

Nick Drake (1948–1974) was an English singer-songwriter and musician, best known for his sombre guitar-based songs. He failed to find a wide audience during his lifetime, but now ranks among the most influential English singer-songwriters of the last 50 years. Drake released his debut album, *Five Leaves Left*, in 1969. None of his first three albums sold more than 5,000 copies on their initial release. Drake suffered from depression and insomnia throughout his life, and these topics were often reflected in his lyrics. On completion of his third album, 1972's *Pink Moon*, he withdrew from both live performance and recording, retreating to his parents' home in rural Warwickshire. He died from an overdose of amitriptyline in 1974 (*grave pictured*). Drake was credited as an influence by numerous artists during the 1980s, including The Dream Academy, who in 1985 reached the UK and US charts with "Life in a Northern Town", a song written for and dedicated to him. By the early 1990s, Drake represented a certain type of "doomed romantic" musician in the UK music press. In 2000, Volkswagen featured the title track from *Pink Moon* in a television advertisement, and within a month Drake had sold more records than he had in the previous 30 years. (**more...**)

In the news

- In the Republic of China (Taiwan), Ma Ying-jeou *(pictured)* is **re-elected** as president and the Kuomintang **retains its majority** in the Legislative Yuan.
- The cruise ship *Costa Concordia* **runs aground and capsizes** off the coast of Tuscany, with at least eleven fatalities.
- Anote Tong **is re-elected** President of Kiribati.
- **Rauf Denktaş**, the founder and the first president of Northern Cyprus, dies at the age of 87.
- In Burma, government officials and the Karen National Union agree on a ceasefire to end an **internal conflict** that began in 1948.

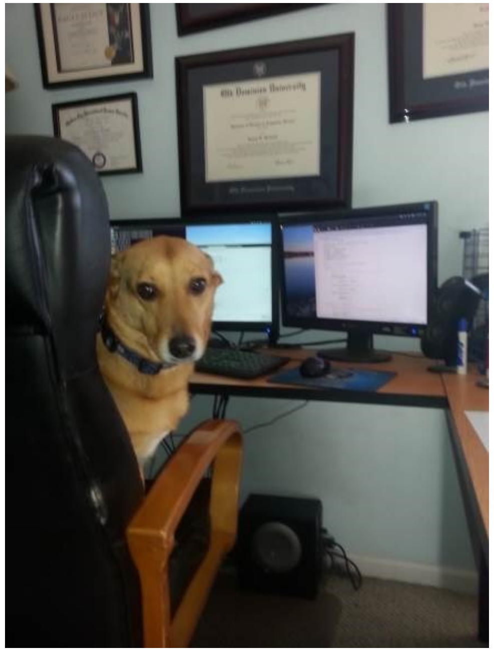
Wikinews – **Recent deaths** – **More current events...**

On this day...

http://web.archive.org/web/20120118110520/http://en.wikipedia.org/wiki/Main_Page January 18th, 2012

Problem!

HERITRIX ≠



The archives contain the Web as seen by crawlers

Why archive?

The Internet Archive has everything!

Why didn’t you back it up?

Participating institutions can hand over their databases.

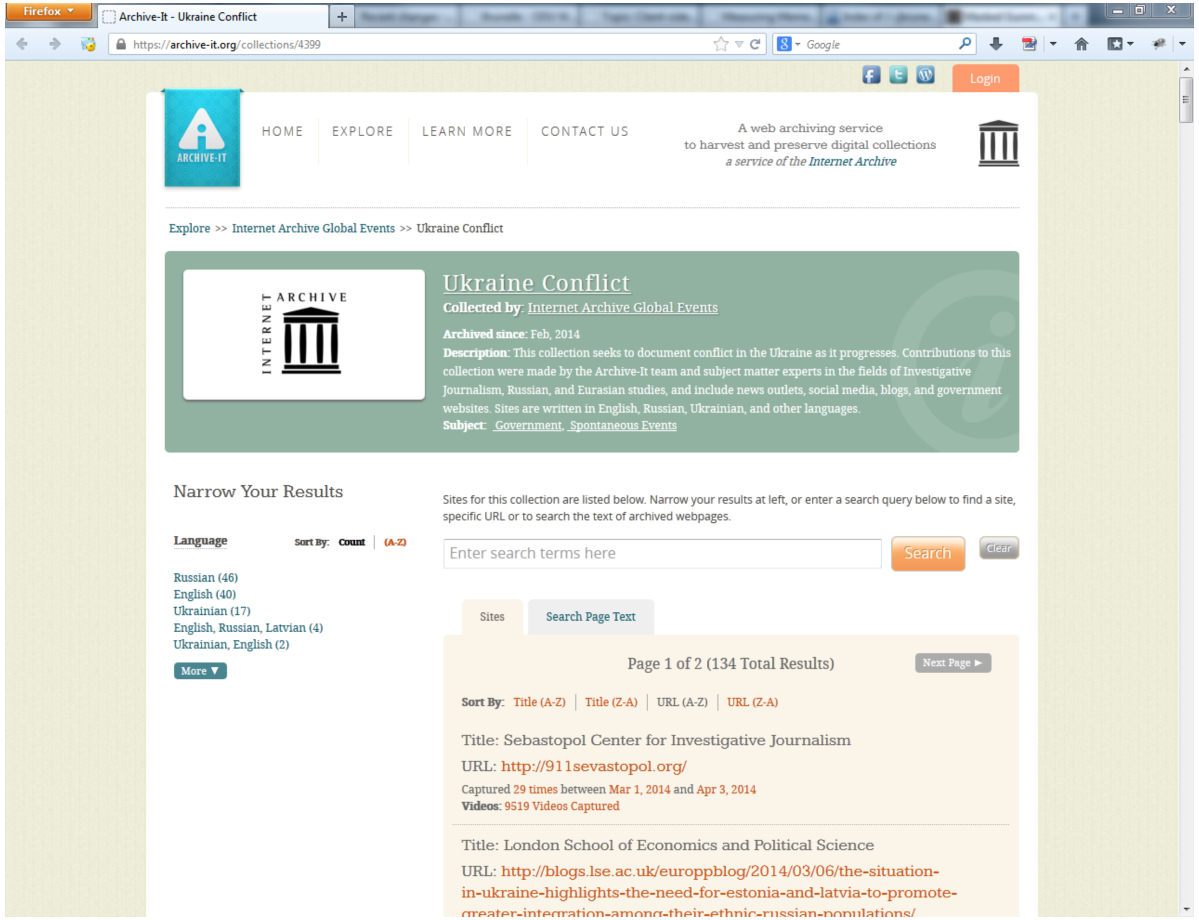
Crimean Conflict

Russian troops captured the Crimean Center for Investigative Journalism

Gunman: "We will try to agree on the correct truthful coverage of events."

http://gijn.org/2014/03/02/masked-gunmen-seize-crimean-investigative-journalism-center/

Archive-It to the rescue!



A web archiving service
to harvest and preserve digital collections
a service of the Internet Archive

Explore >> Internet Archive Global Events >> Ukraine Conflict

Ukraine Conflict

Collected by: Internet Archive Global Events

Archived since: Feb, 2014

Description: This collection seeks to document conflict in the Ukraine as it progresses. Contributions to this collection were made by the Archive-It team and subject matter experts in the fields of Investigative Journalism, Russian, and Eurasian studies, and include news outlets, social media, blogs, and government websites. Sites are written in English, Russian, Ukrainian, and other languages.

Subject: Government, Spontaneous Events

Narrow Your Results

Language Sort By: **Count** | (A-Z)

Russian (46)
English (40)
Ukrainian (17)
English, Russian, Latvian (4)
Ukrainian, English (2)

Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.

Enter search terms here

Page 1 of 2 (134 Total Results)

Sort By: Title (A-Z) | Title (Z-A) | URL (A-Z) | URL (Z-A)

Title: Sebastopol Center for Investigative Journalism

URL: http://911sevastopol.org/

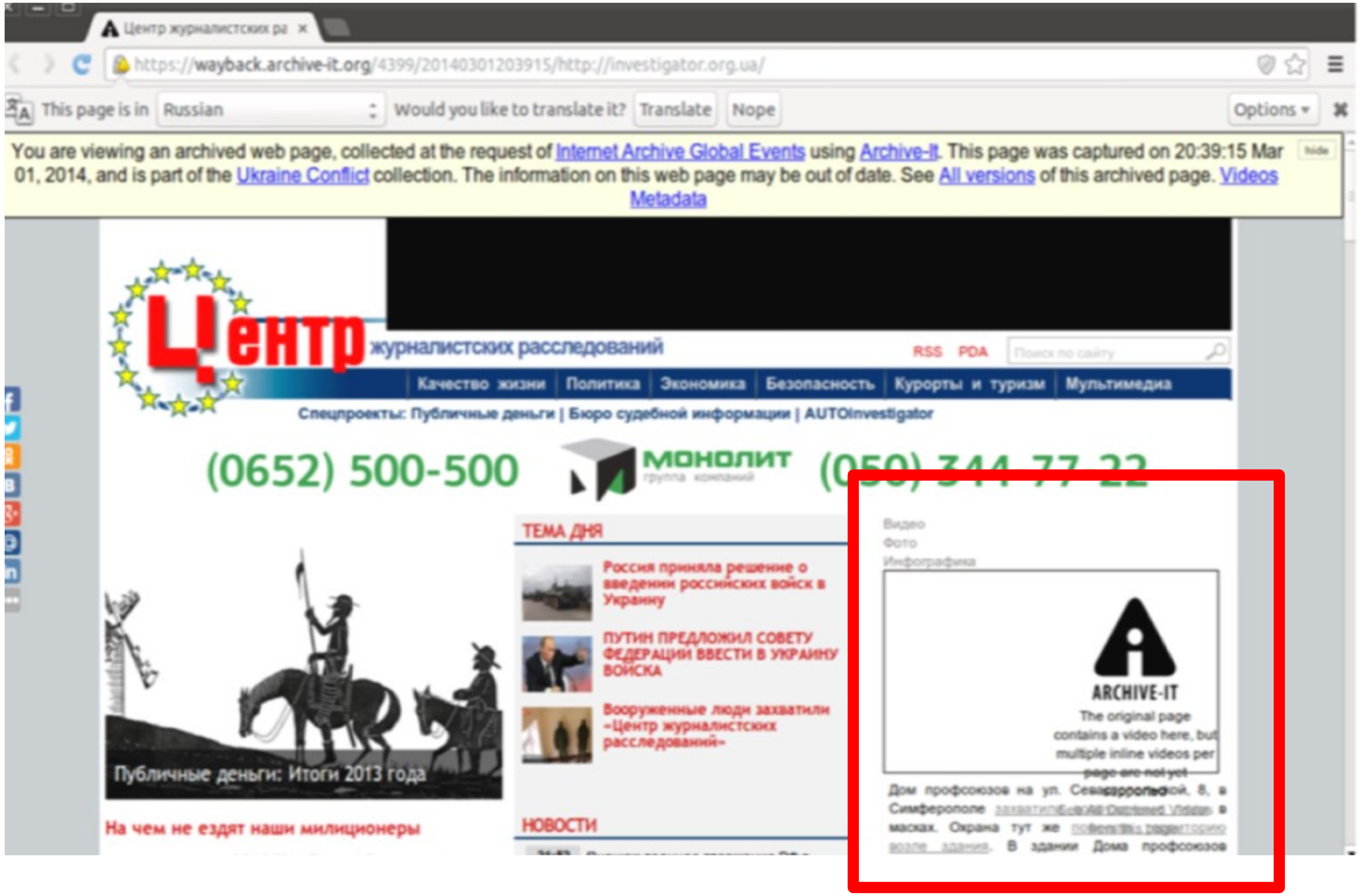
Captured 29 times between Mar 1, 2014 and Apr 3, 2014
Videos: 9519 Videos Captured

Title: London School of Economics and Political Science

URL: http://blogs.lse.ac.uk/europpblog/2014/03/06/the-situation-in-ukraine-highlights-the-need-for-estonia-and-latvia-to-promote-greater-integration-among-their-ethnic-russian-populations/

How well is it archived?

- Masked gunman have your servers
  - anything onsite is gone or altered


Центр журналистских ра
https://wayback.archive-it.org/4399/20140301203915/http://investigator.org.ua/
This page is in
Russian
Would you like to translate it?
Translate
Nope
Options
hide

You are viewing an archived web page, collected at the request of Internet Archive Global Events using Archive-It. This page was captured on 20:39:15 Mar 01, 2014, and is part of the Ukraine Conflict collection. The information on this web page may be out of date. See All versions of this archived page. Videos Metadata

Россия приняла решение о введении российских войск в Украину

ПУТИН ПРЕДЛОЖИЛ СОВЕТУ ФЕДЕРАЦИИ ВВЕСТИ В УКРАИНУ ВОЙСКА

Вооруженные люди захватили «Центр журналистских расследований»

На чем не ездят наши милиционеры

The original page contains a video here, but multiple inline videos per page are not yet

Дом профсоюзов на ул. Севастопольской, 8, в Симферополе захватили люди в масках. Охрана тут же покинула территорию возле здания. В здании Дома профсоюзов

Threat models: http://blog.dshr.org/2011/01/threats-to-preservation.html
Automating assessment of crawl quality: http://www.cs.odu.edu/~mln/pubs/jcdl-2014/jcdl-2014-brunelle-damage.pdf

Any future discussion of the 21st century will involve the web and the web archives

Any future discussion of the 21st century will involve the web and the web archives

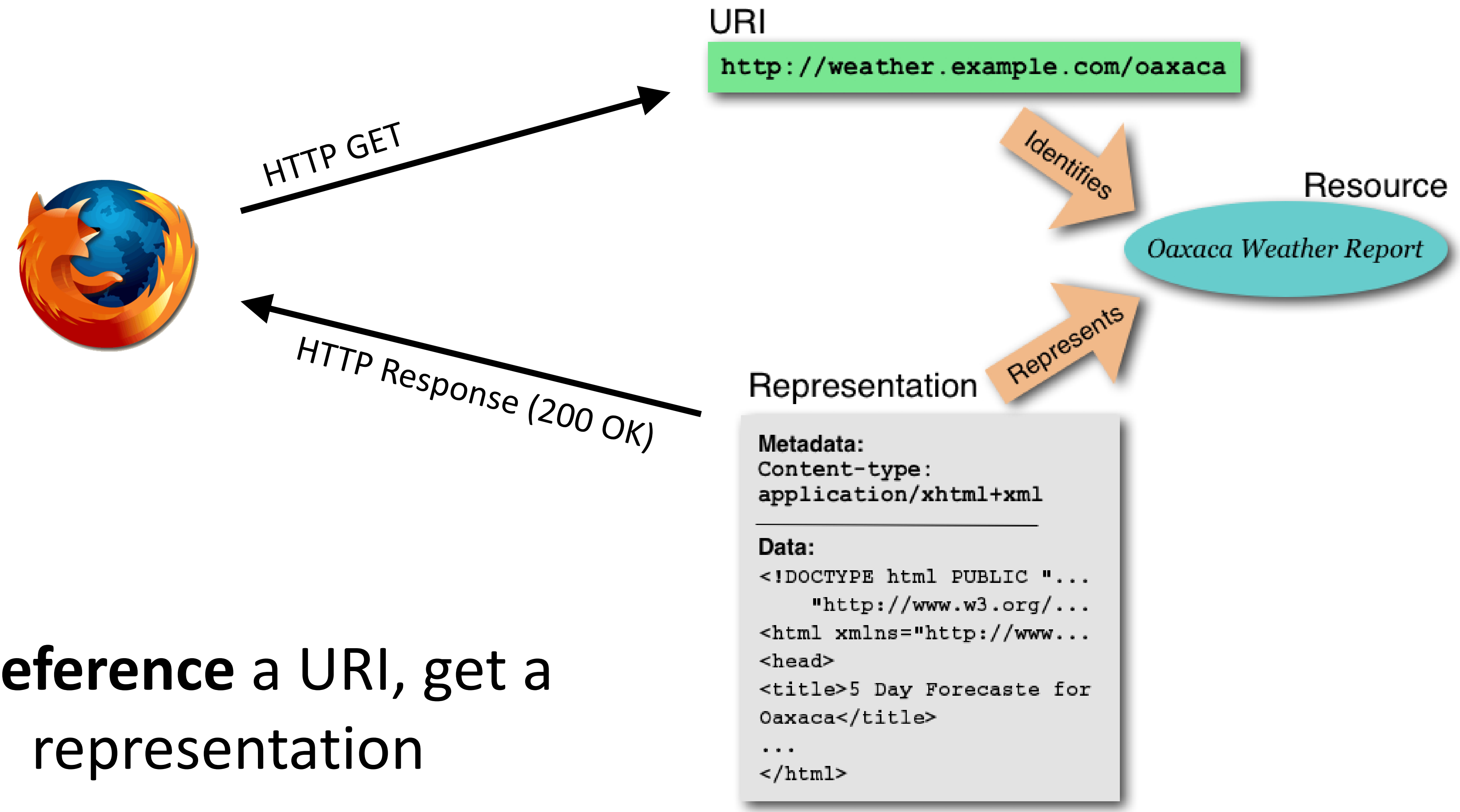
But JavaScript is hard to archive, resulting in archives of content as seen by crawlers rather than as seen by users

