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HE WENT TO VIEW AN ARCHIVED PAGE.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

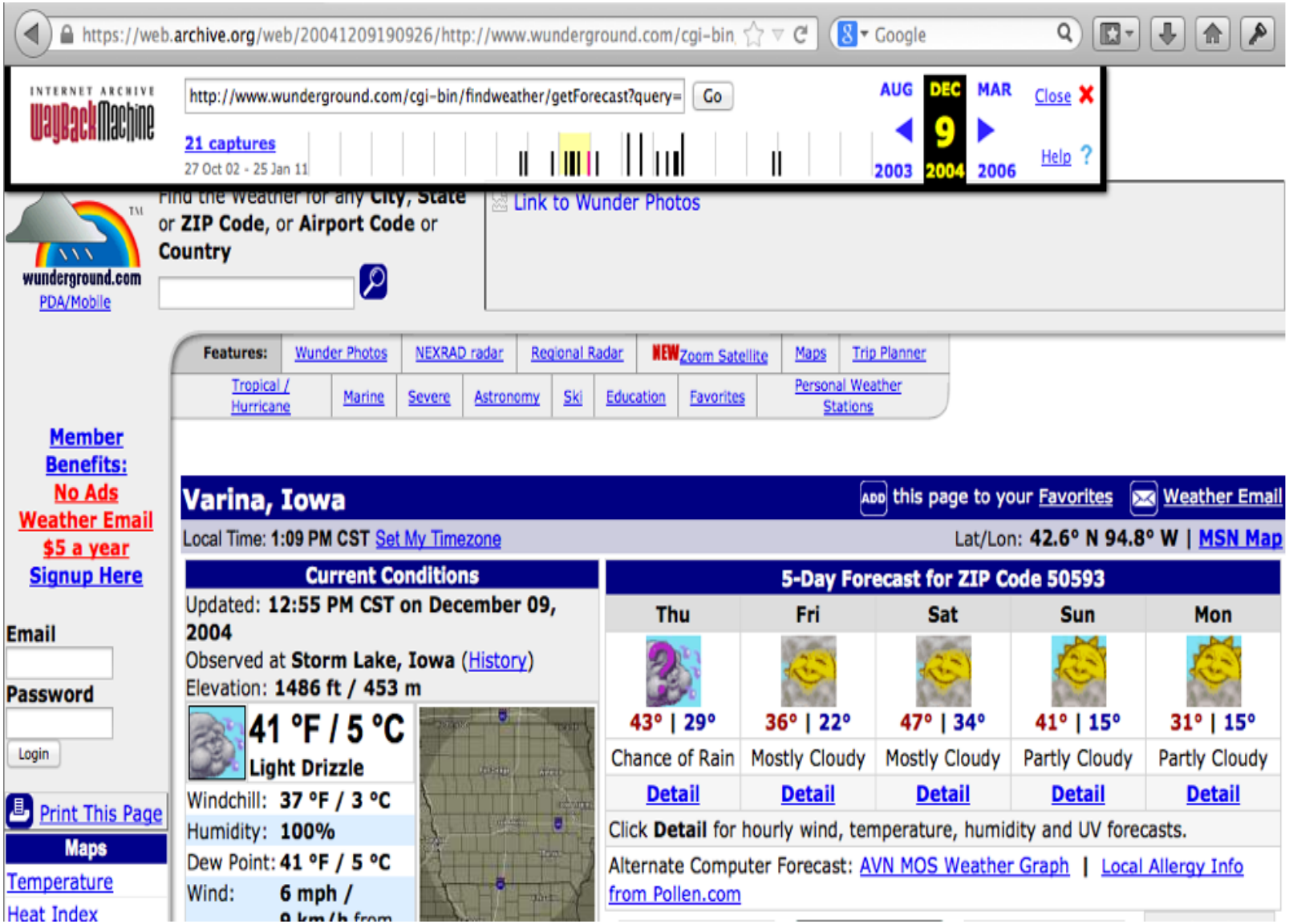
- **How to crawl a site to maximize coherence**
  - Ben Saad et al., JCDL 2011, TPDL 2011
- **Detecting, visualizing temporal defects**
  - Spaniol et al., WICOW 2009, IWAW 2009, VLDB 2009

