



OLD DOMINION
UNIVERSITY
IDEA FUSION

IIPC 2015

APRIL 27–MAY 1, 2015
STANFORD UNIVERSITY

EVALUATING THE TEMPORAL COHERENCE OF ARCHIVED PAGES

SCOTT G. AINSWORTH
MICHAEL L. NELSON
OLD DOMINION UNIVERSITY

HERBERT VAN DE SOMPEL
LOS ALAMOS NATIONAL
LABORATORY



OLD DOMINION
UNIVERSITY
IDEA FUSION

IIPC 2015

APRIL 27–MAY 1, 2015
STANFORD UNIVERSITY

HE WENT TO VIEW AN ARCHIVED PAGE. YOU WON'T BELIEVE WHAT HE SAW NEXT...

SCOTT G. AINSWORTH
MICHAEL L. NELSON
OLD DOMINION UNIVERSITY

HERBERT VAN DE SOMPEL
LOS ALAMOS NATIONAL
LABORATORY

CONTENTS

- Motivation**
- Composite Mementos**
- Coherence Framework**
- Temporal Coherence**
- Future Work**
- Conclusion**

RESEARCH TO DATE

- **How to crawl a site to maximize coherence**
 - Ben Saad et al., JCDL 2011, TPDL 2011
- **Detecting, visualizing temporal defects**
 - Spaniol et al., WICOW 2009, IWAW 2009, VLDB 2009

RESEARCH TO DATE: @WEBSCIDL

- **How much of the web is archived?**
 - Ainsworth et al., JCDL 2011
- **Are the archives stable?**
 - Brunelle et al., JCDL 2013
- **Temporal drift while browsing in an archive?**
 - Ainsworth et al., JCDL 2013
- **Are the missing resources important?**
 - Brunelle et al., JCDL 2014
- *Are the present resources correct?*

AS PRESENTED BY IA

Internet Archive Wayback Machine
<http://www.wunderground.com/cgi-bin/findweather/getForecast?query=> **21 captures** 27 Oct 02 - 25 Jan 11

Find the weather for any **City, State** or **ZIP Code**, or **Airport Code** or **Country**

Member Benefits:
No Ads
Weather Email
\$5 a year
[Signup Here](#)

Email

 Password

[Print This Page](#)
Maps
[Temperature](#)
[Heat Index](#)

Varina, Iowa

Local Time: 1:09 PM CST [Set My Timezone](#) Lat/Lon: 42.6° N 94.8° W | [MSN Map](#)

Current Conditions		5-Day Forecast for ZIP Code 50593				
Updated: 12:55 PM CST on December 09, 2004		Thu	Fri	Sat	Sun	Mon
Observed at Storm Lake, Iowa (History)						
Elevation: 1486 ft / 453 m		43° 29°	36° 22°	47° 34°	41° 15°	31° 15°
41 °F / 5 °C Light Drizzle		Chance of Rain	Mostly Cloudy	Mostly Cloudy	Partly Cloudy	Partly Cloudy
Windchill: 37 °F / 3 °C		Detail	Detail	Detail	Detail	Detail
Humidity: 100%		Click Detail for hourly wind, temperature, humidity and UV forecasts.				
Dew Point: 41 °F / 5 °C		Alternate Computer Forecast: AVN MOS Weather Graph Local Allergy Info				
Wind: 6 mph / 8 km/h from		from Pollen.com				

<http://web.archive.org/web/20041209190926/http://www.wunderground.org/cgi-bin/findWeather/getForecast?query=50593> (now 404, but that's a different story...)

NOT ALL 2004-12-09T19:09:26

Browser address bar: <https://web.archive.org/web/20041209190926/http://www.wunderground.com/cgi-bin>

Wayback Machine: <http://www.wunderground.com/cgi-bin/findweather/getForecast?query=> Go

Wayback Machine calendar: DEC 9 2004

Search: Find the weather for any City, State or ZIP Code, or Airport Code or Country

missing

-15 hours

Member Benefits: No Ads, Weather Email \$5 a year, Signup Here

Email, Password, Login

Print This Page, Maps, Temperature, Heat Index

Varina, Iowa

Local Time: 1:09 PM CST [Set My Timezone](#) Lat/Lon: 42.6° N 94.8° W | [MSN Map](#)

Current Conditions		5-Day Forecast for ZIP Code 50593				
Updated: 12:55 PM CST on December 09, 2004		Thu	Fri	Sat	Sun	Mon
Observed at Storm Lake, Iowa (History)						
Elevation: 1486 ft / 453 m		43° 29°	36° 21°	15°	31° 15°	
41 °F / 5 °C Light Drizzle		Chance of Rain	Mostly Clear	Cloudy	Partly Cloudy	
Windchill: 37 °F / 3 °C		Detail	Detail	Detail	Detail	
Humidity: 100%		Click Detail for hourly wind, temperature, humidity and UV forecasts.				
Dew Point: 41 °F / 5 °C		Alternate Computer Forecast: AVN MOS Weather Graph Local Allergy Info from Pollen.com				
Wind: 6 mph / 8 km/h from						

CLEAR OR CLOUDY?

Current Conditions
CST on December 09,
Lake, Iowa ([History](#))
453 m

5 °C
3 °C
5 °C



5-Day Forecast

Thu	Fri
	
42° 50°	36° 20°
Chance of Rain	Mostly Clear
Detail	Detail

Click **Detail** for hourly wind, temperature, and precipitation.

Alternate Computer Forecast: [All from Pollen.com](#)

+9 months

Meet Locals



QUESTIONS

- **How prevalent is temporal incoherence?**
- **Can temporal coherence be improved by using multiple archives?**
- **Can temporal coherence be improved by introducing memento selection heuristics?**