Any future discussion of the 21st century will involve the web and the web archives

But JavaScript is hard to archive, resulting in archives of content as seen by crawlers rather than as seen by users

Goal: Mitigate the impact of JavaScript on the archives by making crawlers behave like users

W3C Web Architecture

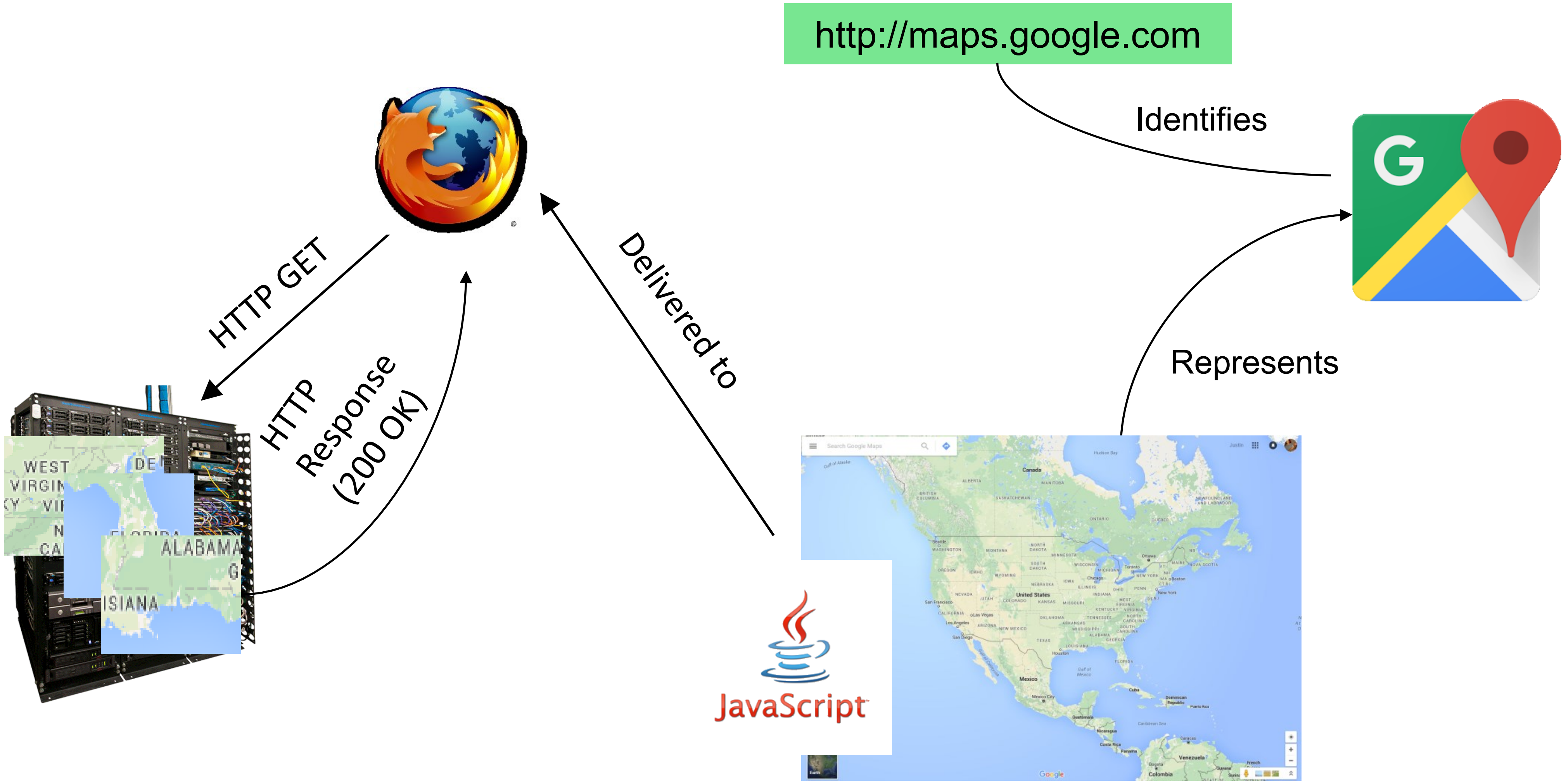


```
Metadata:
Content-type:
application/xhtml+xml

Data:
<!DOCTYPE html PUBLIC "...
    "http://www.w3.org/...
<html xmlns="http://www...
<head>
<title>5 Day Forecaste for
Oaxaca</title>
...
</html>
```

Dereference a URI, get a representation

JavaScript Impact on the Web Architecture



http://maps.google.com

Identifies

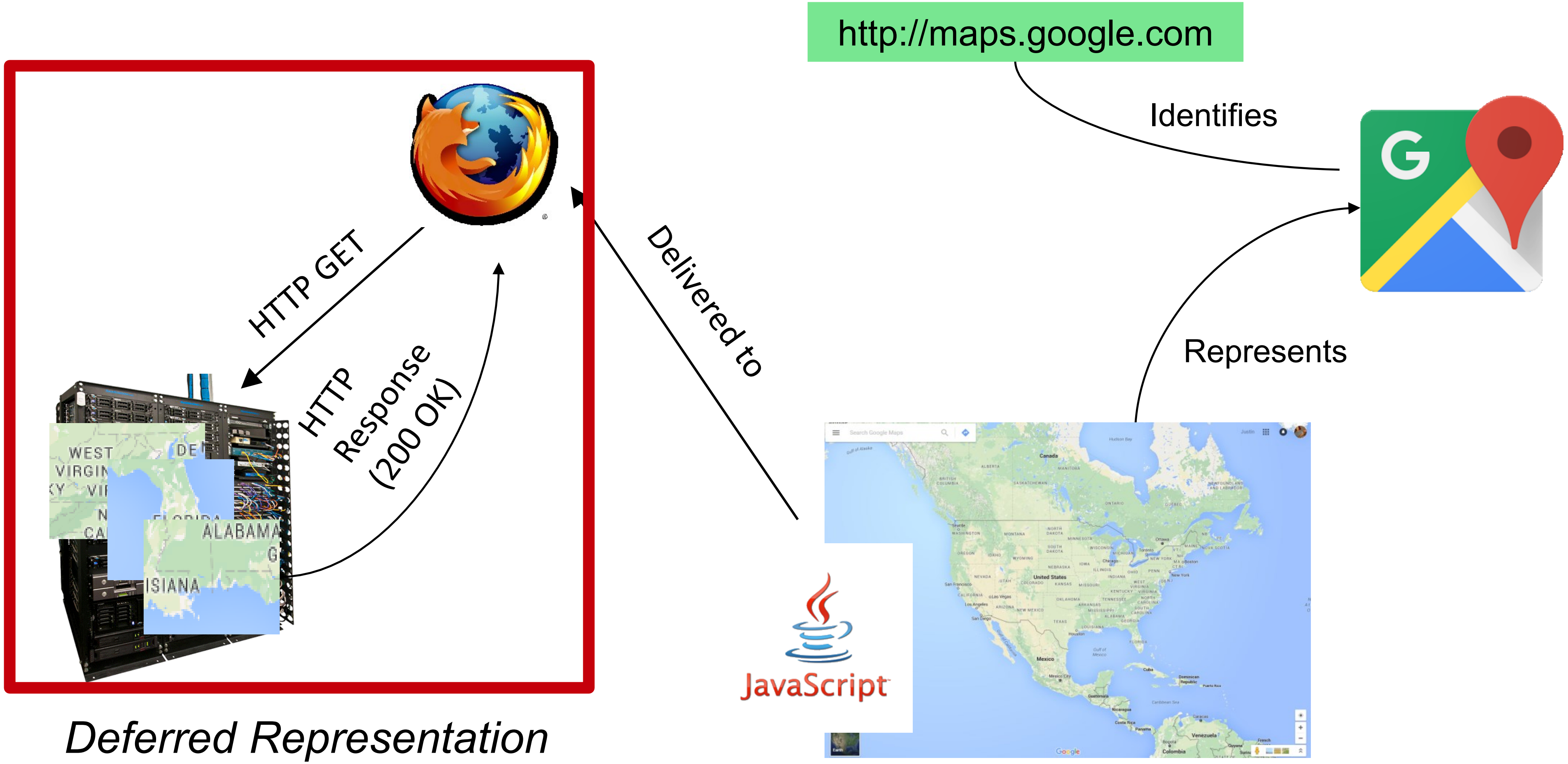
Represents

Delivered to

HTTP GET

HTTP Response (200 OK)

JavaScript makes requests for new resources after the initial page load



http://maps.google.com

Identifies

Represents

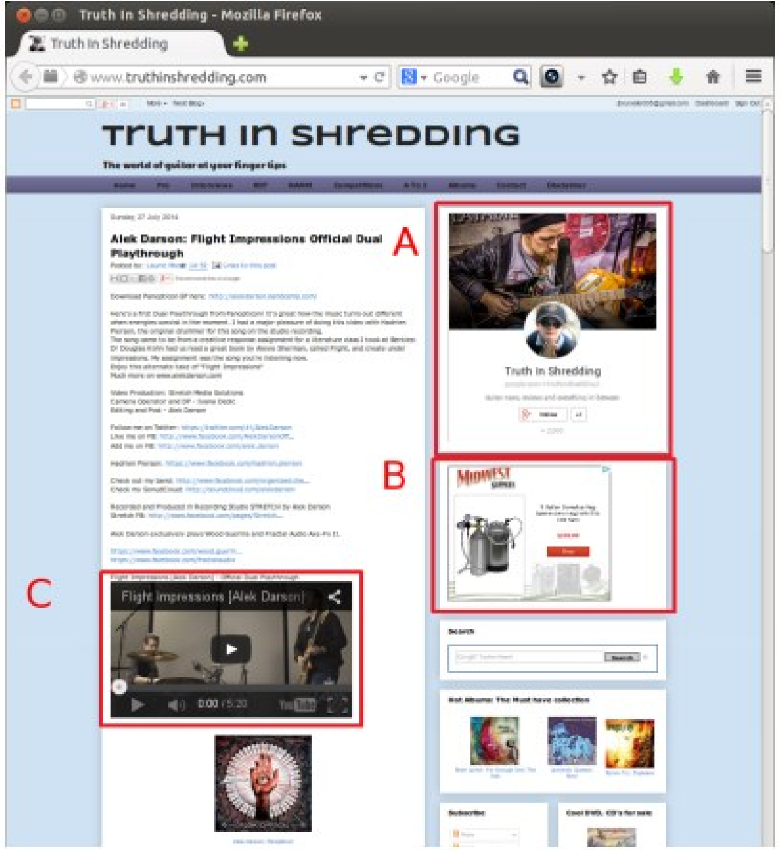
Delivered to

HTTP GET

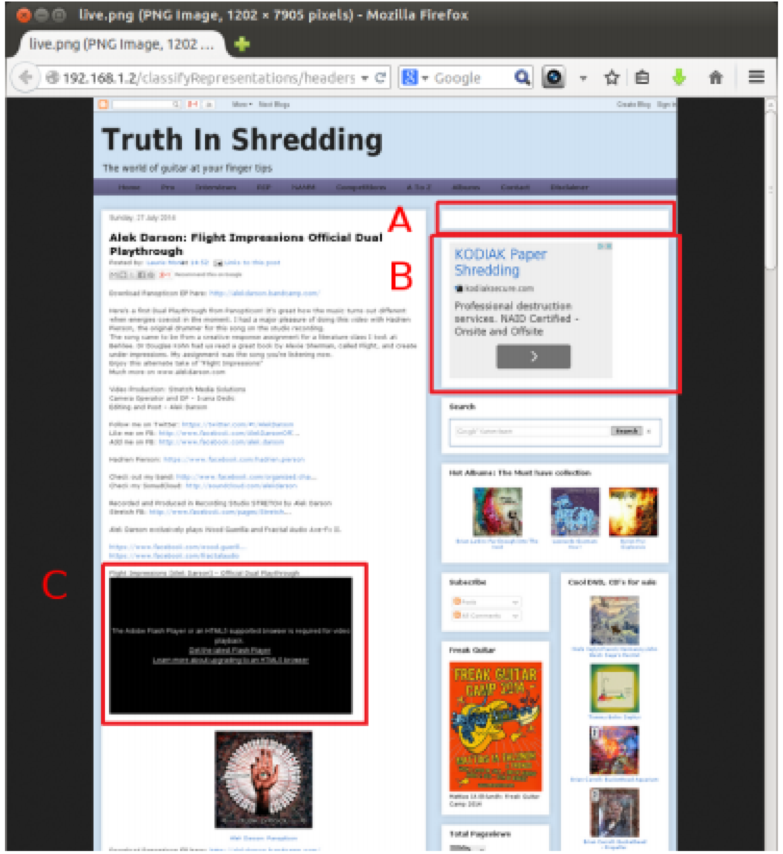
HTTP Response (200 OK)

Deferred Representation

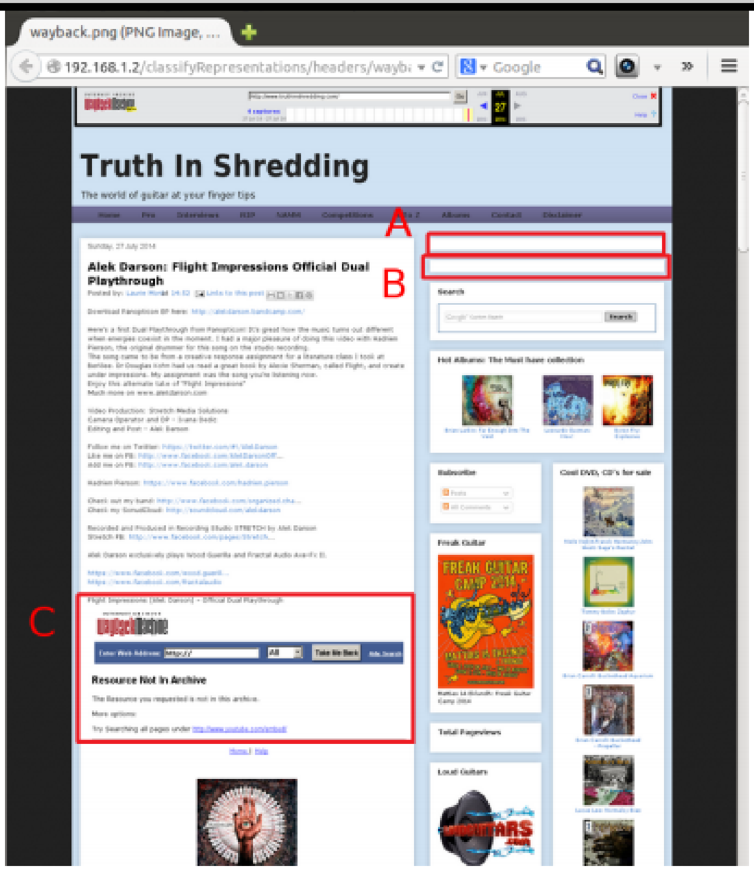
Live: JavaScript PhantomJS: JavaScript Heritrix: *No* JavaScript


Truth In Shredding - Mozilla Firefox
Truth In Shredding
www.truthinshredding.com
Truth In Shredding
The world of guitar at your finger tips
Alek Darson: Flight Impressions Official Dual Playthrough
A
B
C
Flight Impressions [Alek Darson]
Truth In Shredding
MIDWEST
0:00 / 5:20