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

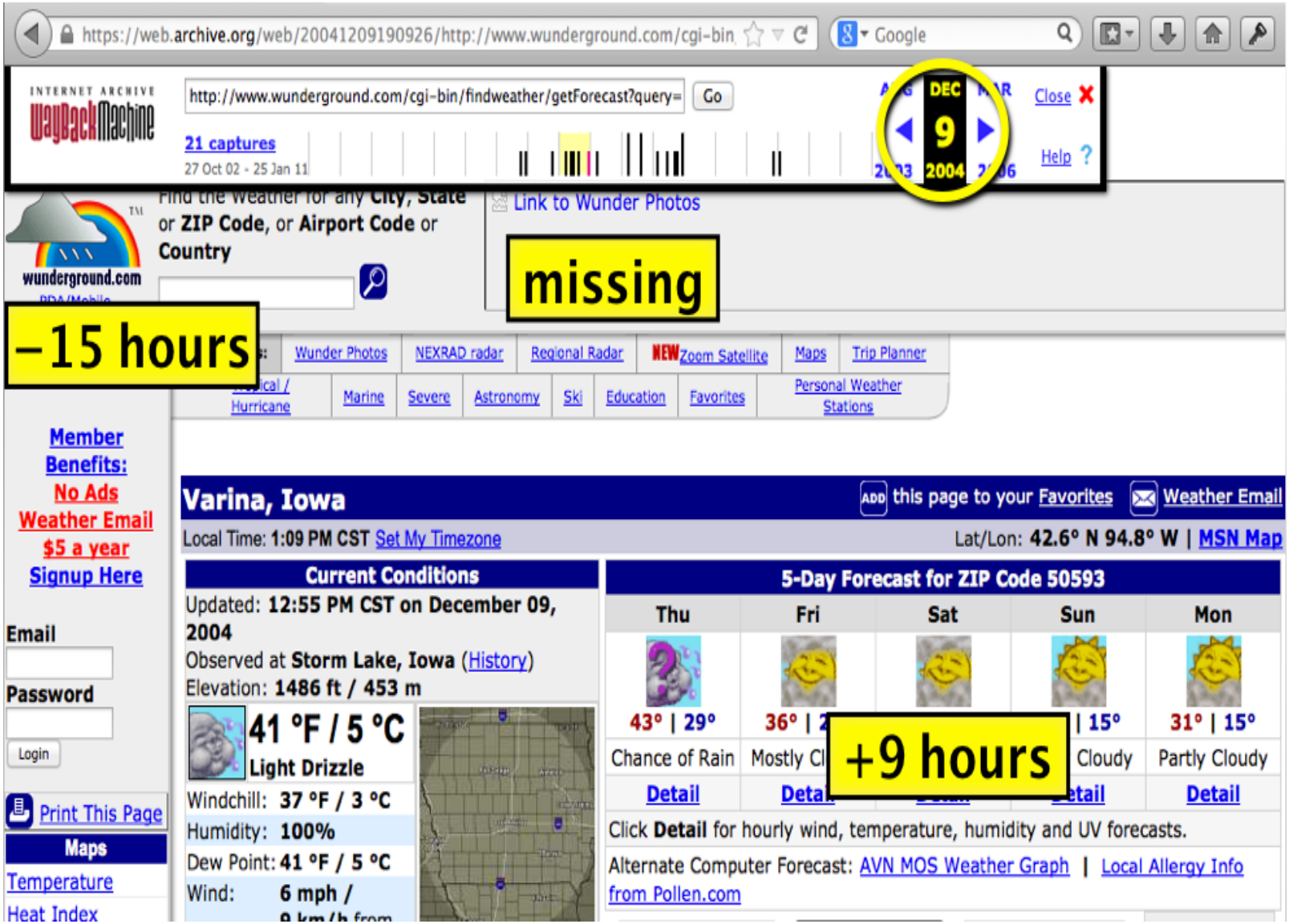
- ***Are the present resources correct?***

AS PRESENTED BY IA



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



CLEAR OR CLOUDY?