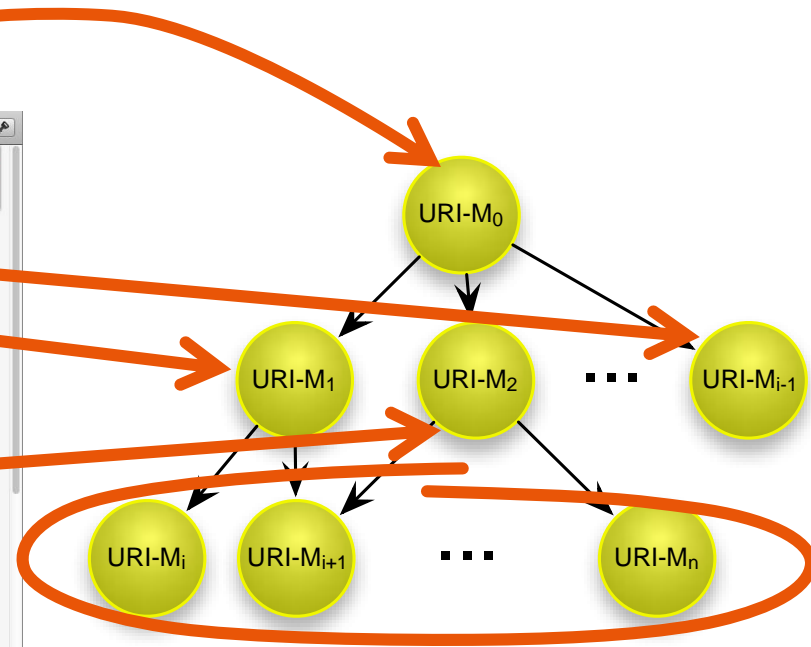
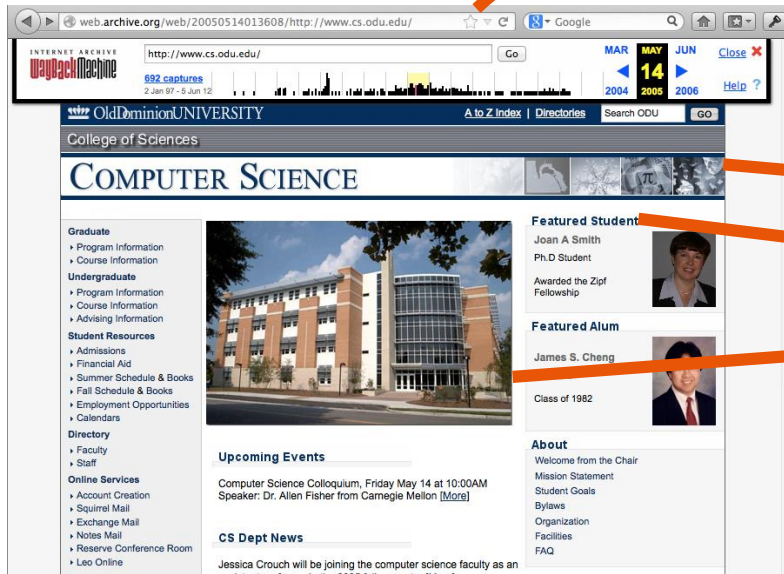
CONTENTS

- Motivation**
- Composite Memento**
- Coherence Framework**
- Temporal Coherence**
- Future Work**
- Conclusion**

COMPOSITE MEMENTO

PRESENTATION

STRUCTURE



CONTENTS

- Motivation**
- Composite Memento**
- Coherence Framework**
- Temporal Coherence**
- Future Work**
- Conclusion**

COHERENCE STATES

- **Prima Facie Coherent**

Evidence that the memento **existed** in its archived state when the root was acquired.

- **Prima Facie Violative**

Evidence ... **did not exist** ...

- **Possibly Coherent**

Evidence ... **might have existed** ...

- **Probably Violative**

Evidence ... **probably did not exist** ...

CONSIDER THIS PAGE...

```
<html>  
  
</html>
```

WITH THESE RESPONSE HEADERS

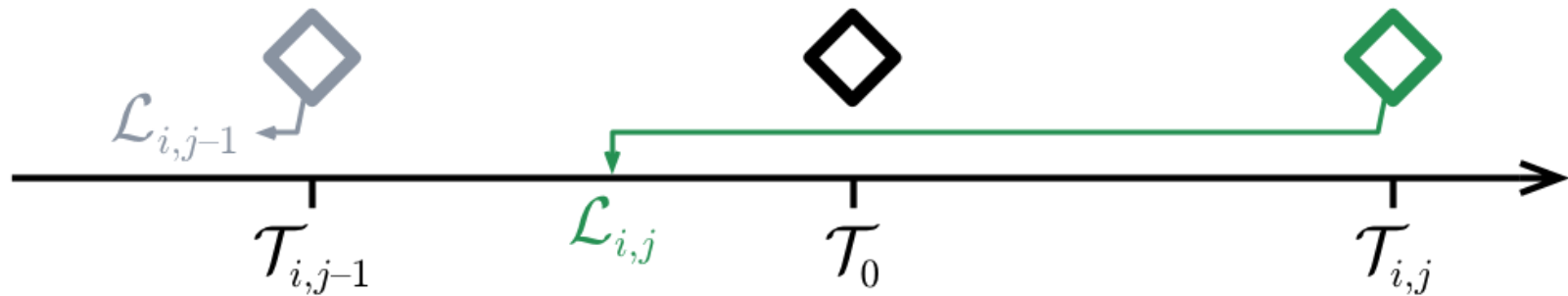
```
HTTP/1.1 200 OK
Server: Tengine/2.0.3
Date: Mon, 27 Apr 2015 22:03:32 GMT
Content-Type: image/jpeg
Content-Length: 15632
Connection: keep-alive
Memento-Datetime: Tue, 07 Feb 2006 00:58:23 GMT
Link: <Memento links deleted...>
X-Archive-Orig-server: Apache/1.3.26 (Unix) ApacheJServ/1.1.2 PHP/4.3.4
X-Archive-Orig-etag: "4978-3d10-3e4d822e"
X-Archive-Orig-content-length: 15632
X-Archive-Orig-accept-ranges: bytes
X-Archive-Orig-date: Tue, 07 Feb 2006 00:58:20 GMT
X-Archive-Orig-content-type: image/jpeg
X-Archive-Orig-last-modified: Fri, 14 Feb 2003 23:56:30 GMT
X-Archive-Orig-connection: close
<other headers deleted>
```

PRIMA FACIE COHERENT

Bracket Pattern:

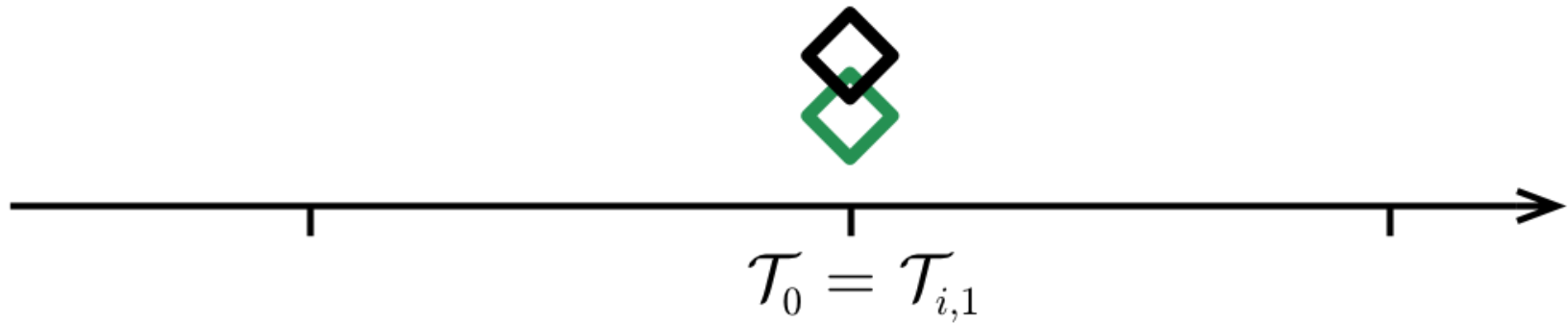
Memento-Datetime + Last-Modified

(yes, Last-Modified is sometimes wrong, but many of those cases can be detected)