Live Resource

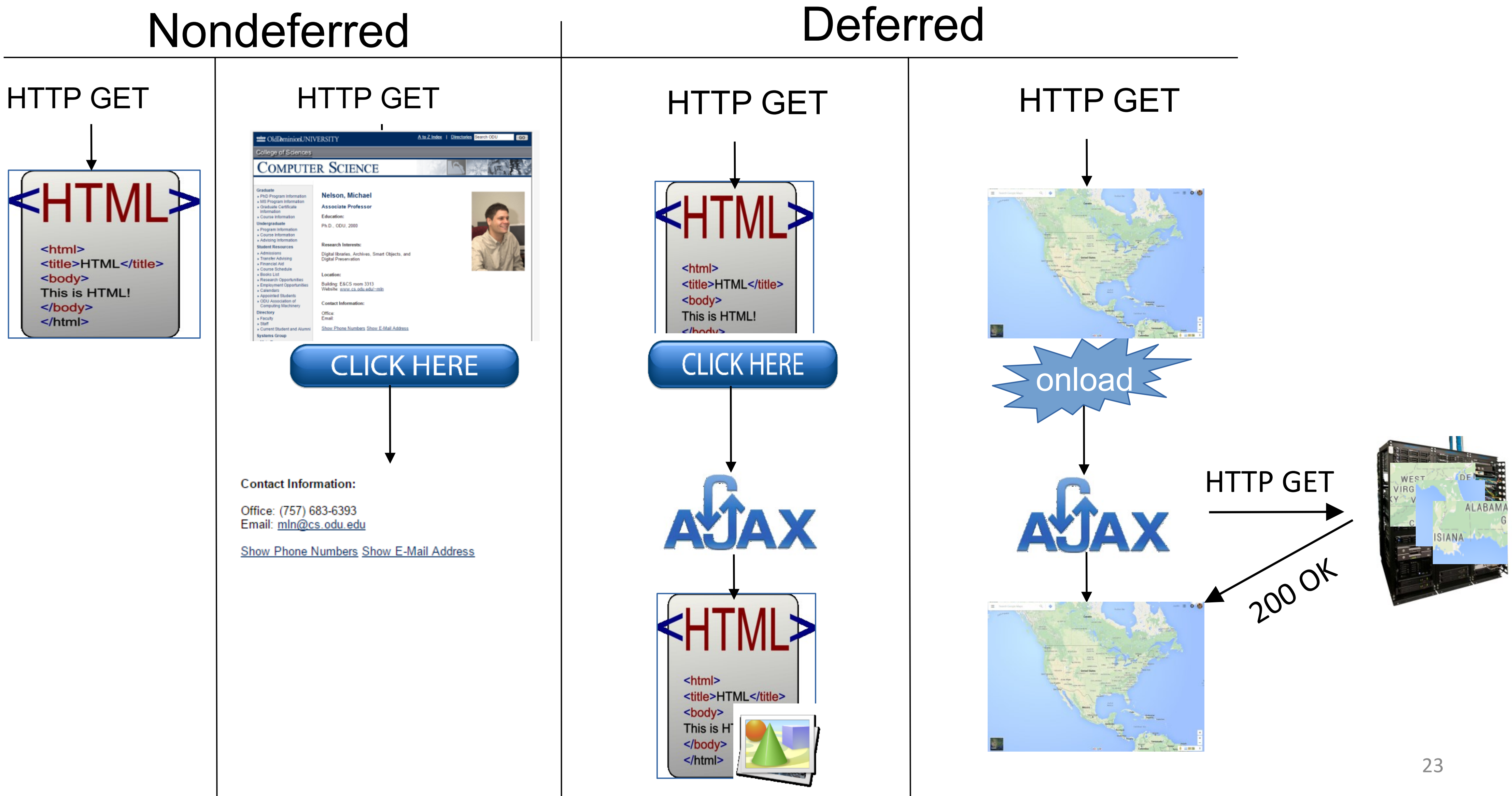

live.png (PNG Image, 1202 × 7905 pixels) - Mozilla Firefox
live.png (PNG Image, 1202 ...
192.168.1.2/classifyRepresentations/headers
Truth In Shredding
The world of guitar at your finger tips
Alek Darson: Flight Impressions Official Dual Playthrough
A
B
C
KODIAK Paper Shredding
kodiaksecure.com
Professional destruction services. NAID Certified - Onsite and Offsite
FREAK GUITAR CAMP 2014

PhantomJS Crawled


wayback.png (PNG Image, ...
192.168.1.2/classifyRepresentations/headers/wayb
Truth In Shredding
The world of guitar at your finger tips
Alek Darson: Flight Impressions Official Dual Playthrough
A
B
C
Resource Not In Archive
FREAK GUITAR CAMP 2014

Heritrix Crawled, Wayback replayed

JavaScript != Deferred



OldDominionUNIVERSITY A to Z Index | Directories Search ODU GO

College of Sciences

COMPUTER SCIENCE

Graduate
- PhD Program Information
- MS Program Information
- Graduate Certificate Information
- Course Information

Undergraduate
- Program Information
- Course Information
- Advising Information

Student Resources
- Admissions
- Transfer Advising
- Financial Aid
- Course Schedule
- Books List
- Research Opportunities
- Employment Opportunities
- Calendars
- Appointed Students
- ODU Association of Computing Machinery

Directory
- Faculty
- Staff
- Current Student and Alumni

Systems Group

Nelson, Michael

Associate Professor

Education:

Ph.D., ODU, 2000

Research Interests:

Digital libraries, Archives, Smart Objects, and Digital Preservation

Location:

Building: E&CS room 3313
Website: www.cs.odu.edu/~mln

Contact Information:

Office:
Email:

Show Phone Numbers Show E-Mail Address

Contact Information:

Office: (757) 683-6393
Email: mln@cs.odu.edu

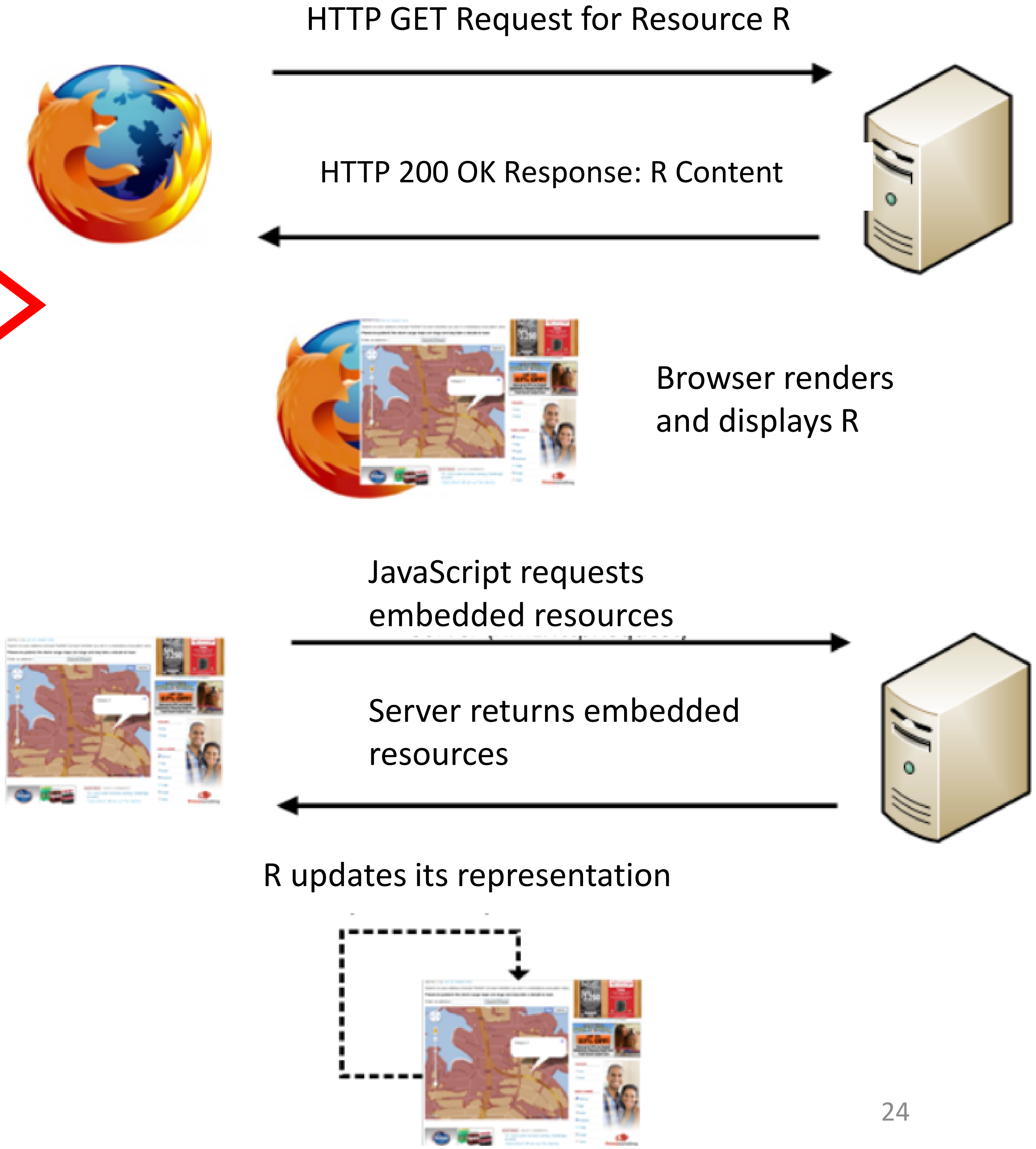
Show Phone Numbers Show E-Mail Address

Web Browsing Process

Archival Tools stop here

WebCite

HERITRIX



HTTP GET Request for Resource R

HTTP 200 OK Response: R Content

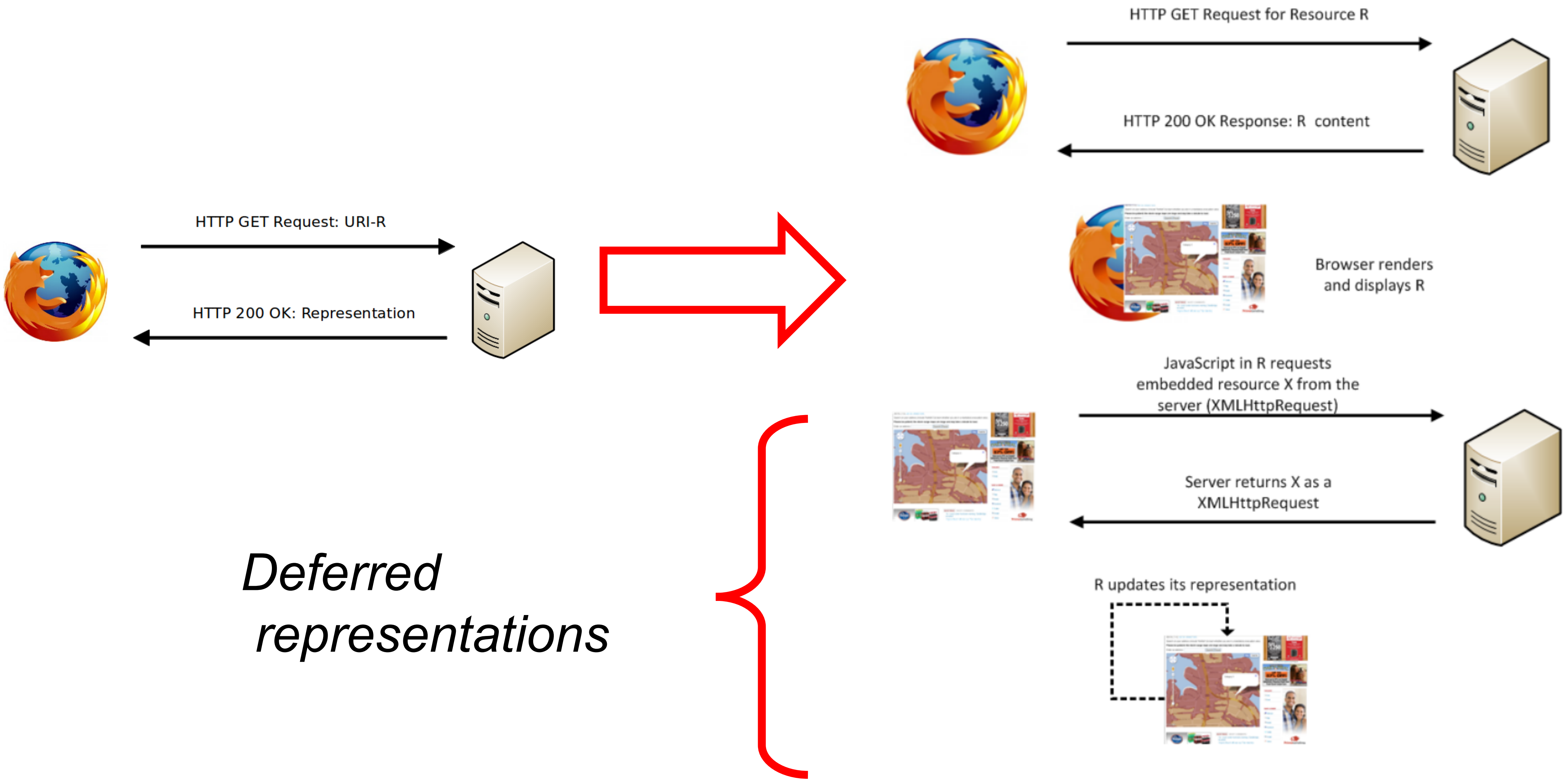
Browser renders and displays R

JavaScript requests embedded resources

Server returns embedded resources

R updates its representation

Web Browsing Process



HTTP GET Request: URI-R

HTTP 200 OK: Representation

HTTP GET Request for Resource R

HTTP 200 OK Response: R content

Browser renders and displays R

JavaScript in R requests embedded resource X from the server (XMLHttpRequest)

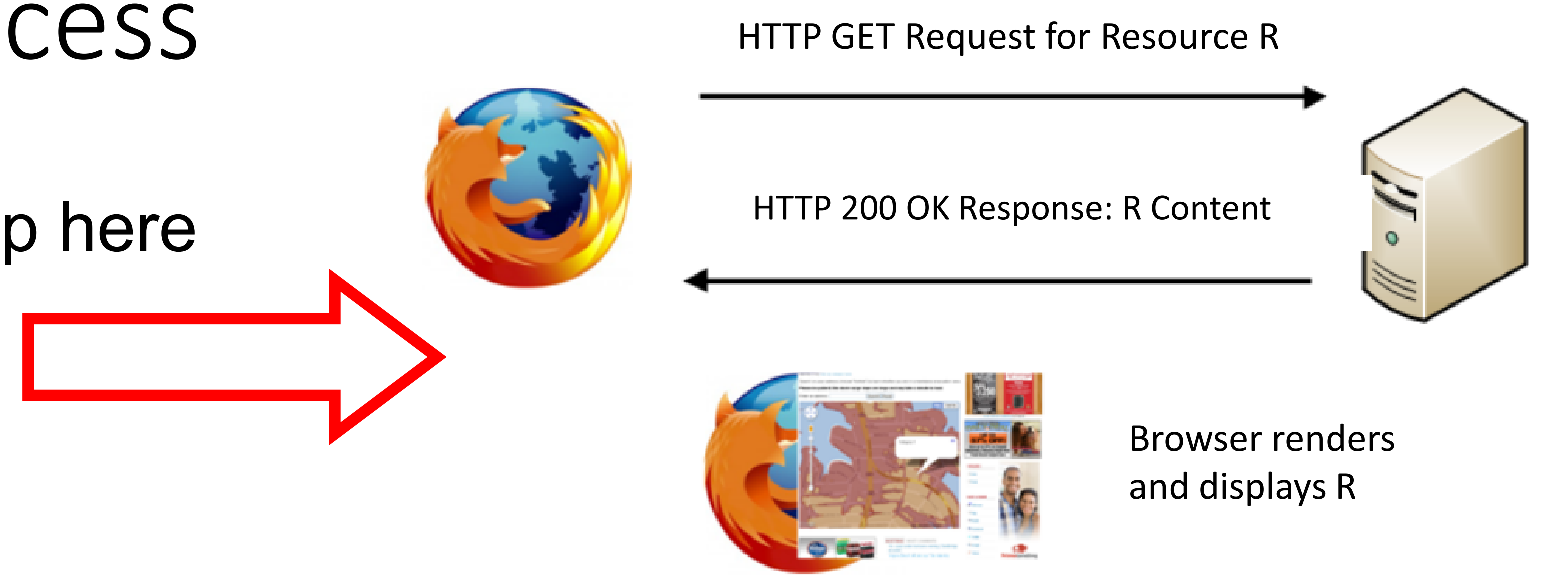
Server returns X as a XMLHttpRequest

R updates its representation

Deferred representations

Web Browsing Process

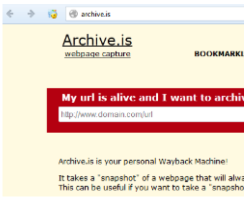
Archival Tools stop here



HTTP GET Request for Resource R

HTTP 200 OK Response: R Content

Browser renders and displays R



archive.is

Archive.is
webpage capture

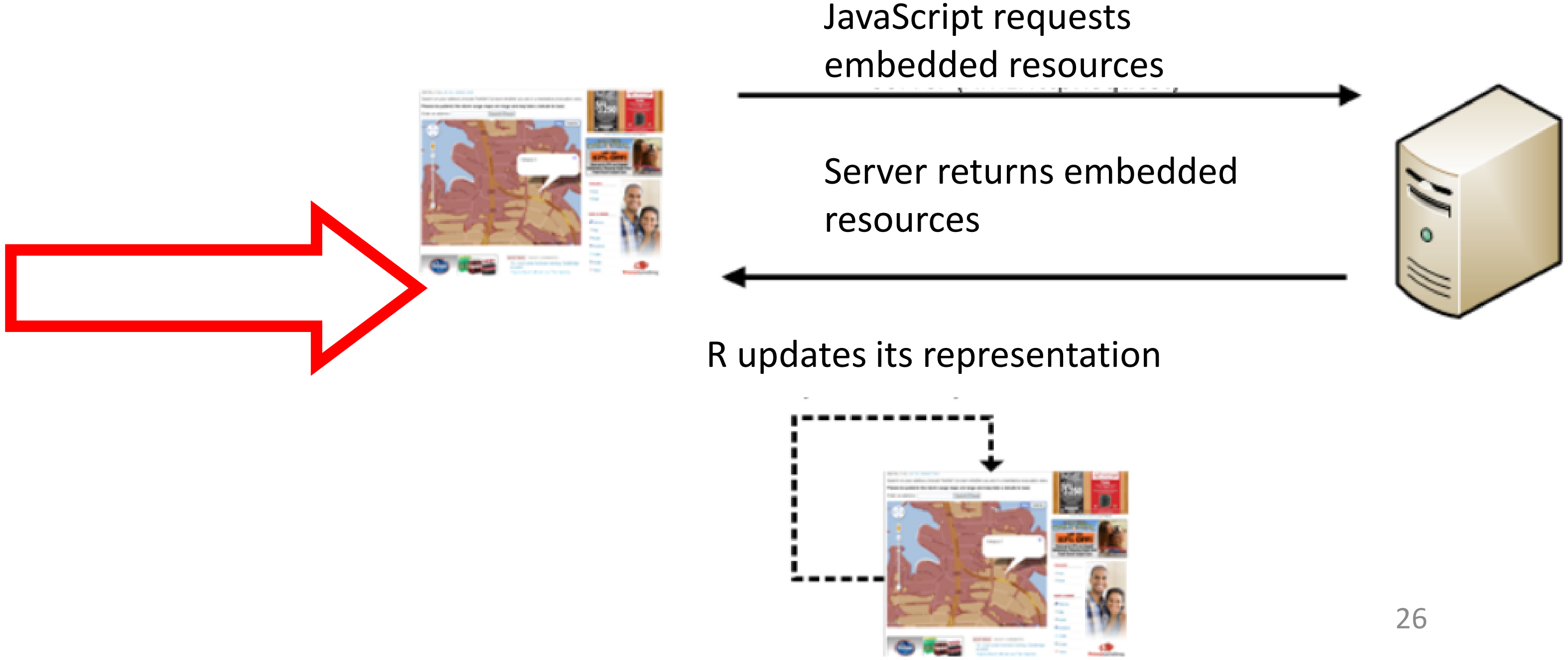
BOOKMARKL

My url is alive and I want to archiv

http://www.domain.com/url

Archive.is is your personal Wayback Machine!