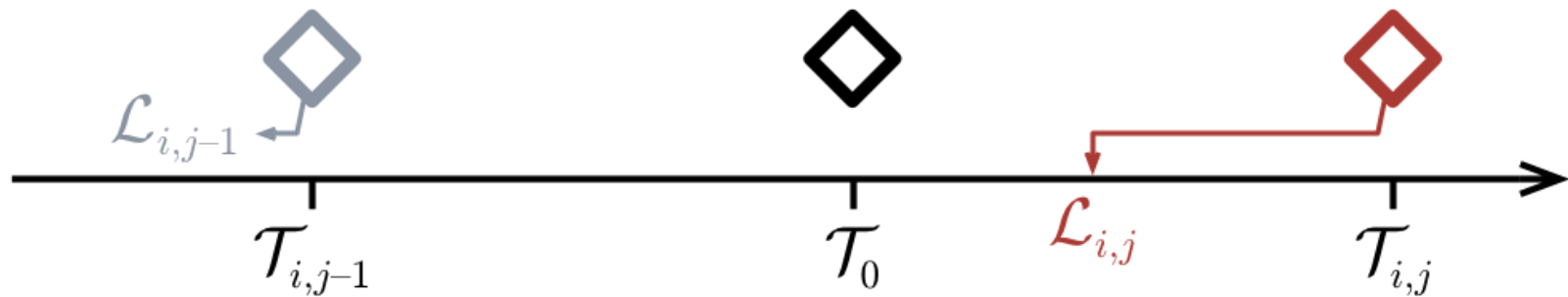
PRIMA FACIE COHERENT

Equal Pattern: simultaneous capture
(with an optionally tunable “bubble of simultaneity”)



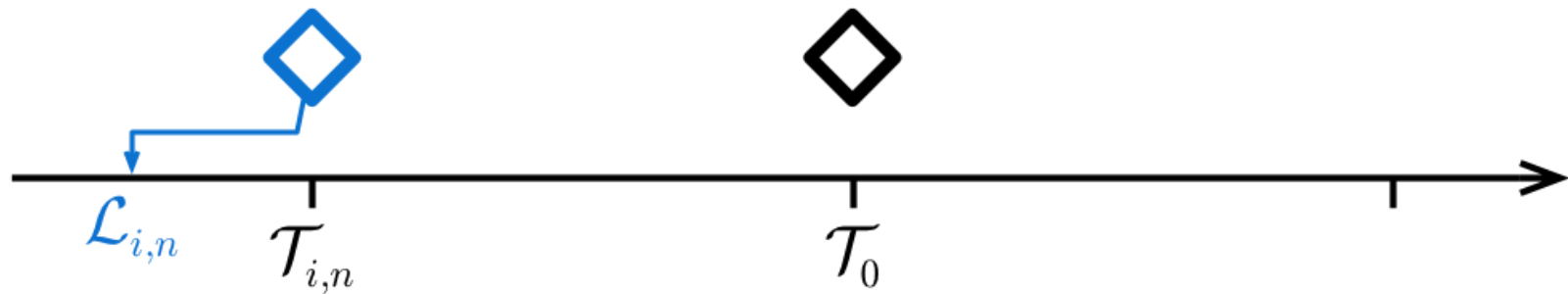
PRIMA FACIE VIOLATIVE

Closest memento created and acquired after the root was acquired



POSSIBLY COHERENT

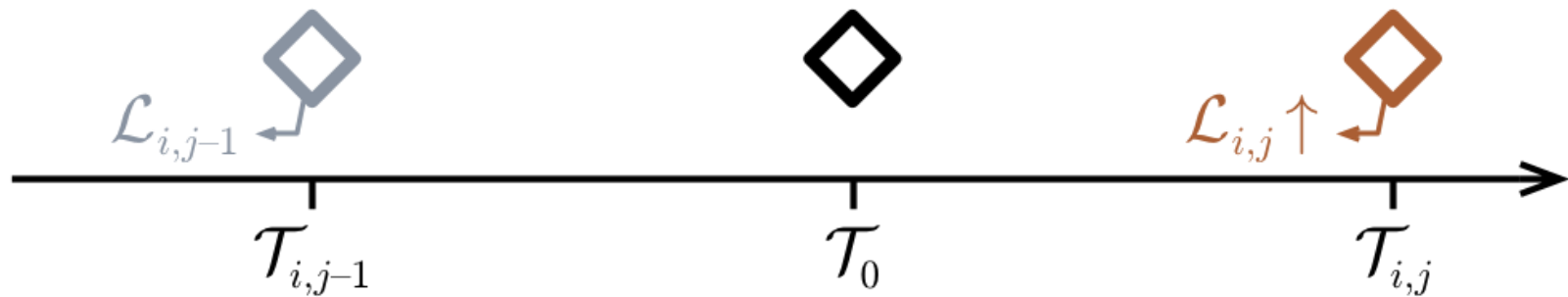
Closest (or only) memento captured before the root



PROBABLY VIOLATIVE

Closest (or only) memento captured after the root but no Last-Modified (possibly indicating a dynamically generated representations)

(for both PC & PV, you could do content comparison if there are 2 mementos that straddle the root page)



CONTENTS

- Motivation**
- Composite Memento**
- Coherence Framework**
- Temporal Coherence**
- Future Work**
- Conclusion**

TEMPORAL COHERENCE

web.archive.org/web/20050514013608/http://www.cs.odu.edu/

INTERNET ARCHIVE
WaybackMachine

692 captures

Old Dominion UNIVERSITY

College of Sciences

COMPUTER SCIENCE

Graduate

- Program Information
- Course Information

Undergraduate

- Program Information
- Course Information
- Advising Information

Student Resources

- Admissions
- Financial Aid
- Summer Schedule & Books
- Fall Schedule & Books
- Employment Opportunities
- Calendars

Directory

- Faculty
- Staff

Online Services

- Account Creation
- Squirrel Mail
- Exchange Mail
- Notes Mail
- Reserve Conference Room
- Leo Online

Featured Student

Joan A Smith
Ph.D Student
Awarded the Zipf Fellowship

Featured Alum

James . Cheng
BS in CS
Class of 1982

Upcoming Events

Computer Science Colloquium, Friday May 14 at 10:00AM
Speaker: Dr. Allen Fisher from Carnegie Mellon [\[More\]](#)

CS Dept News

Jessica Crouch will be joining the computer science faculty as an assistant professor in the 2005 fall semester.

About

- Welcome from the Chair
- Mission Statement
- Student Goals
- Bylaws
- Organization
- Facilities
- FAQ

TEMPORAL COHERENCE

2005-05-14

01:36:08

The screenshot shows the website for the Computer Science department at Old Dominion University. Annotations with orange boxes and arrows indicate time intervals between various elements:

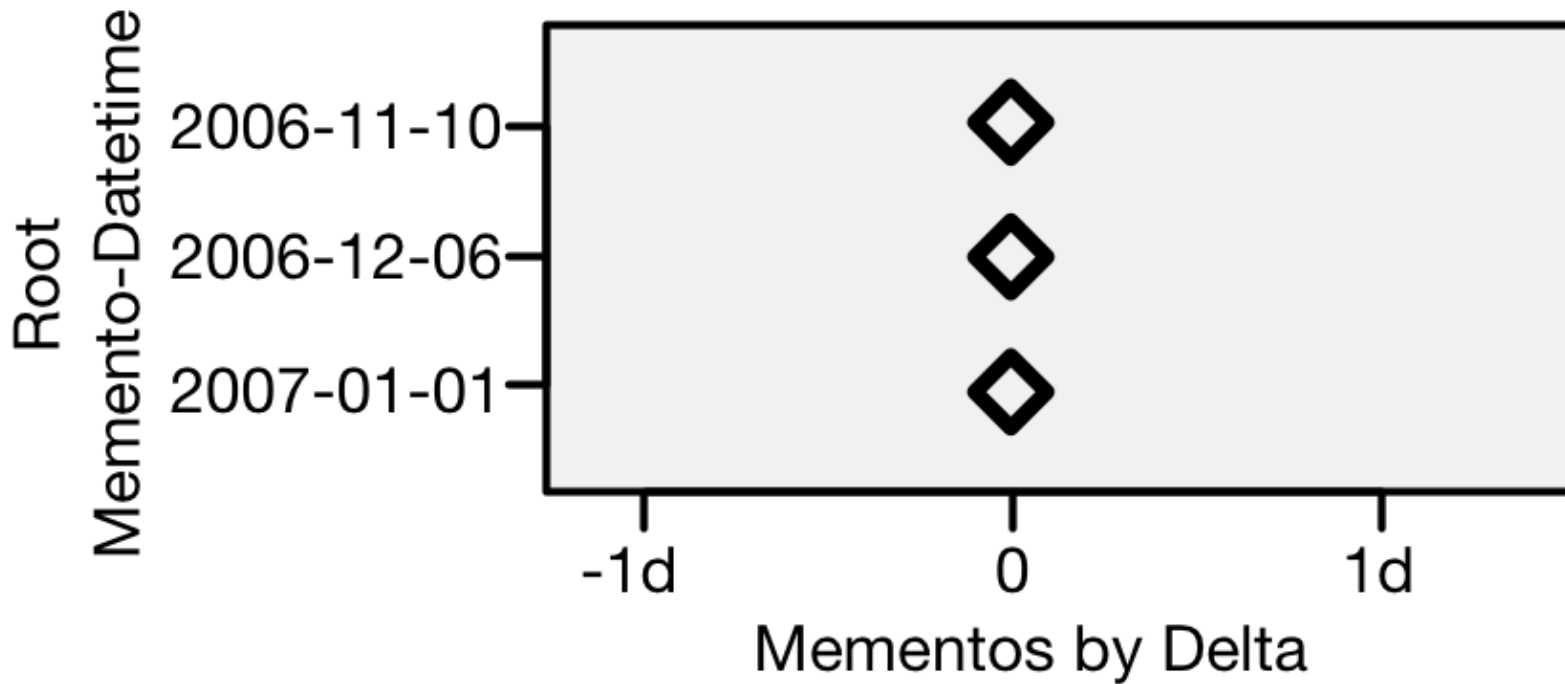
- +9 days**: Between the Wayback Machine search bar and the main navigation bar.
- +18 days**: Between the main navigation bar and the featured student section.
- +7 months**: Between the featured student section and the featured alum section.
- +18 days**: Between the featured alum section and the upcoming events section.
- +2.1 years**: Between the upcoming events section and the bottom of the page.

EMBEDDED RESOURCES

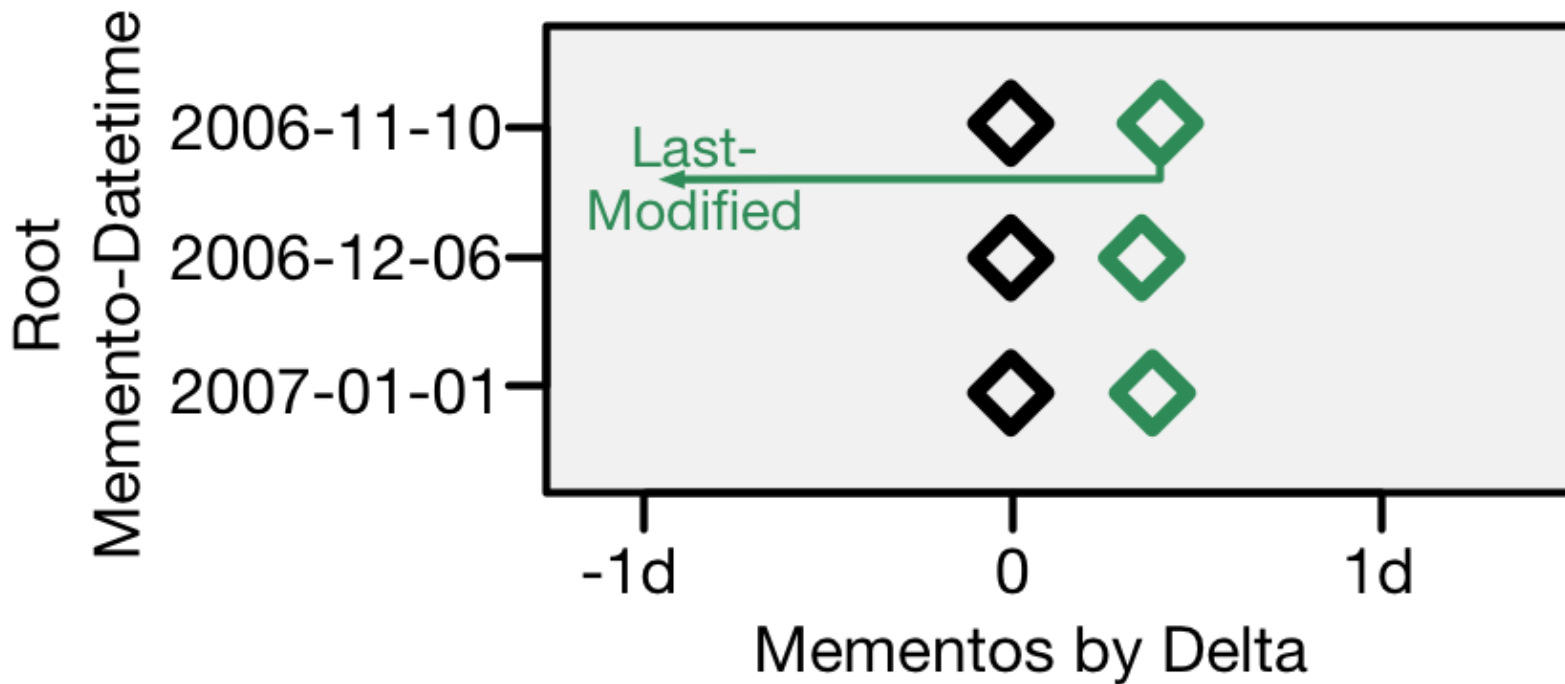
Resource	Memento-Datetime	Delta	Resource	Memento-Datetime	Delta
http://www.cs.odu.edu	2005-05-14 01:36:08		spacer.gif	2005-06-01 16:23:10	18.6 d
mm_menu.js	2005-05-23 02:39:12	9.0 d	jimcheng.gif	2005-06-01 16:37:39	18.6 d
style.css	2005-05-23 02:39:12	9.0 d	imgwith.gif	2005-06-01 16:58:50	18.6 d
gfx-logo-odu-crown.gif				06-01 21:21:45	18.8 d
ddmenu_ddown.js				2-21 20:14:25	7.3 mo
university.js				2-21 20:15:14	7.3 mo
rmenu_1st_about.png				2-21 21:01:12	7.3 mo
rmenu_bottom_229.gif				2-28 17:47:41	7.5 mo
shadow-bl.gif				2-28 19:43:48	7.5 mo
ecsbdg.jpg				2-28 19:54:29	7.5 mo
shadow-br.gif				06-12 02:36:07	2.1 years
gfx-btn-go-dblue.gif				06-21 02:35:17	2.1 years
shadow-tr.gif				Not Found	
header-right1.gif					