It takes a "snapshot" of a webpage that will alwa
This can be useful if you want to take a "snapsho



JavaScript requests embedded resources

Server returns embedded resources

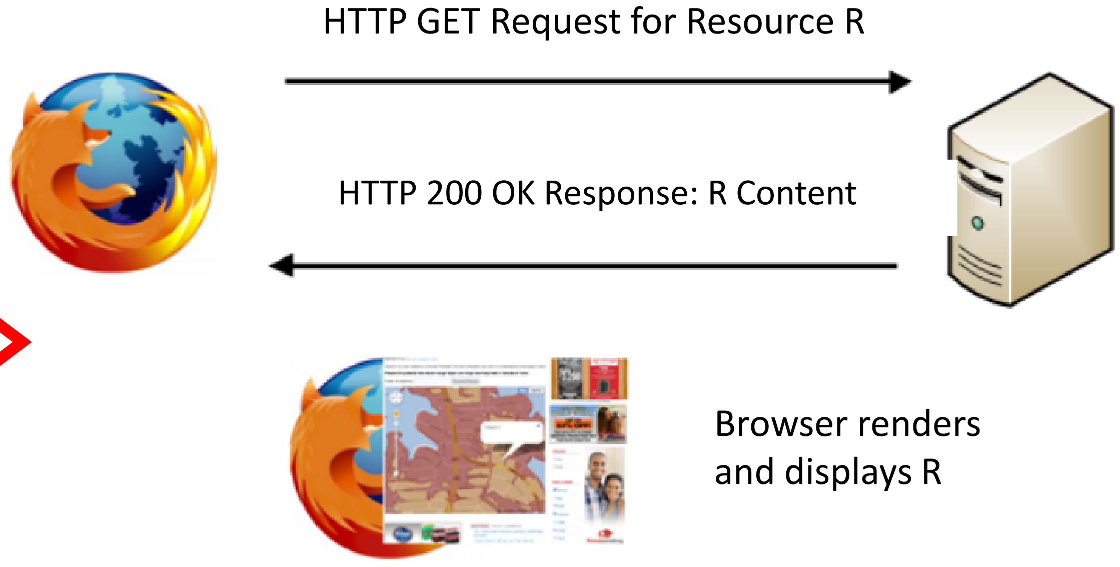
R updates its representation

Web Browsing Process

Archival Tools stop here

WebCite

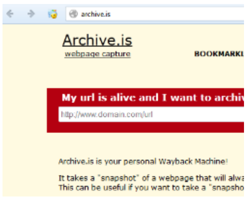
HERITRIX



HTTP GET Request for Resource R

HTTP 200 OK Response: R Content

Browser renders and displays R



archive.is

Archive.is
webpage capture

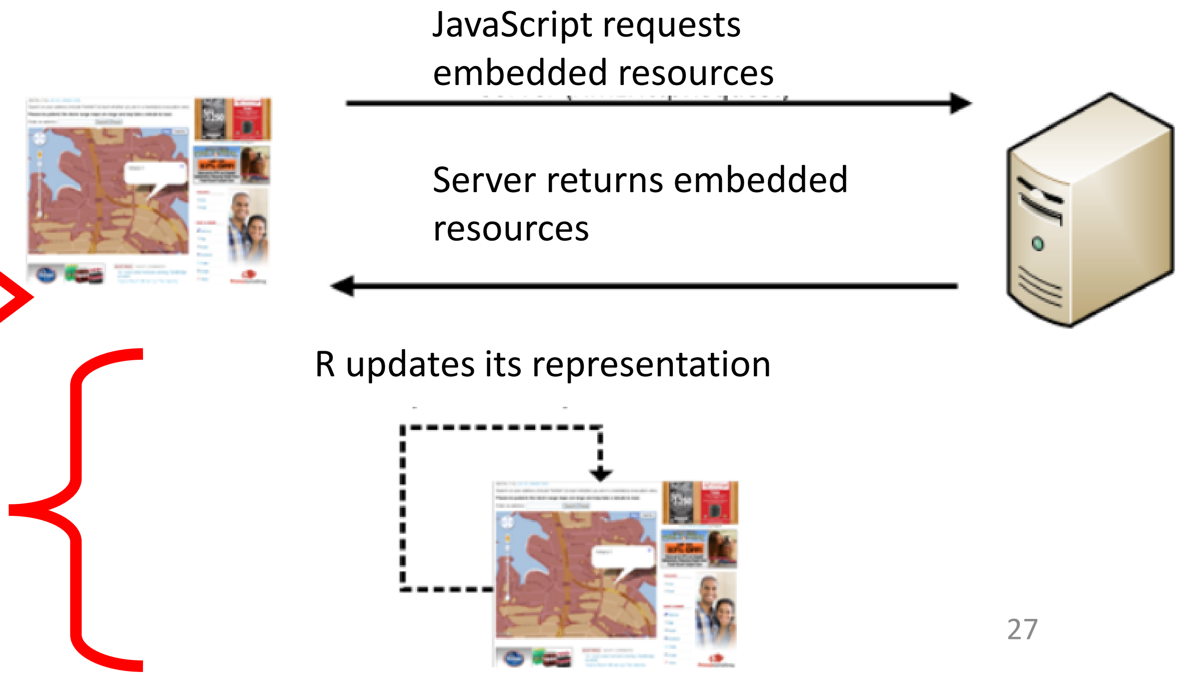
BOOKMARKL

My url is alive and I want to archi

http://www.domain.com/url

Archive.is is your personal Wayback Machine!

It takes a "snapshot" of a webpage that will alwa
This can be useful if you want to take a "snapsho



JavaScript requests embedded resources

Server returns embedded resources

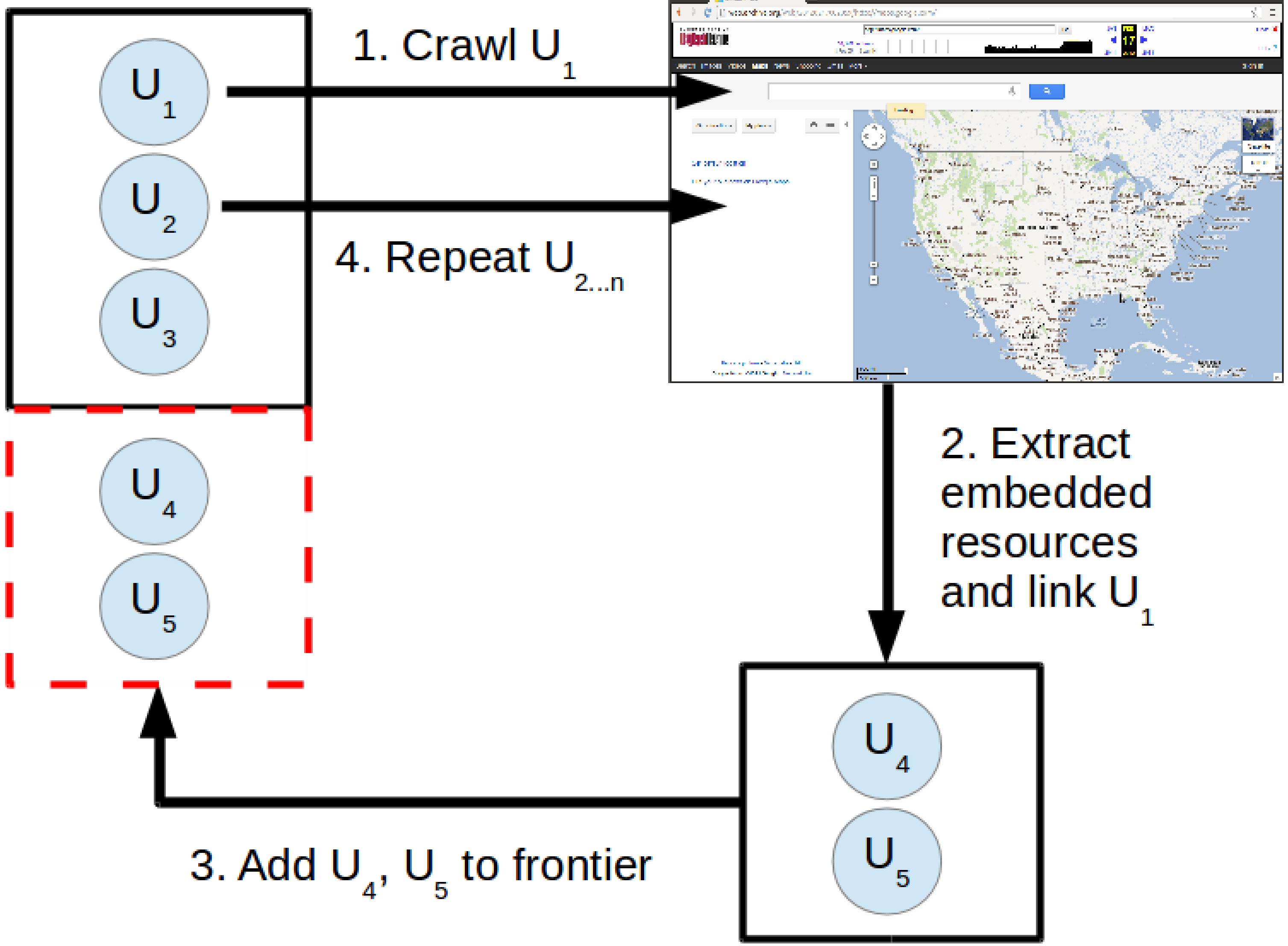
R updates its representation

Archival approach not defined!

Current Workflow

- Dereference URI-Rs
- Archive representation
- Extract embedded
- URI-Rs
- Repeat

Crawl Frontier



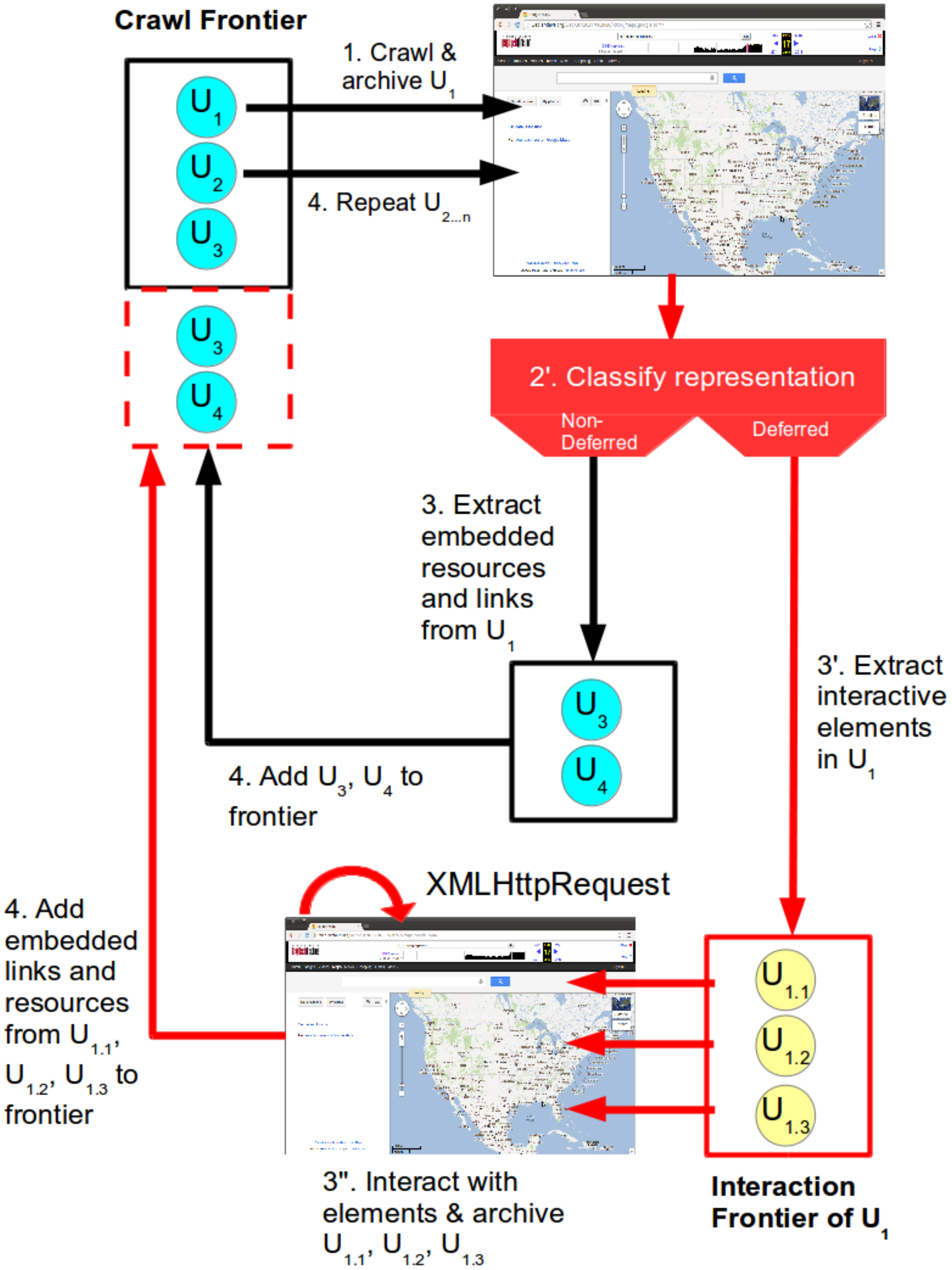
1. Crawl U_1

2. Extract embedded resources and link U_1

3. Add U_4, U_5 to frontier

4. Repeat $U_{2...n}$

Two-Tiered Crawling



Crawl Frontier

1. Crawl & archive U_1

4. Repeat $U_{2...n}$

2'. Classify representation

3. Extract embedded resources and links from U_1

4. Add U_3, U_4 to frontier

3'. Extract interactive elements in U_1

XMLHttpRequest

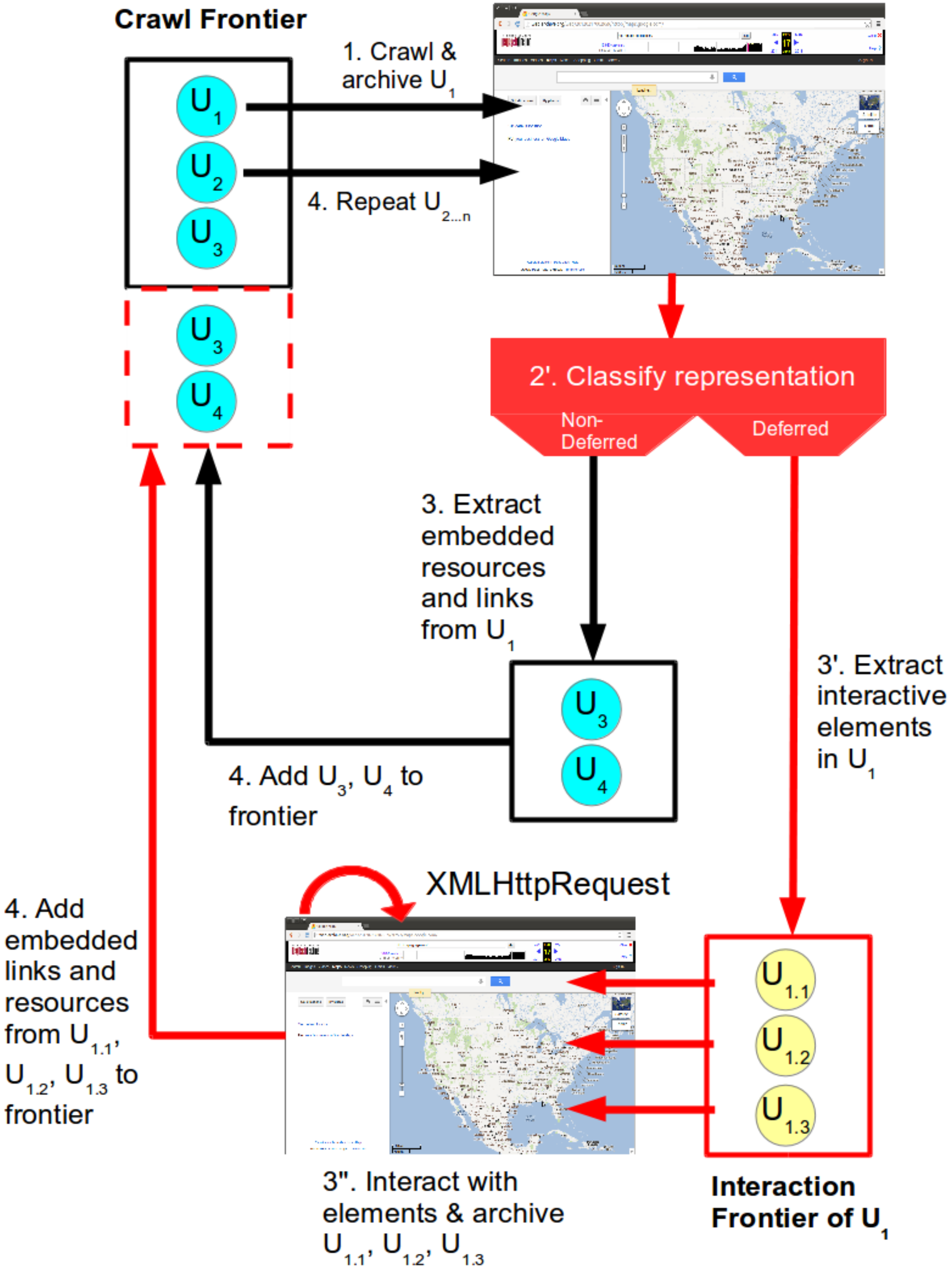
4. Add embedded links and resources from $U_{1.1}$, $U_{1.2}$, $U_{1.3}$ to frontier