Embedded Resources	26
Mean Delta	125.9 days
Standard Deviation	207.7 days
Minimum Delta	9.0 days
Maximum Delta	2.1 years

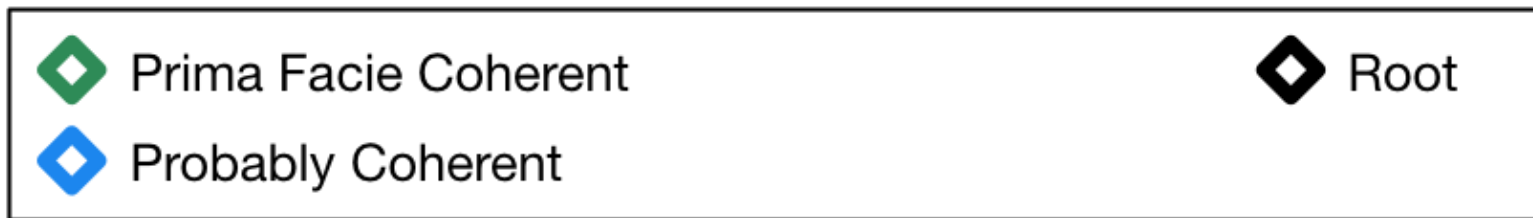
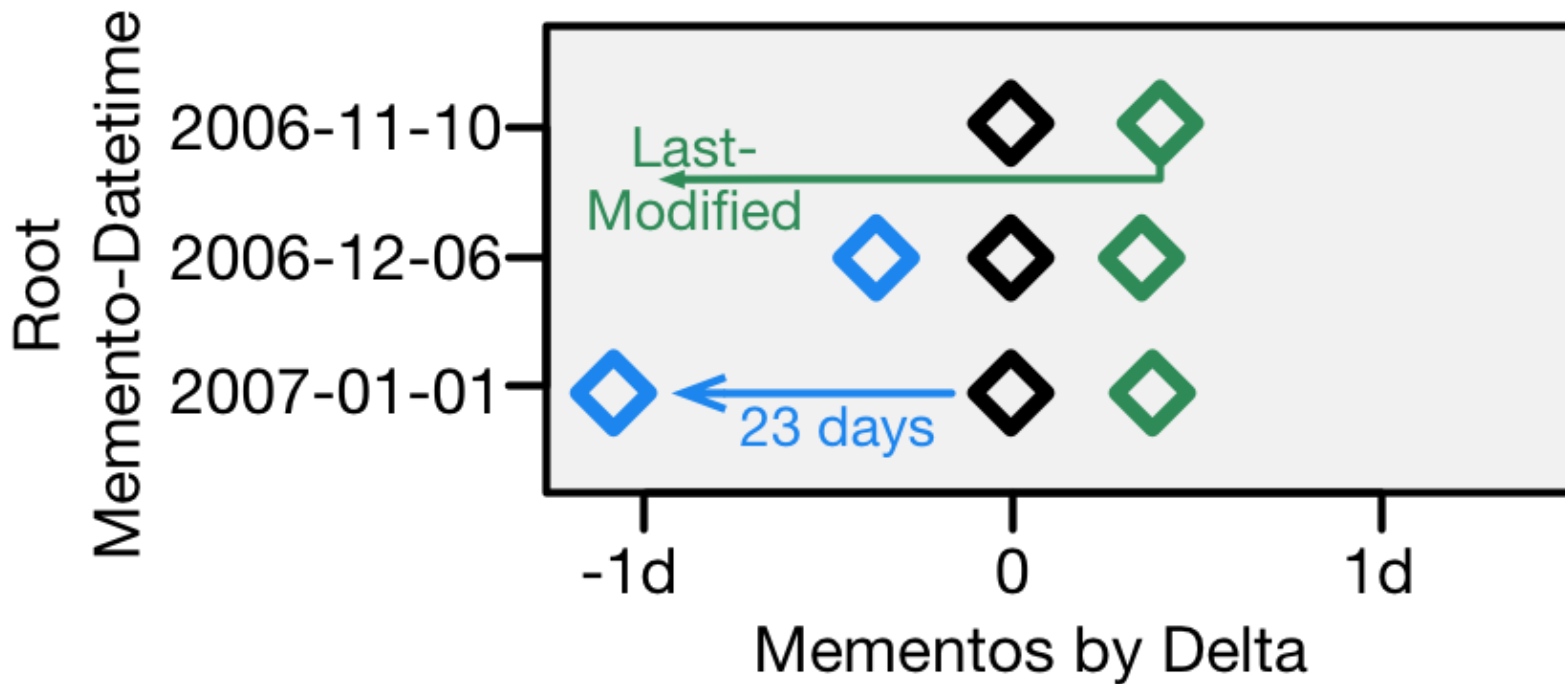
REPRESENTING COHERENCE



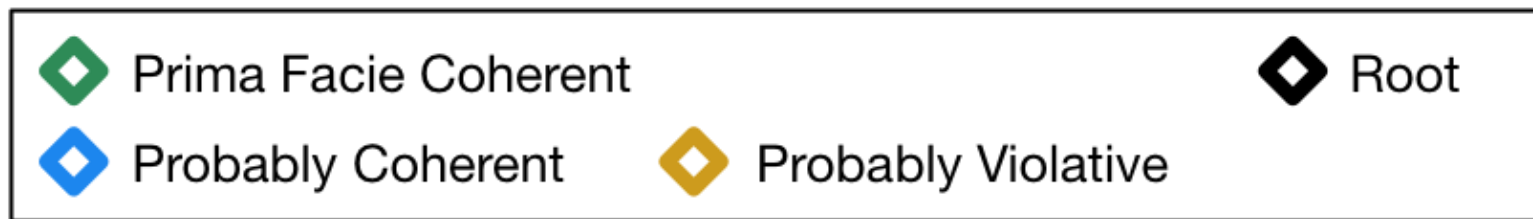
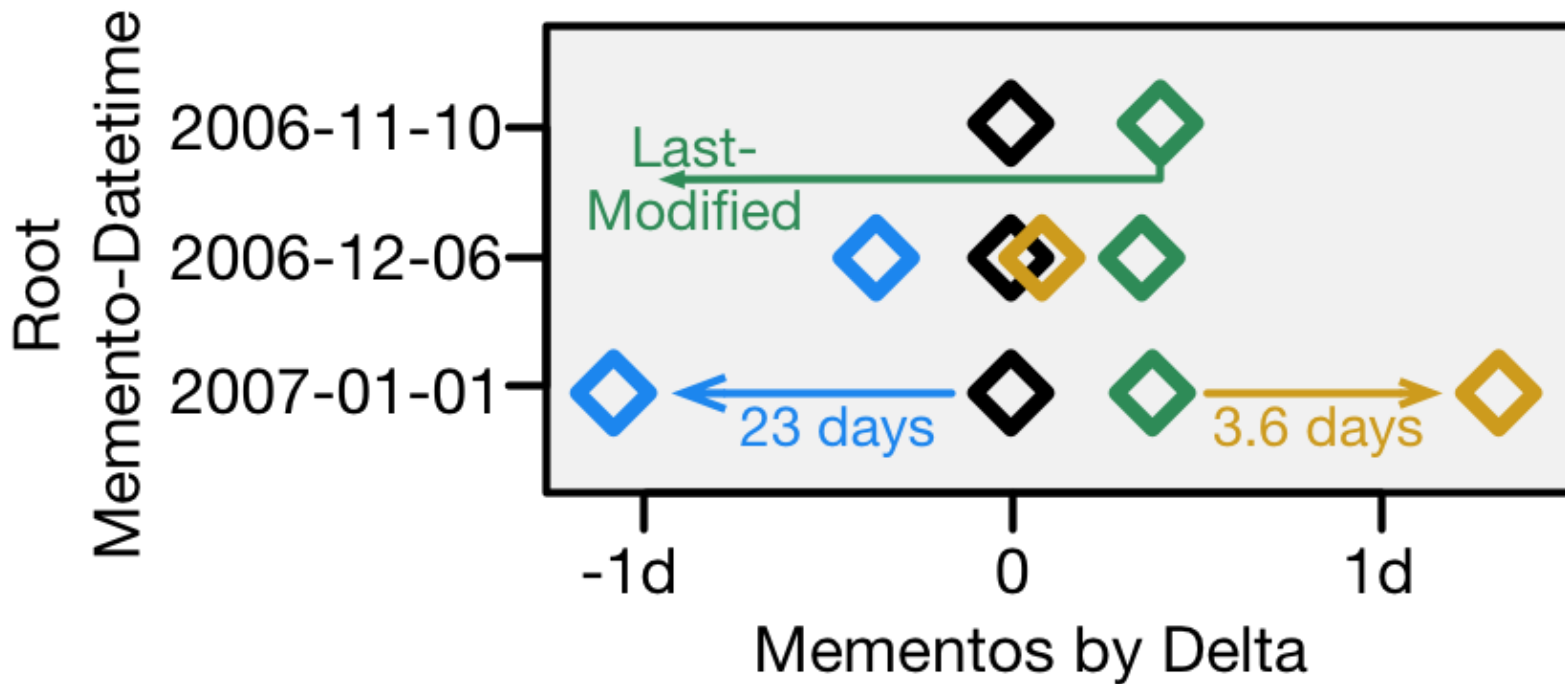
REPRESENTING COHERENCE



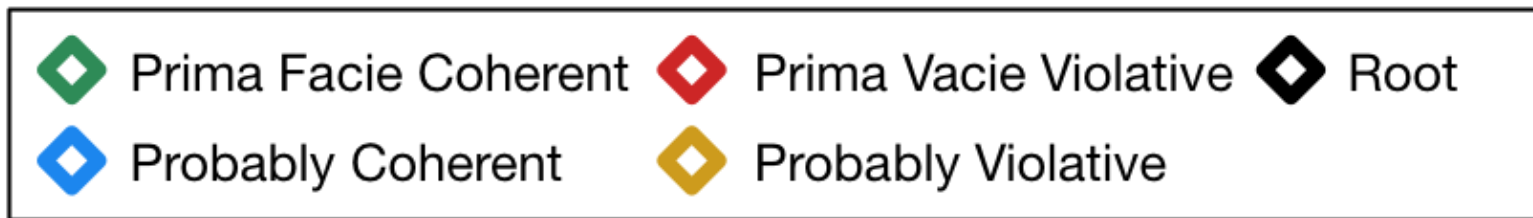
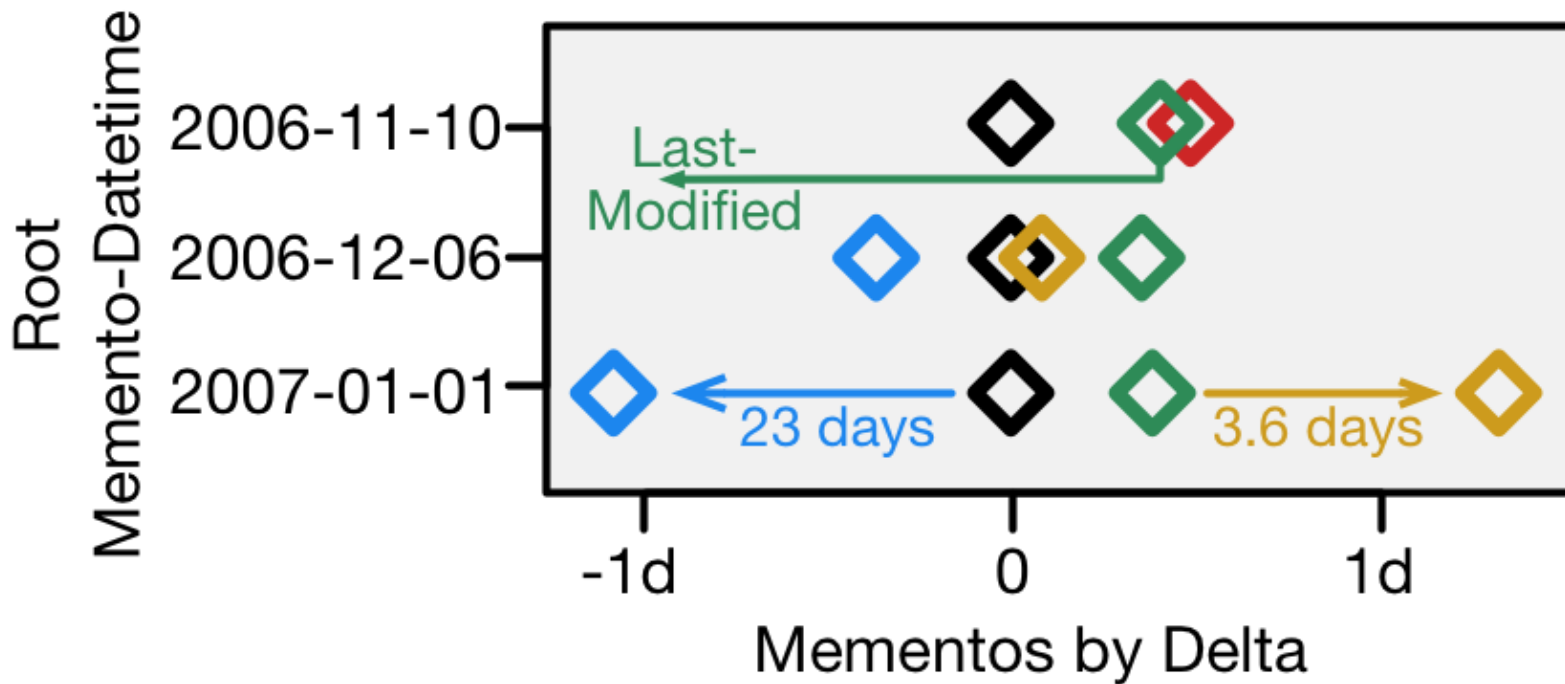
REPRESENTING COHERENCE