3". Interact with elements & archive $U_{1.1}$, $U_{1.2}$, $U_{1.3}$

Interaction Frontier of U_1

"Archiving Deferred Representations Using a Two-Tiered Crawling Approach", *iPRES2015*

"Adapting the Hypercube Model to Archive Deferred Representations at Web-Scale", *Technical Report, arXiv:1601.05142, 2016*


Crawl Frontier

1. Crawl & archive U_1

4. Repeat $U_{2...n}$

2'. Classify representation

Non-Deferred

Deferred

3. Extract embedded resources and links from U_1

4. Add U_3, U_4 to frontier

3'. Extract interactive elements in U_1

XMLHttpRequest

4. Add embedded links and resources from $U_{1.1}$, $U_{1.2}$, $U_{1.3}$ to frontier

3". Interact with elements & archive $U_{1.1}$, $U_{1.2}$, $U_{1.3}$

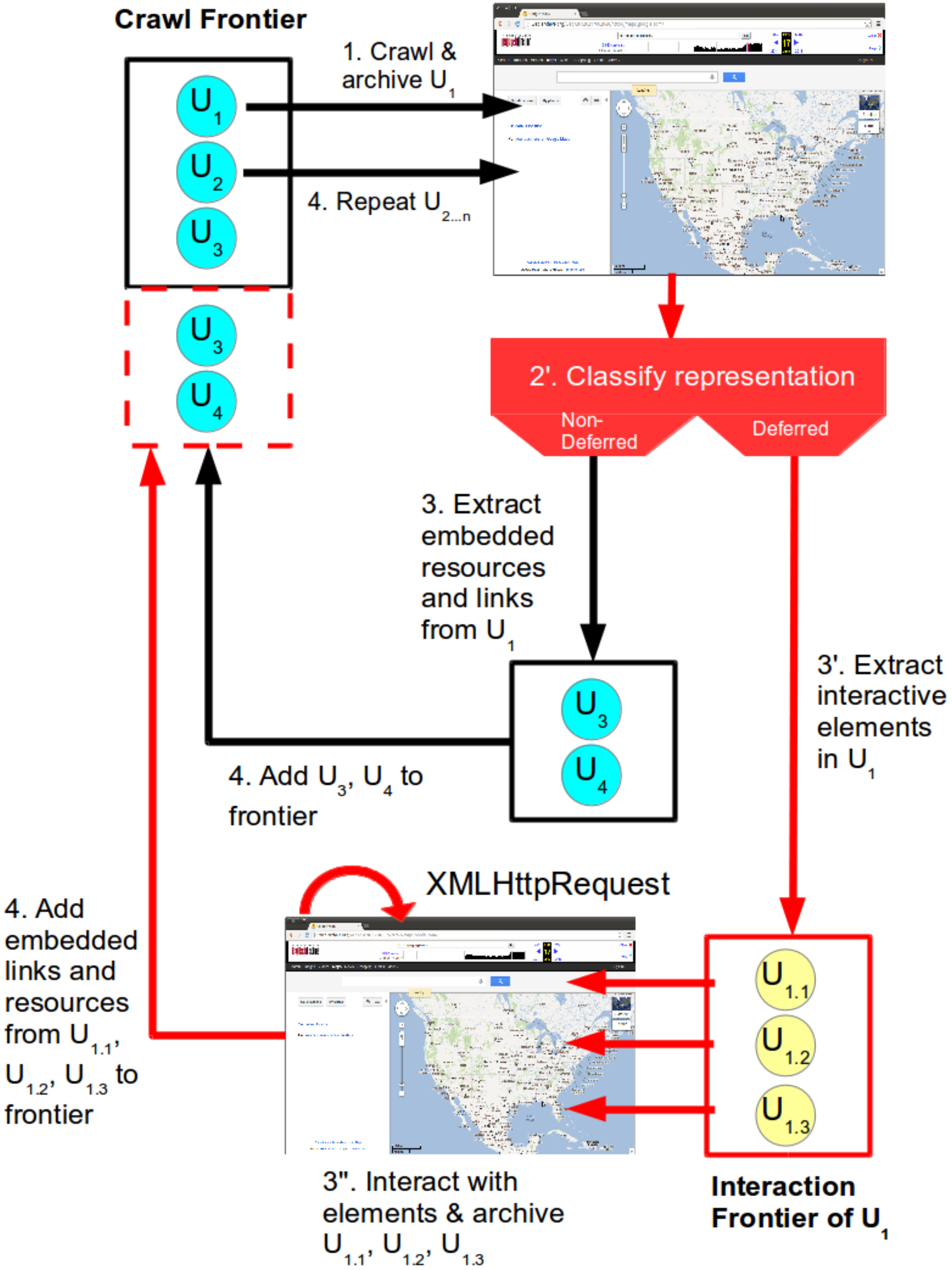
Interaction Frontier of U_1

rrent workflow not suitable for deferred
presentations

<script> tags alone are not indicative of a deferred
epresentation. JavaScript can be played back in the
archives!

Two-tiered crawling approach to optimize
performance

Use PhantomJS to run JavaScript, interact with the
epresentation


Crawl Frontier
U1
U2
U3
U3
U4
1. Crawl & archive U1
4. Repeat U2...n
2'. Classify representation
Non-Deferred
Deferred
3. Extract embedded resources and links from U1
U3
U4
4. Add U3, U4 to frontier
3'. Extract interactive elements in U1
XMLHttpRequest
4. Add embedded links and resources from U1.1, U1.2, U1.3 to frontier
U1.1
U1.2
U1.3
3". Interact with elements & archive U1.1, U1.2, U1.3
Interaction Frontier of U1

More URI-Rs in the crawl frontier

Runs more slowly bu more deeply

rrent workflow not suitable for deferred resentations

script> tags alone are not indicative of a deferred epresentation. JavaScript can be played back in the rchives!

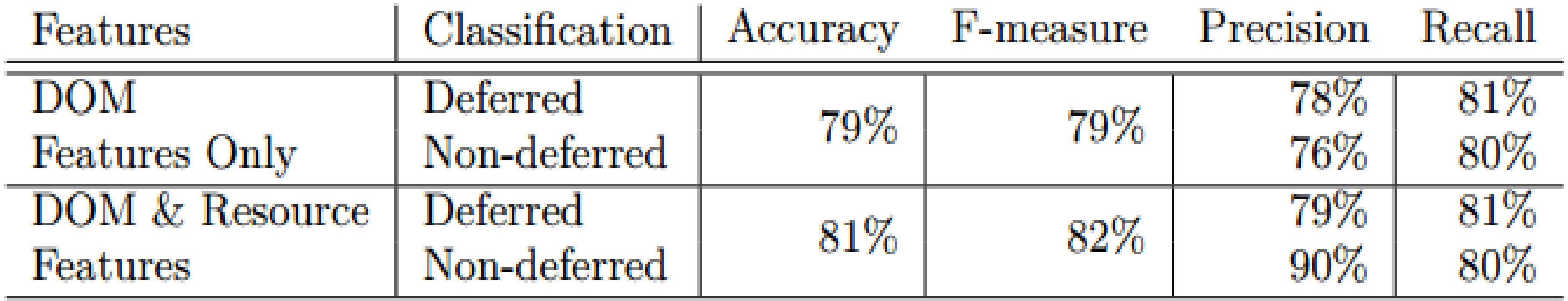
wo-tiered crawling approach to optimize erformance

Jse PhantomJS to run JavaScript, interact with the epresentation

Classifying deferred representations

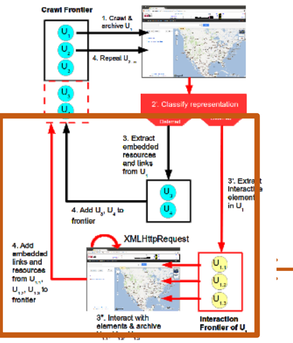
- Manually classify 440 URIs (generated from random bitlys) as deferred or non-deferred; build classifier based on 12 different features (8 DOM-based, 4 resource-based)
- On a 10,000 URI set (random bitlys, including 440 from before) compare crawl speed & discovered frontier size with and without classifier
-
- Data set & code available at:
- https://github.com/jbrunelle/classifyDeferred/
- https://github.com/jbrunelle/DataSets

Classifier accuracy improved slightly when monitoring HTTP requests

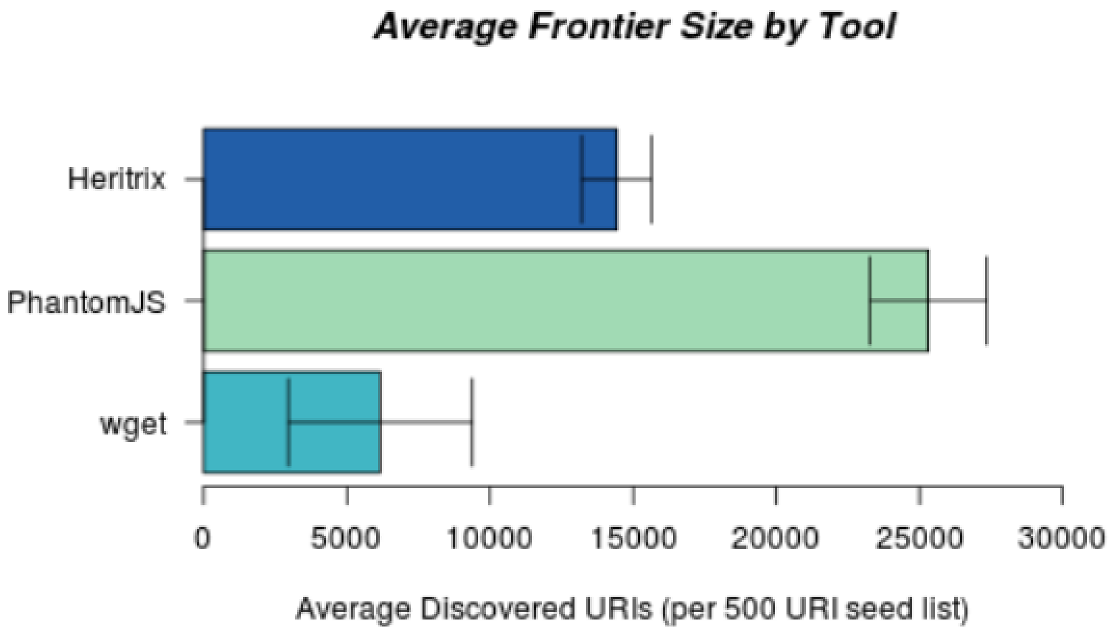


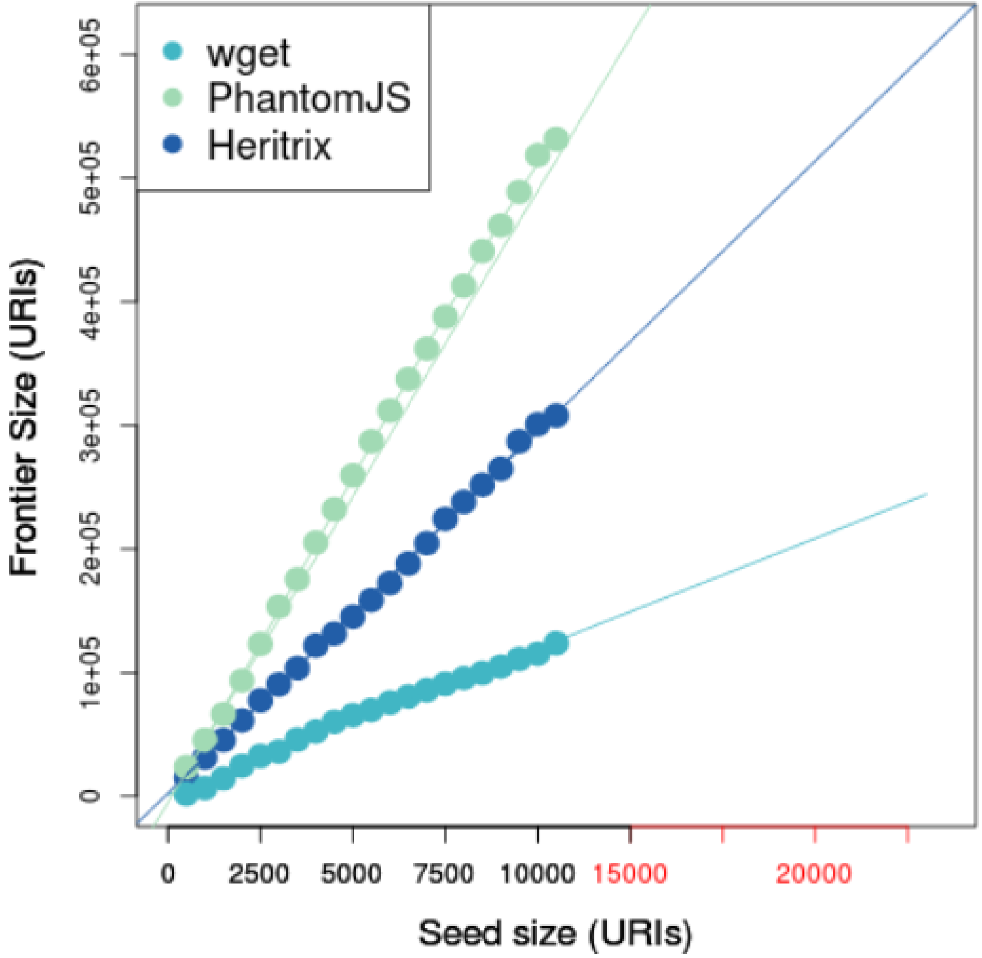
Features	Classification	Accuracy	F-measure	Precision	Recall
DOM Features Only	Deferred	79%	79%	78%	81%
	Non-deferred			76%	80%
DOM & Resource Features	Deferred	81%	82%	79%	81%
	Non-deferred			90%	80%

Table 8: Classification success statistics for DOM-only and DOM and Resource feature sets.