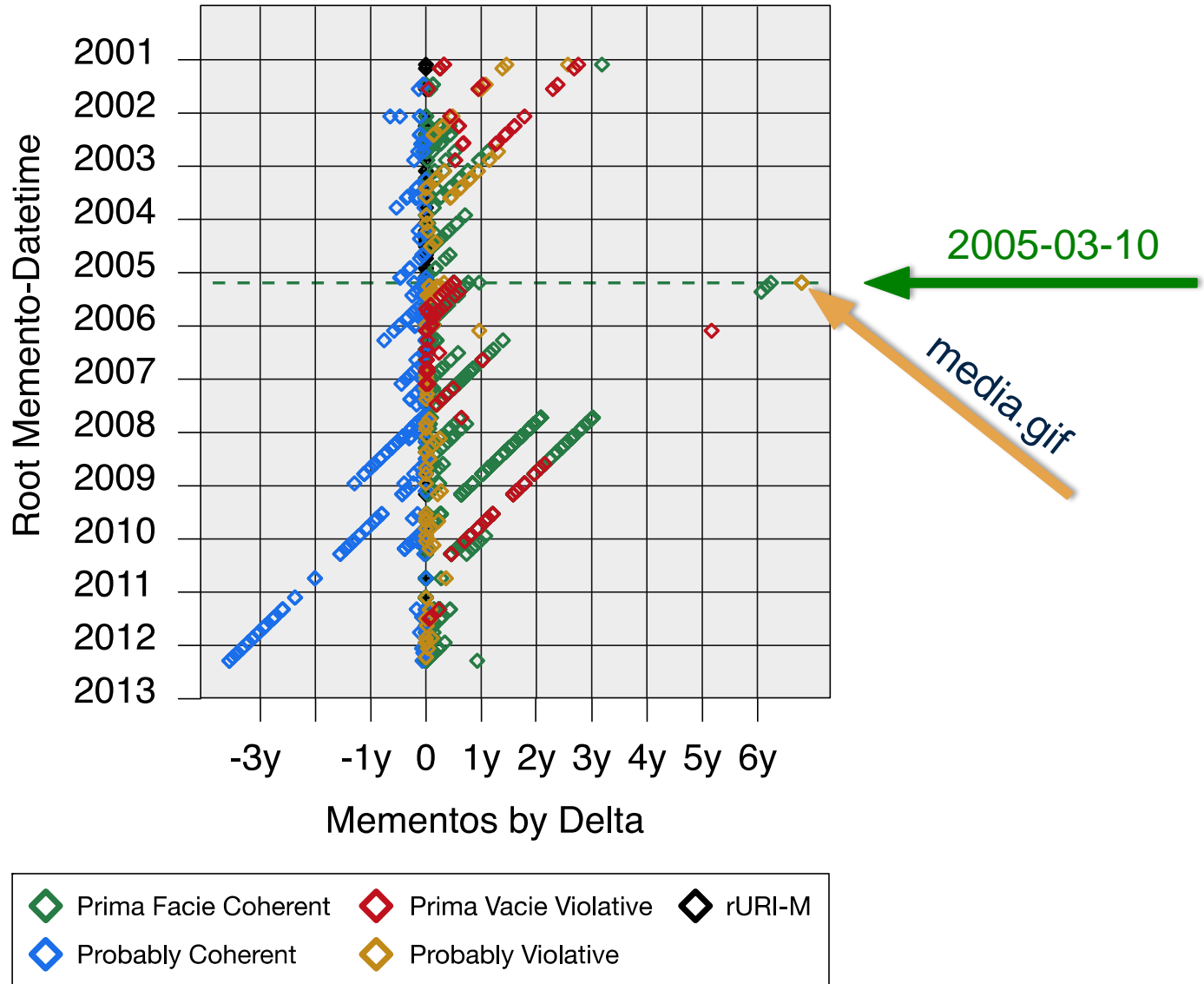
REPRESENTING COHERENCE



REPRESENTING COHERENCE



THE FULL CHART



EXPERIMENT: DATA SET

- 4,000 sample URI-Rs (data set from JCDL 2011)
- Single and Multiple Archives
- Two Heuristics:
 - Minimum distance (current default Wayback behavior)
 - choose closest Memento-Datetime
 - Bracket (proposed here)
 - use combination of Memento-Datetime + Last-Modified
- Download all TimeMaps
- Download all root mementos
- Download all embedded resources

EXPERIMENT: SAMPLING

- For each root URI-R TimeMap, choose a single memento per month
- Extract embedded URI-Rs
- Download TimeMaps for embedded URI-Rs
- Download heuristically best URI-Ms
- Repeat recursively

ROOT URI-R STATISTICS

Archival Data

Root URI-Rs archived	2,756 • 68.9%
In multiple archives	1,180 • 29.5%
Mean archives per URI-R	1.58
Mean mementos per URI-R	124.57

URI-M Status

200 OK	82,425 • 93.6%
503 Service Unavailable	4,444 • 5.0%
404 Not found	583 • 0.7%
403 Forbidden	388 • 0.4%
Others	214 • 0.3%

EMBEDDED URI-R STATISTICS

Archival Data

Embedded URI-Rs	1,623,127
per root URI-M	19.7
Embedded URI-Ms available	1,332,993 • 93.6%
per root URI-M	15.1

URI-M Failure Reasons

Not archived	312,641 • 83.9%
404 Not found	44,852 • 12.0%
403 Forbidden	6,116 • 1.6%
503 Service Unavailable	5,442 • 1.5%
Others	3,508 • 0.9%

COMPOSITE MEMENTO (ROOT) COMPLETENESS & COHERENCE

Completeness (and Missing)

Description	MinDist Single	MinDist Multi	Bracket Single	Bracket Multi
Mean Complete	76.1%	80.2%	76.2%	80.3%
Mean Missing	23.9%	19.8%	23.8%	19.7%

Coherence

Multiple archives: +completeness, -coherence?