Crawl Frontier
U1
U2
U3
1. Crawl & archive U1
4. Repeat U2..n
U3
U4
2'. Classify representation
Deferred
3. Extract embedded resources and links from U1
U3
U4
4. Add U3, U4 to frontier
3'. Extract interactive elements in U1
XMLHttpRequest
4. Add embedded links and resources from U1,1, U1,2, U1,3 to frontier
U1,1
U1,2
U1,3
3''. Interact with elements & archive
Interaction Frontier of U1

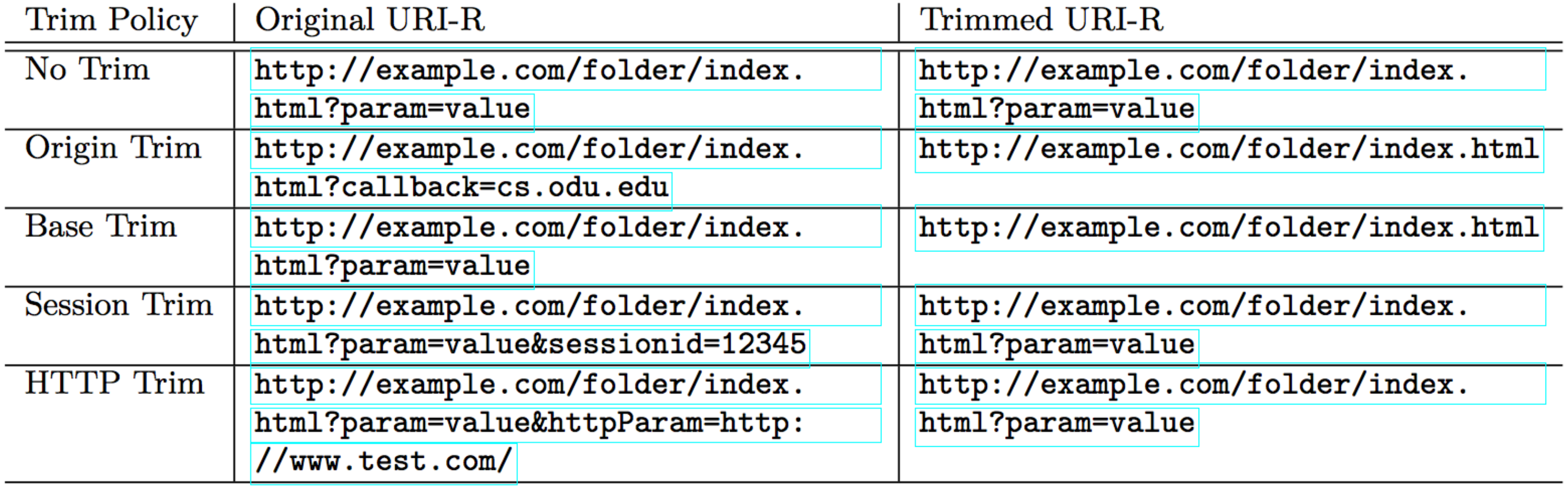
Performance: Frontier Size





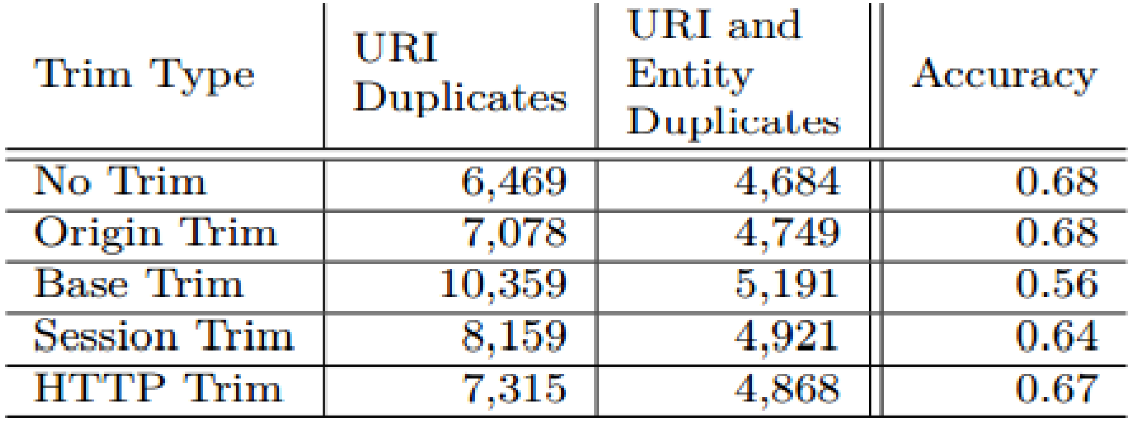
PhantomJS creates a 1.5x larger crawl frontier than Heritrix

Are all those URIs the same?



Trim Policy	Original URI-R	Trimmed URI-R
No Trim	http://example.com/folder/index.html?param=value	http://example.com/folder/index.html?param=value
Origin Trim	http://example.com/folder/index.html?callback=cs.odu.edu	http://example.com/folder/index.html
Base Trim	http://example.com/folder/index.html?param=value	http://example.com/folder/index.html
Session Trim	http://example.com/folder/index.html?param=value&sessionid=12345	http://example.com/folder/index.html?param=value
HTTP Trim	http://example.com/folder/index.html?param=value&httpParam=http://www.test.com/	http://example.com/folder/index.html?param=value

Table 3: Examples of the URI trim policies.

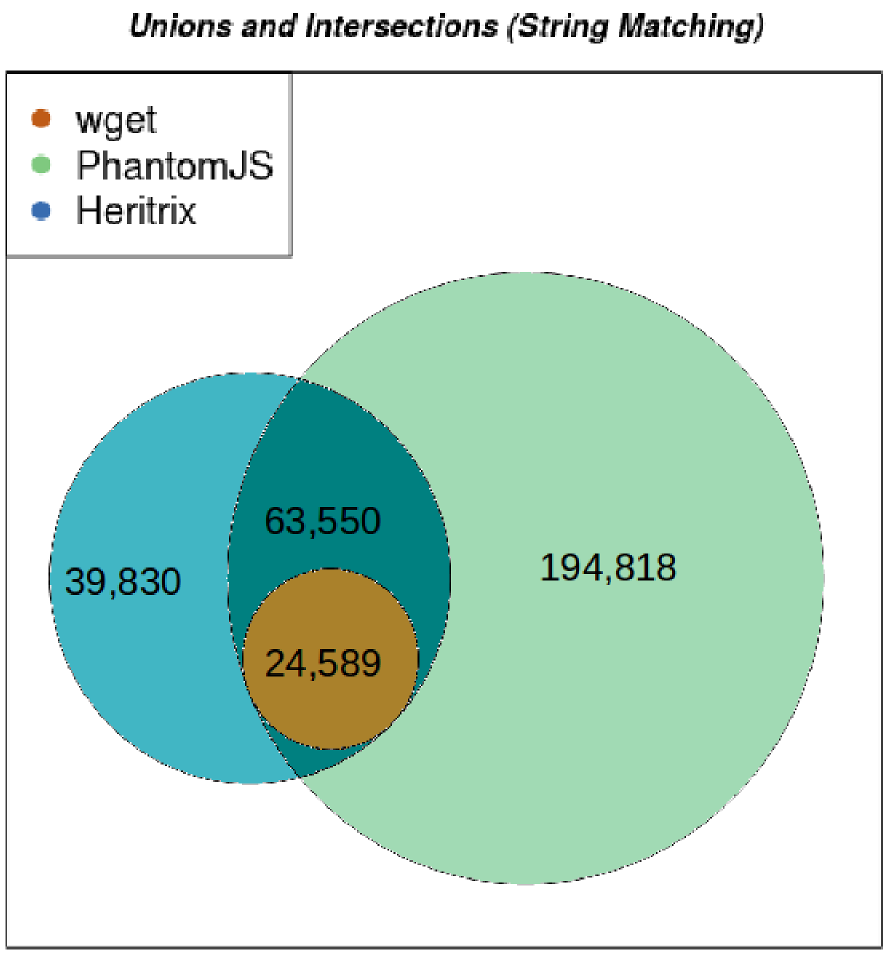


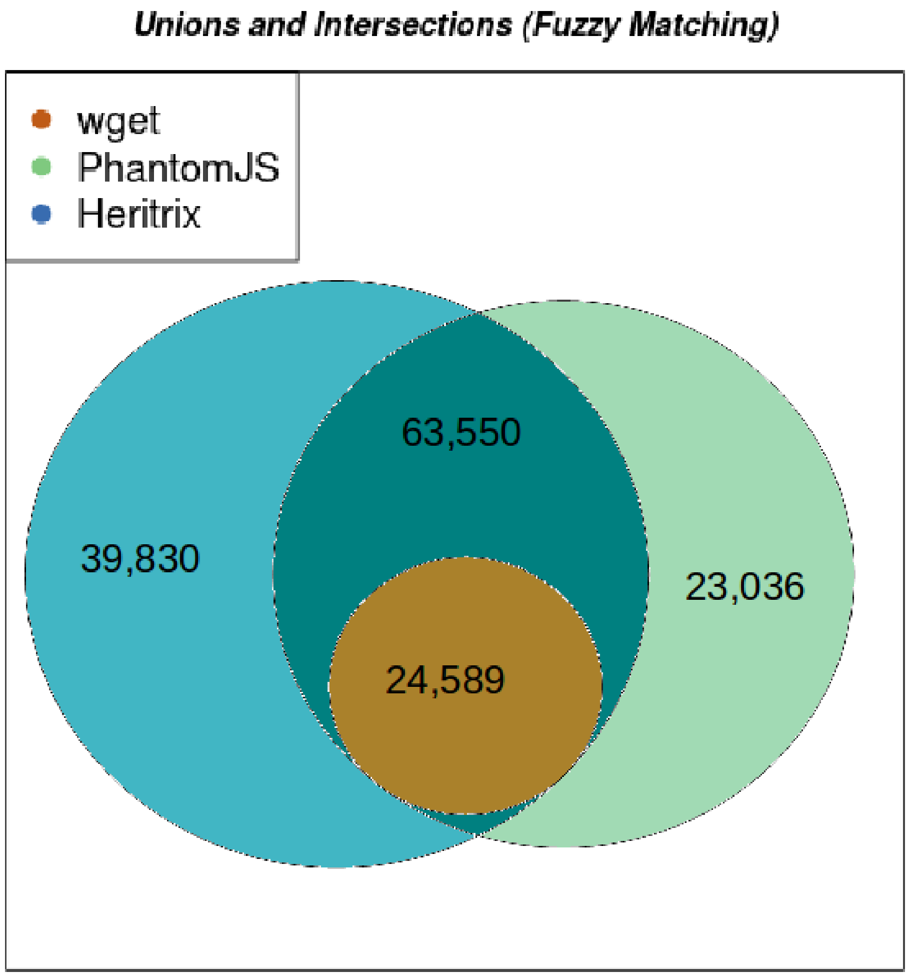
Trim Type	URI Duplicates	URI and Entity Duplicates	Accuracy
No Trim	6,469	4,684	0.68
Origin Trim	7,078	4,749	0.68
Base Trim	10,359	5,191	0.56
Session Trim	8,159	4,921	0.64
HTTP Trim	7,315	4,868	0.67

Table 4: Detected duplicate URIs, entity bodies, and the overlap between the two using the five URI string trimming policies.

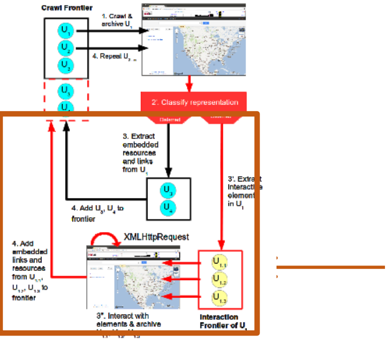
TP = URIs match & entities match
TN = neither URI nor entity matches
P + N = 19,522

Trimming shrinks the PhantomJS Frontier

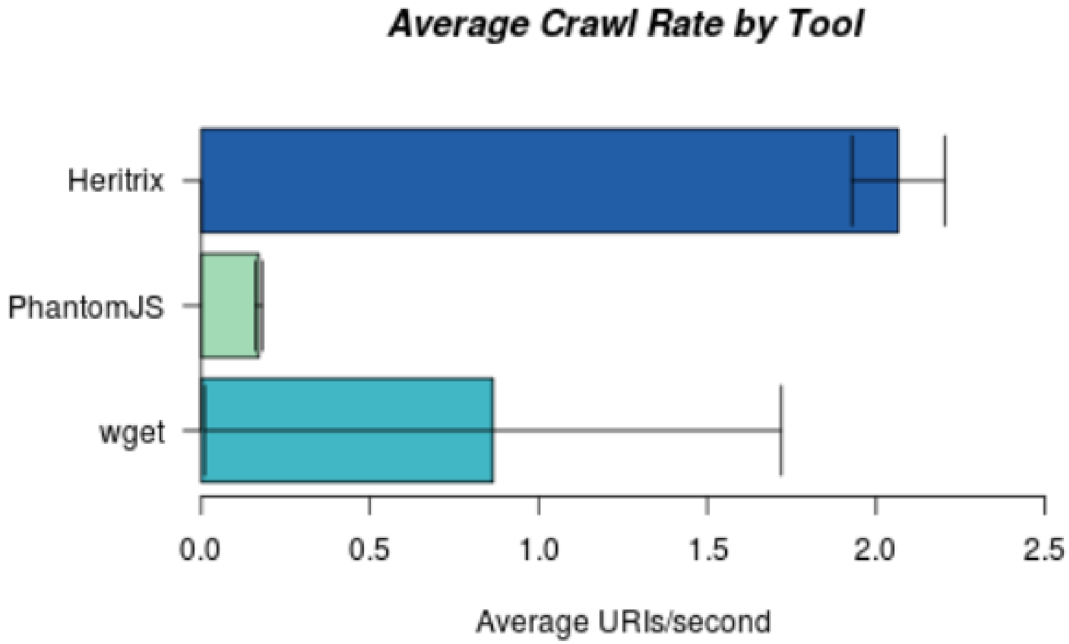


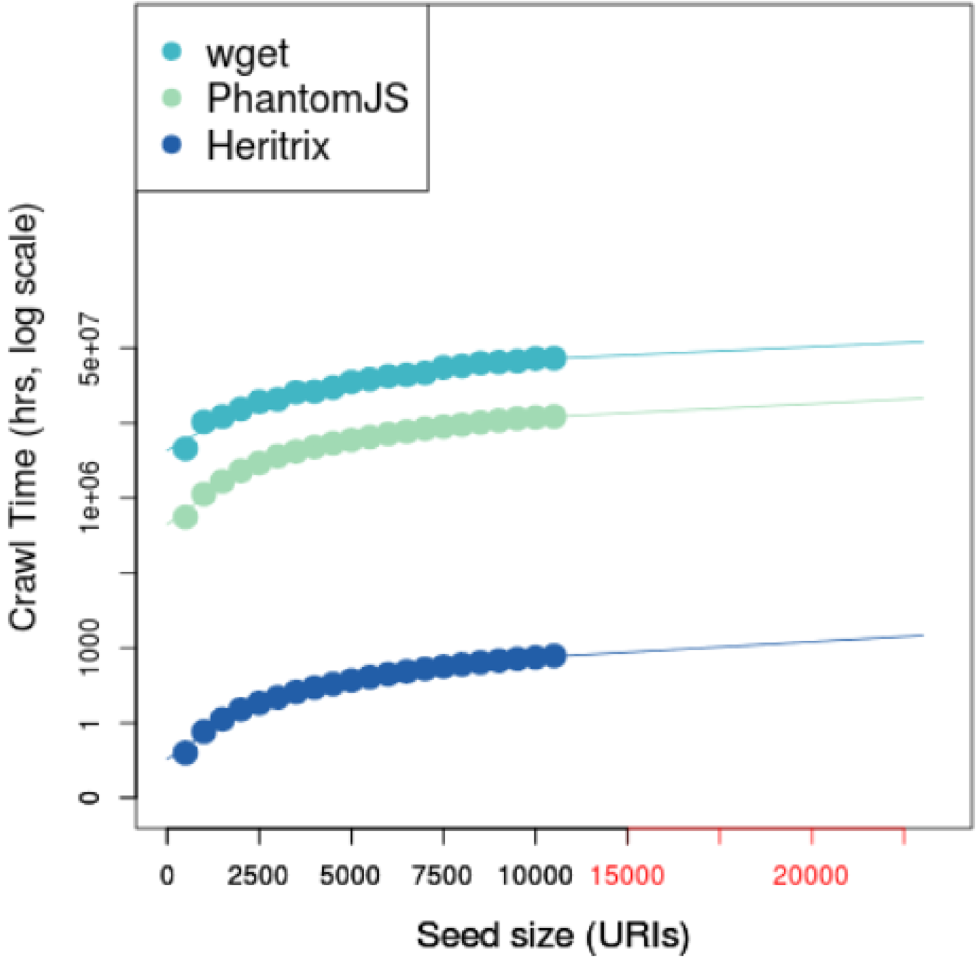


(Base policy shown)


Crawl Frontier
1. Crawl & archive U1
4. Repeat U2..n
U1
U2
U3
U3
U4
2'. Classify representation
Deferred
3. Extract embedded resources and links from U1
3'. Extract interactive elements in U1
4. Add U3, U4 to frontier
XMLHttpRequest
4. Add embedded links and resources from U1.1, U1.2, U1.3 to frontier
U1.1
U1.2
U1.3
3". Interact with elements & archive
Interaction Frontier of U1

Performance: Crawl Speed

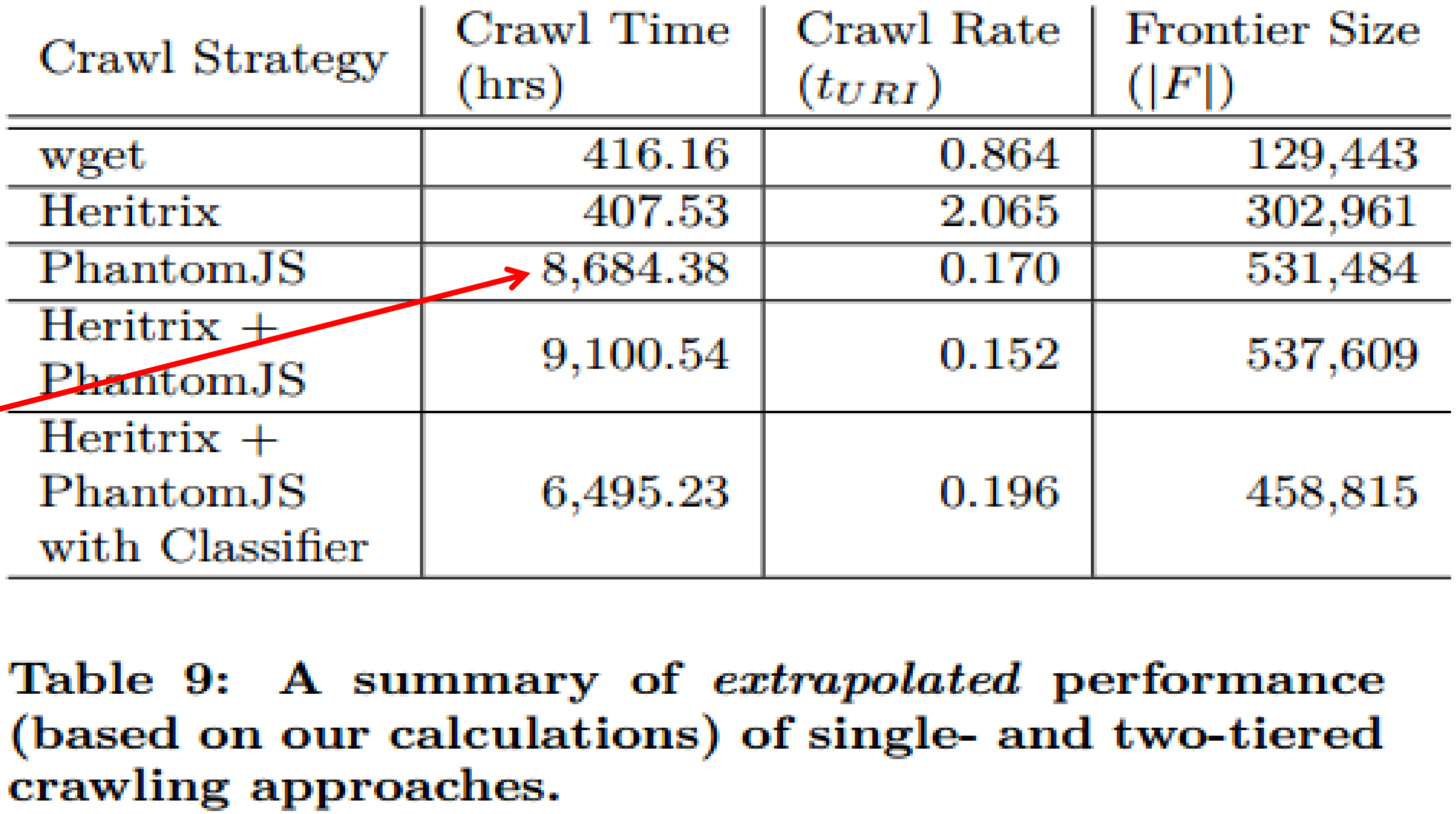




Heritrix: ~2 URIs/second

PhantomJS: ~4 seconds/URI

How long would it take to crawl *everything*?



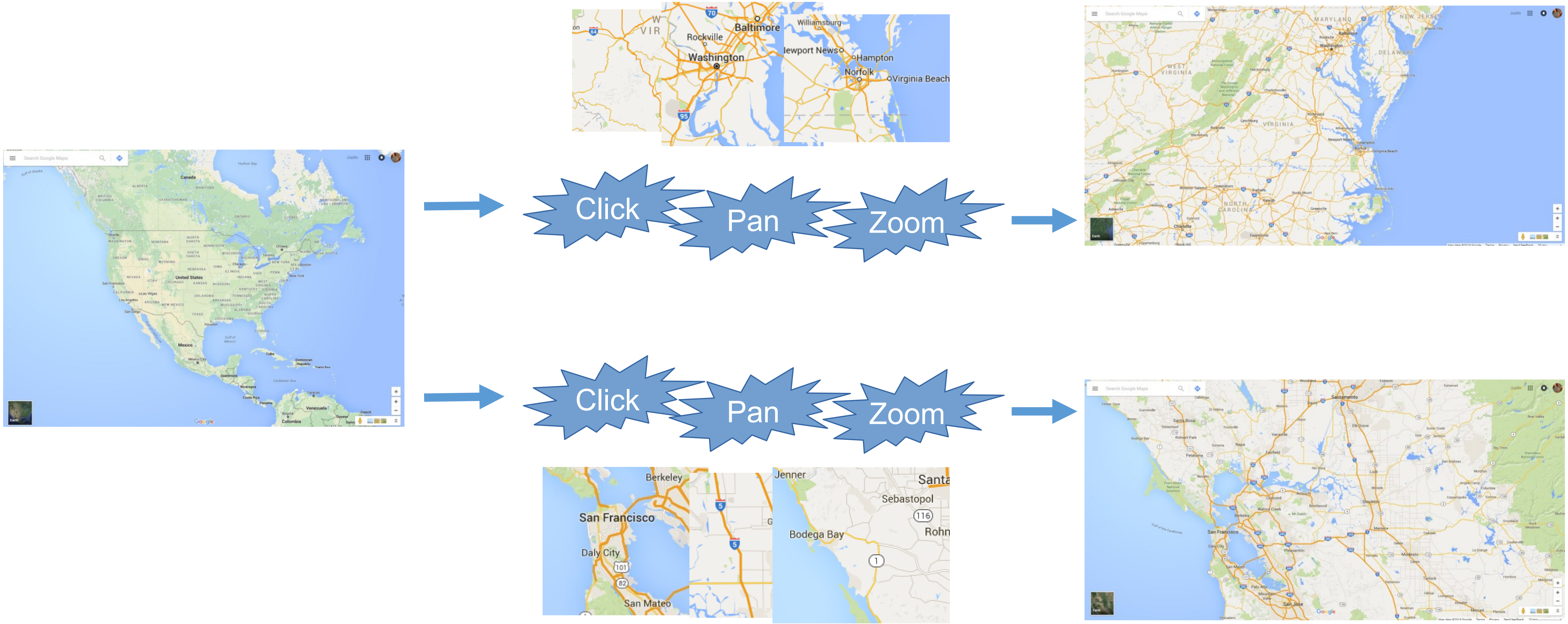
Crawl Strategy	Crawl Time (hrs)	Crawl Rate (t_{URI})	Frontier Size ($\|F\|$)
wget	416.16	0.864	129,443
Heritrix	407.53	2.065	302,961
PhantomJS	8,684.38	0.170	531,484
Heritrix + PhantomJS	9,100.54	0.152	537,609
Heritrix + PhantomJS with Classifier	6,495.23	0.196	458,815

Table 9: A summary of *extrapolated* performance (based on our calculations) of single- and two-tiered crawling approaches.

nearly a year!

(obviously parallelization would help)

Descendants = States of deferred representations reached through client-side events



Click Pan Zoom

Click Pan Zoom

Finding descendants

- Return to the same 440 URIs from before
-
- Use VisualEvent to identify interactive elements
- http://ws-dl.blogspot.com/2015/06/2015-06-26-phantomjsvisualevent-or.html
-
- Adapting work on state equivalency based on DOM equivalency, we define state equivalency as requiring the same embedded resources

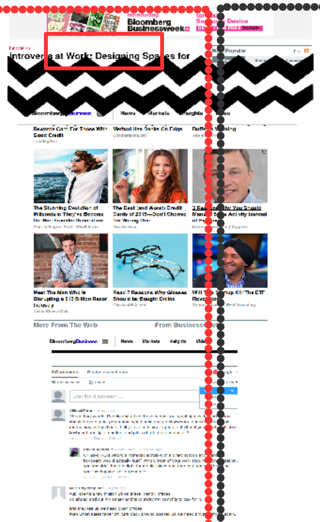
- Report & code:
- http://arxiv.org/abs/1601.05142
- https://github.com/jbrunelle/clientSideState
-
-

Interaction Trees are 2 Levels Deep

s_0

s_1

s_2



Bloomberg Businessweek

Introverts at Work: Designing Spaces for

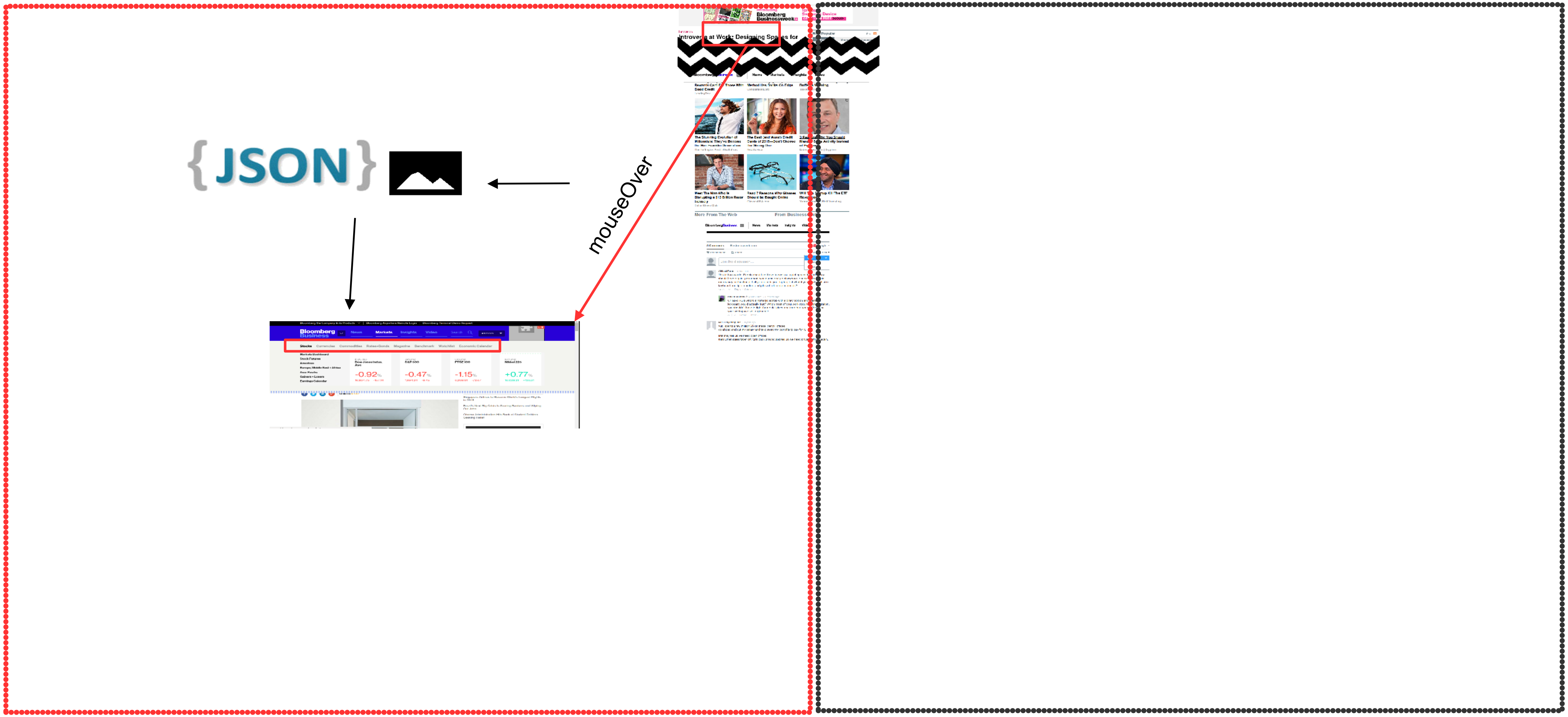
http://www.bloomberg.com/bw/articles/2014-06-16/open-plan-offices-for-people-who-hate-open-plan-offices

Interaction Trees are 2 Levels Deep

s_0

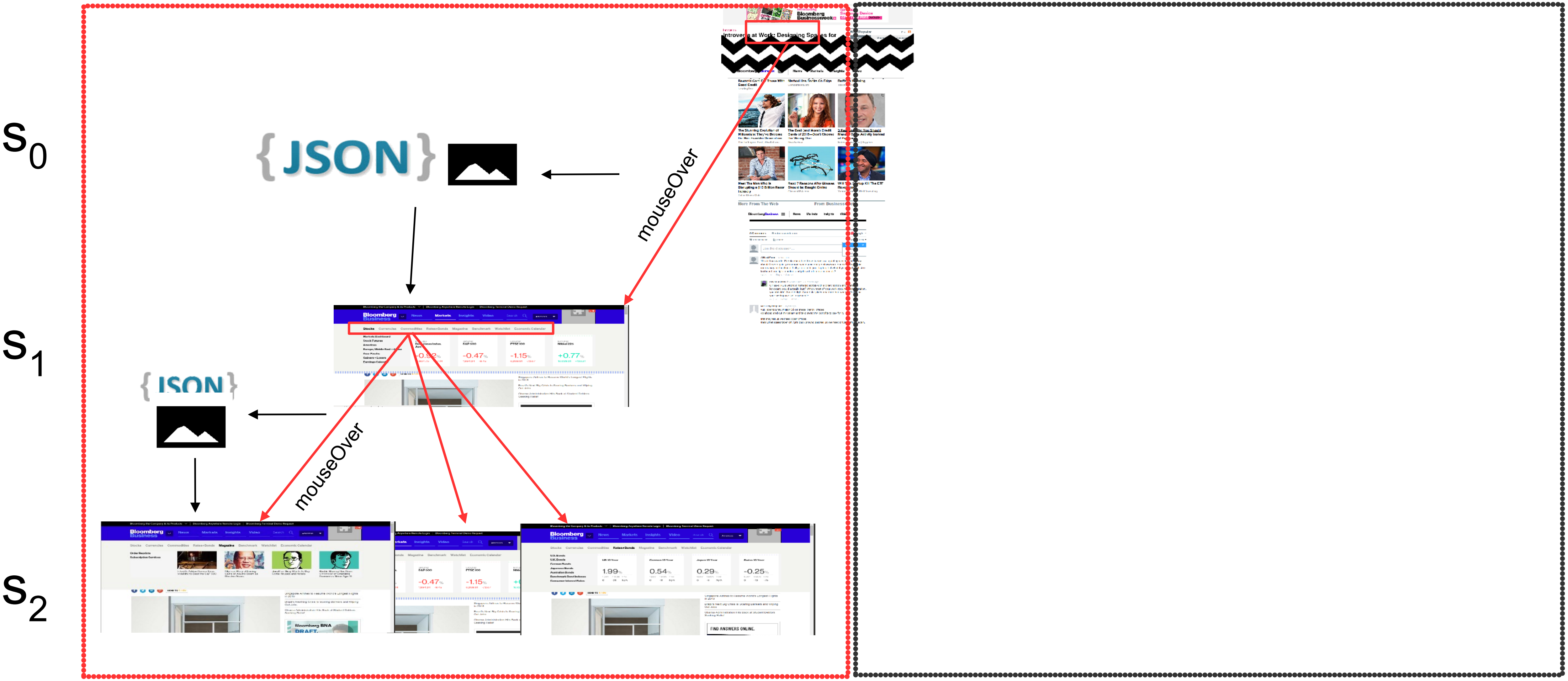
s_1

s_2



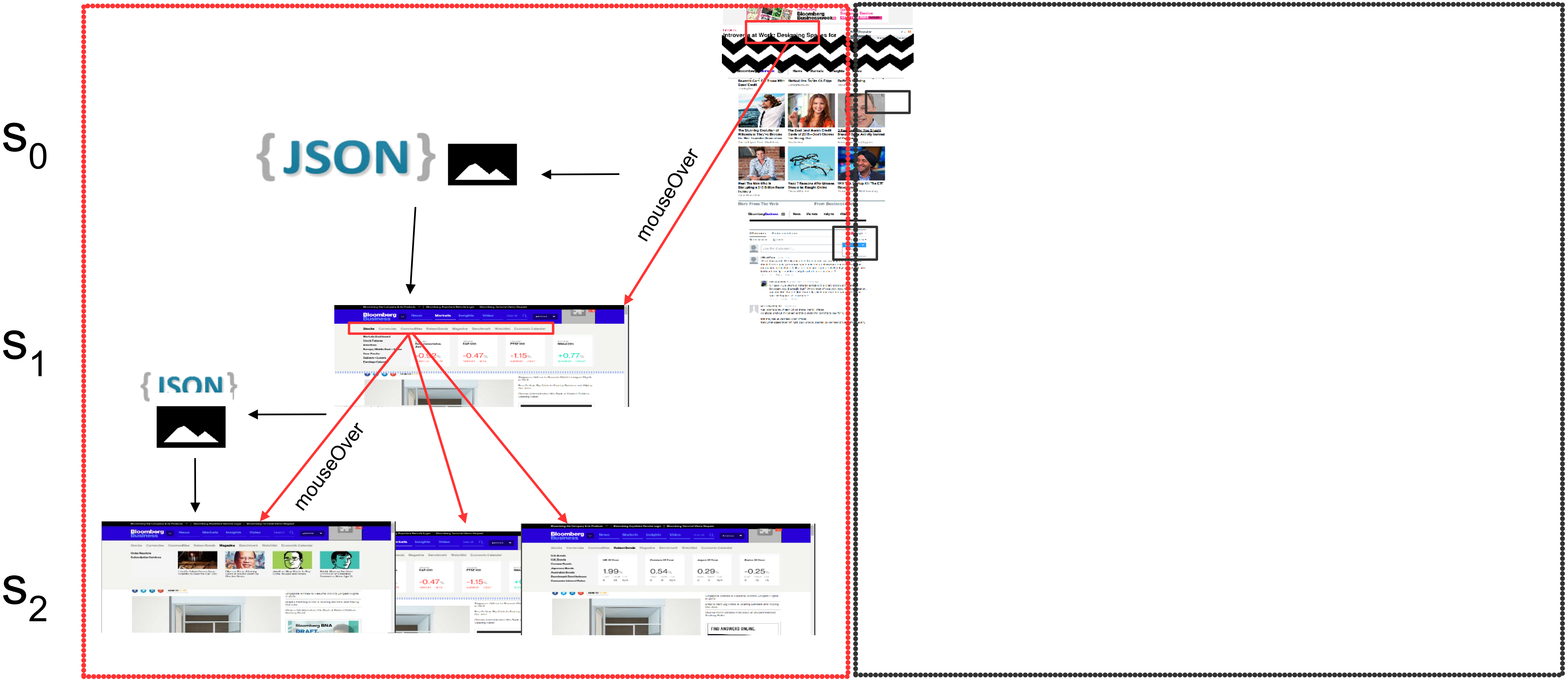
http://www.bloomberg.com/bw/articles/2014-06-16/open-plan-offices-for-people-who-hate-open-plan-offices