Description	MinDist Single	MinDist Multi	Bracket Single	Bracket Multi
Mean Prima Facie Coherent	41.0%	40.9%	54.7%	54.6%
Mean Possibly Coherent	27.3%	28.7%	12.8%	14.2%
Mean Probably Violative	2.5%	5.3%	2.5%	5.3%
Mean Prima Facie Violative	5.3%	5.3%	6.2%	6.2%

At least 5% of pages can be shown to have temporal violations!

EMBEDDED MEMENTO COHERENCE

Description	MinDist Single	MinDist Multi	Bracket Single	Bracket Multi
Prima Facie Coherent	622,565	621,447	864,736	859,625
Possibly Coherent	497,405	466,046	244,104	215,585
Probably Violative	104,376	53,734	104,339	53,694
Prima Facie Violative	100,760	103,662	114,062	117,469
Totals	1,325,106	1,244,889	1,327,241	1,246,373

Description	MinDist Single	MinDist Multi	Bracket Single	Bracket Multi
Prima Facie Coherent	47.0%	49.9%	65.2%	69.0%
Possibly Coherent	37.5%	37.4%	18.4%	17.3%
Probably Violative	7.9%	4.3%	7.9%	4.3%
Prima Facie Violative	7.6%	8.3%	8.6%	9.4%

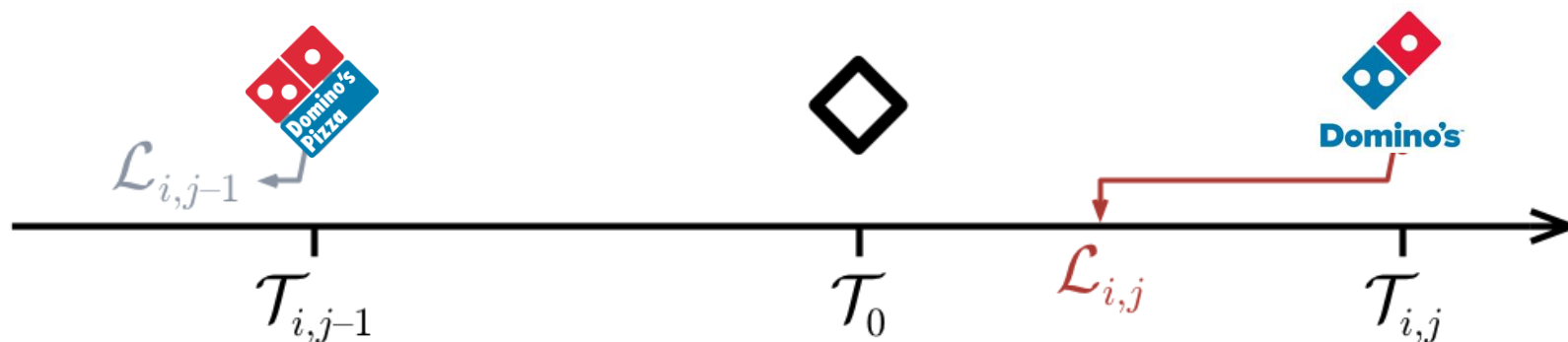
At least 7% of embedded resources are used violatively!

CONTENTS

- Motivation**
- Related work**
- Preliminary work**
- Temporal Coherence**
- Future work**
- Conclusion**

MINOR OR MAJOR VIOLATIONS?

- This is a temporal violation. But is it meaningful?



- How to judge?
 - Most archives transform HTML
 - Not all archives support export of original file
 - How to measure similarity on binary files?
 - early results: very few cases of equivalent binaries

HOW TO CONVEY COHERENCE?

Time Travel [About API Privacy Terms](#)

http://stanford.edu

2010-04-28 10:34:32 Find Reconstruct 2010-04-27 16:17:59 GMT

4 months before 12 days before a day before 3 hours before 17 minutes before 2 minutes before 9 seconds before 2010-04-27 10:34:32

wayback.archive-it.org (39)
web.archive.org (2)
wayback.vodafone.is (3)

The page below assembled using **44 Mementos** from **3 archives**, spanning a **year**

STANFORD UNIVERSITY

Web People Maps A-Z Index Search...

About Stanford Admission Academics Research Life On Campus

Show Expanded Menus

Stanford scenes

EVENTS

APR 27 Panel: SEC v. Goldman Sachs 5 p.m.
Hal Holbrook: Mark Twain Tonight! 8 p.m.

APR 28 Wellness Fair 10 a.m.

APR 30 Conference: The Future of Journalism

UNIVERSITY NEWS

Admit Weekend 2010
Campus rolls out the Cardinal carpet for more than 1,300 prospective freshmen and their families.

Ash and aircraft
Danger to airliners from volcano's debris is difficult to assess, Stanford expert says.

Sun in high definition
Recently launched Stanford telescope sends crystal-clear

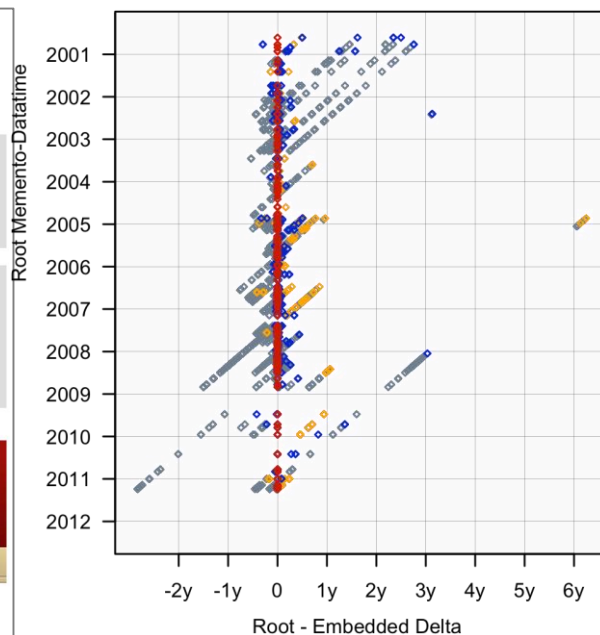
GATEWAYS FOR...

- Students
- Faculty & Staff
- Alumni
- Parents
- Visitors & Neighbors

TOP DESTINATIONS

SCHOOLS

- Business
- Earth Sciences
- Education



How to convey coherence & contributing archive?

How to scale to > 100 embedded mementos?

POLICIES & HEURISTICS

- **Tradeoffs:**
 - Fast: minimize distance
 - Accurate: maximize coherence
 - Complete: query all (not just top k) archives in order to maximize completeness

CONTENTS

- Motivation**
- Composite Memento**
- Coherence Framework**
- Temporal Coherence**
- Future Work**
- Conclusion**

CONCLUSION

- **Defined four classes of temporal coherence for describing relationship between root & embedded mementos**
 - Prima Facie {Coherent|Violative}
 - Possibly Coherent / Probably Violative
- **Determine classes using a combination of HTTP metadata, primarily Memento-Datetime & Last-Modified**
- **At least 5% of IA pages have 1 or more temporal violations**
- **Using multiple archives increases completeness, but with a possible loss of coherence**
- **Determining semantic impact of violations and UI issues (status, policy choices) are areas of future research**