Interaction Trees are 2 Levels Deep



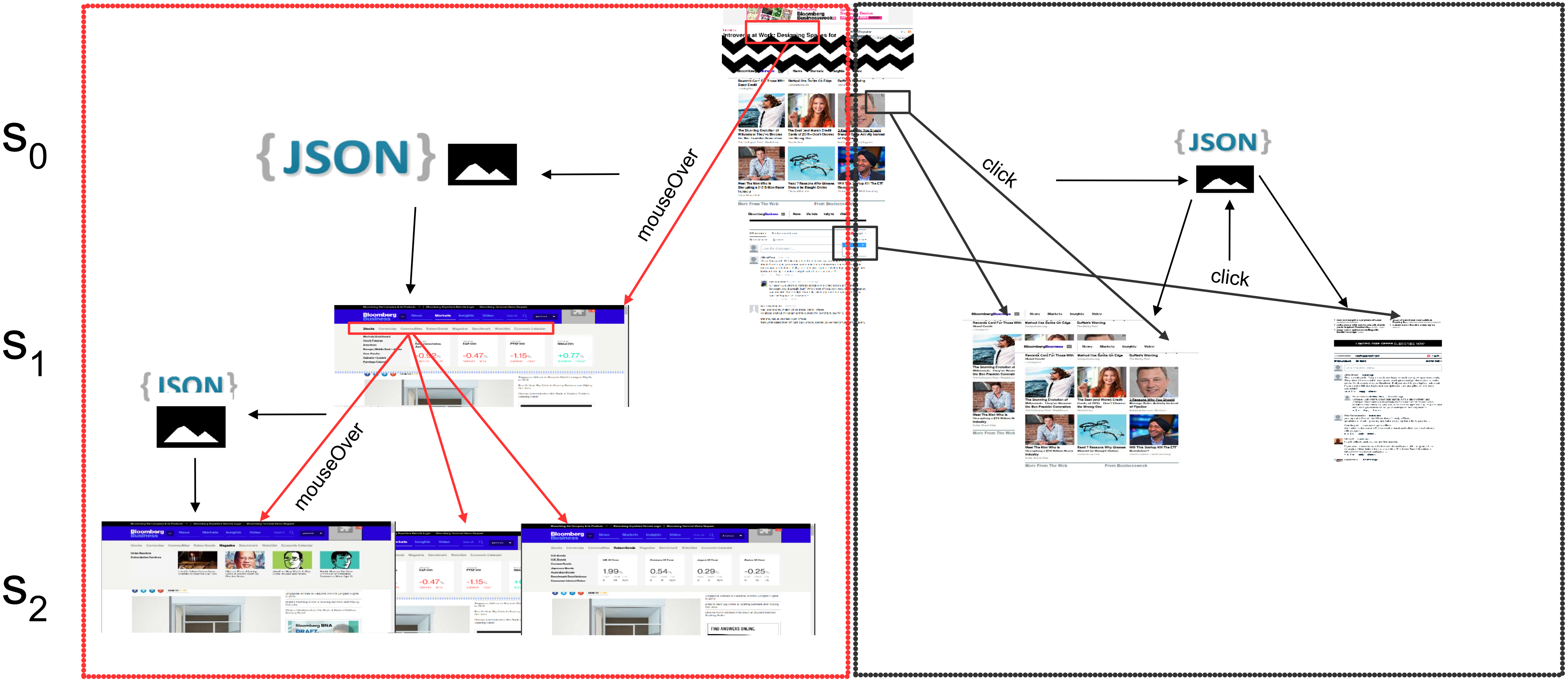
http://www.bloomberg.com/bw/articles/2014-06-16/open-plan-offices-for-people-who-hate-open-plan-offices

Interaction Trees are 2 Levels Deep



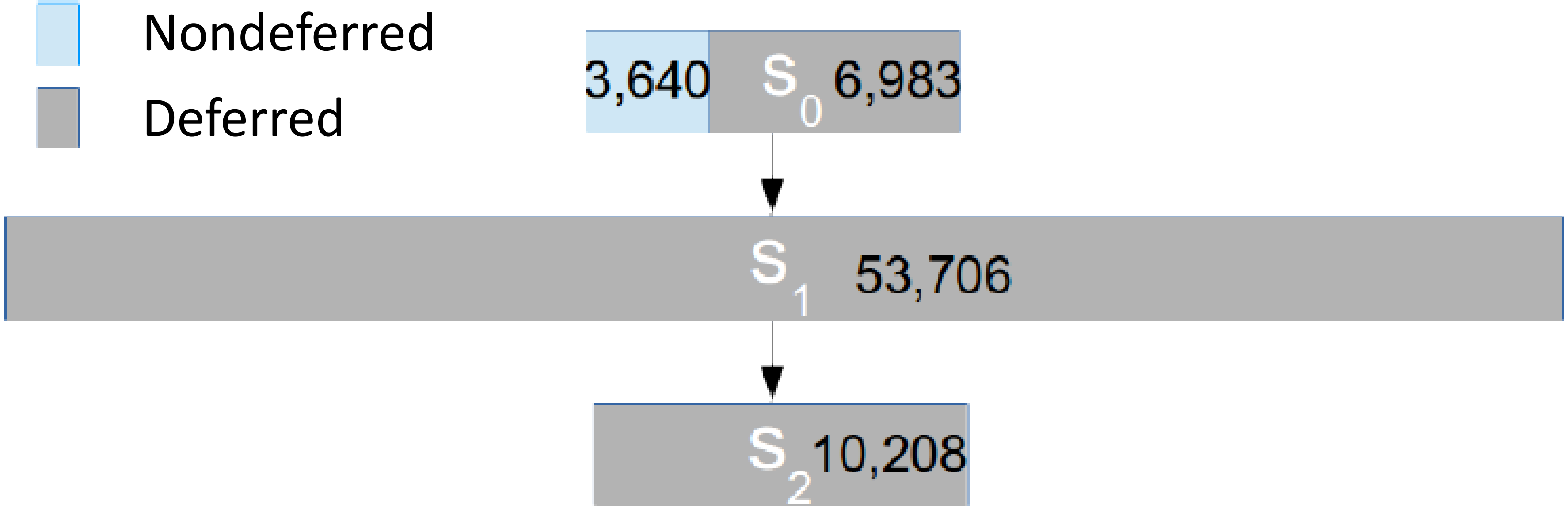
http://www.bloomberg.com/bw/articles/2014-06-16/open-plan-offices-for-people-who-hate-open-plan-offices

Interaction Trees are 2 Levels Deep



http://www.bloomberg.com/bw/articles/2014-06-16/open-plan-offices-for-people-who-hate-open-plan-offices

Expanding the Crawl Frontier



Nondeferred

Deferred

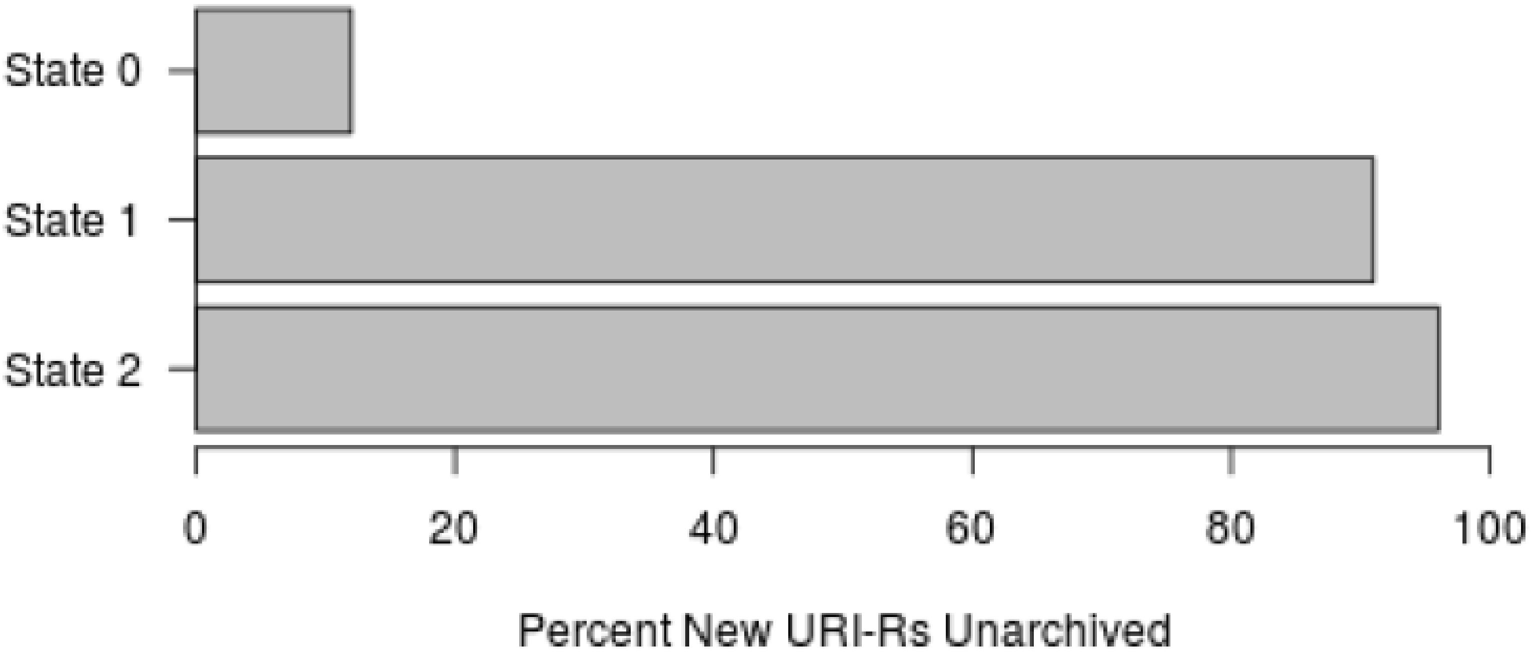
3,640 S_0 6,983

S_1 53,706

S_2 10,208

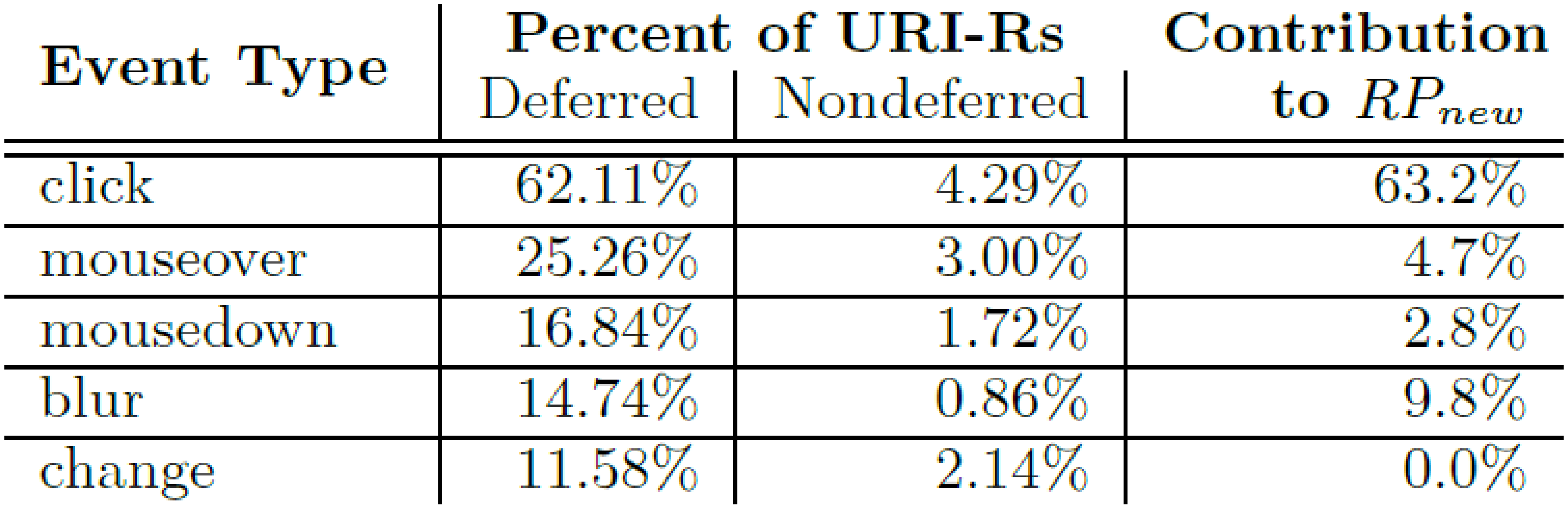
Level s_1 provides the greatest benefit to the crawl frontier

Crawling Descendants



New embedded resources at levels s_1 are largely unarchived

Expanding the crawl frontier



Event Type	Percent of URI-Rs		Contribution
	Deferred	Nondeferred	to RP_{new}
click	62.11%	4.29%	63.2%
mouseover	25.26%	3.00%	4.7%
mousedown	16.84%	1.72%	2.8%
blur	14.74%	0.86%	9.8%
change	11.58%	2.14%	0.0%

Click events lead to the most descendants

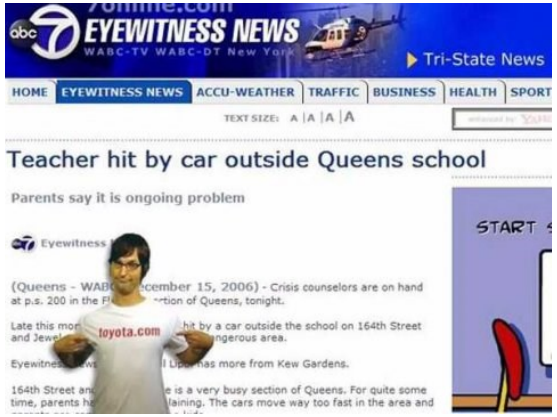
Future Work

- Modeling user interactions, tendencies, and simulation
  - form filling
  - click & navigation likelihood
  - Added frontier 92% unarchived
- Archival Halting Problem: How much is enough?
  - Mapping Applications – How many pans and zooms gets all the Norfolk, VA Google map tiles?
  - How many CNN.com pages get all the Google Ads?
  - Game walkthrough metaphor? (insert url here)
- Playing back WARCs with IIPC metadata of deferred representations and descendants

Contributions

- Defined:
  - *deferred representations*: representations that need client-side processing to load all required embedded resources
  - *descendants*: representation states reachable only via client-side events
- Two-tiered crawling of deferred representations
  - 10.5 times slower
  - 1.5 times larger frontier
  - 2 levels of descendants
- 2 levels are sufficient for descendants
  - Added frontier 92% unarchived
- More info:
  - http://arxiv.org/abs/1508.02315
  - http://arxiv.org/abs/1601.05142

He's one of the busiest men in town. While his door may say Office Hours 2 to 4, he's actually on call 24 hours a day.
The doctor is a scientist, a diplomat, and a friendly sympathetic human being all in one, no matter how long and hard his schedule.
According to a recent Nationwide survey:
MORE DOCTORS SMOKE CAMELS
THAN ANY OTHER CIGARETTE
Your "T-Zone" Will Tell You...
T for Taste...
T for Throat...
that's your proving ground for any cigarette. See if Camels don't suit your "T-Zone" to a "T."
CAMEL
CAMELS Costlier Tobaccos



Teacher hit by car outside Queens school

Parents say it is ongoing problem

Eyewitness

(Queens - WAB ecember 15, 2006) - Crisis counselors are on hand at p.s. 200 in the F rtion of Queens, tonight.

Late this mor hit by a car outside the school on 164th Street and Jewe ngerous area.

Eyewitne l L has more from Kew Gardens.

164th Street an e is a very busy section of Queens. For quite some time, parents h laining. The cars move way too fast in the area and