QUESTIONS

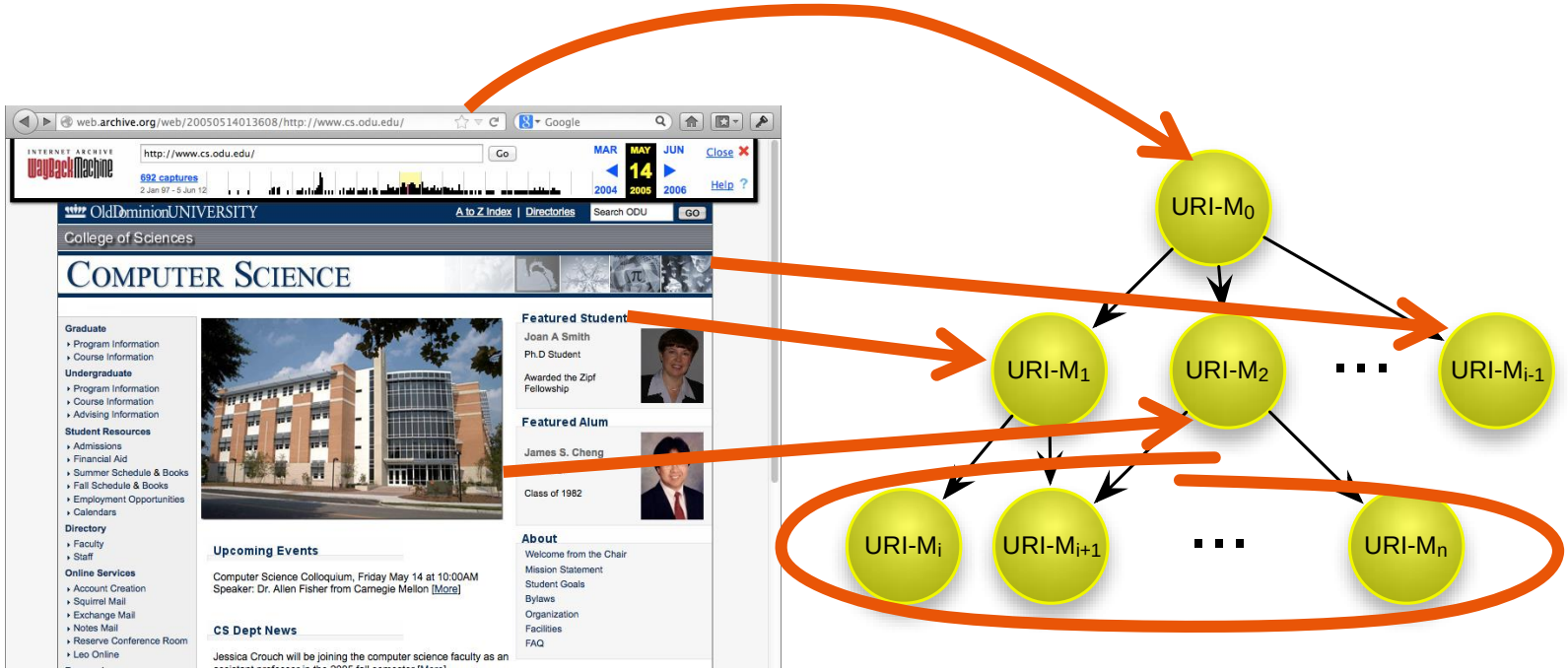
- **How prevalent is temporal incoherence?**
- **Can temporal coherence be improved by using multiple archives?**
- **Can temporal coherence be improved by introducing memento selection heuristics?**

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>
<img src="foo.jpeg">
</html>
```

WITH THESE RESPONSE HEADERS

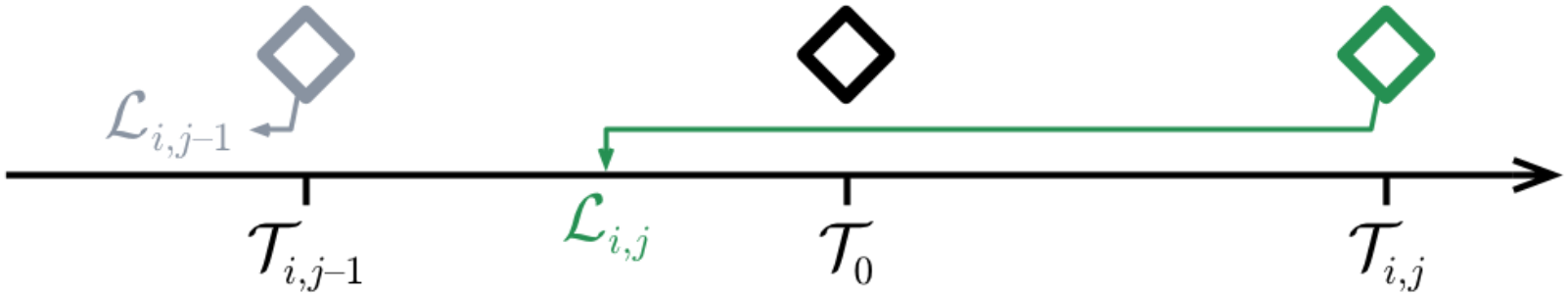
```
HTTP/1.1 200 OK
Server: Tengine/2.0.3
Date: Mon, 27 Apr 2015 22:03:32 GMT
Content-Type: image/jpeg
Content-Length: 15632
Connection: keep-alive
Memento-Datetime: Tue, 07 Feb 2006 00:58:23 GMT
Link: <Memento links deleted...>
X-Archive-Orig-server: Apache/1.3.26 (Unix) ApacheJServ/1.1.2 PHP/4.3.4
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

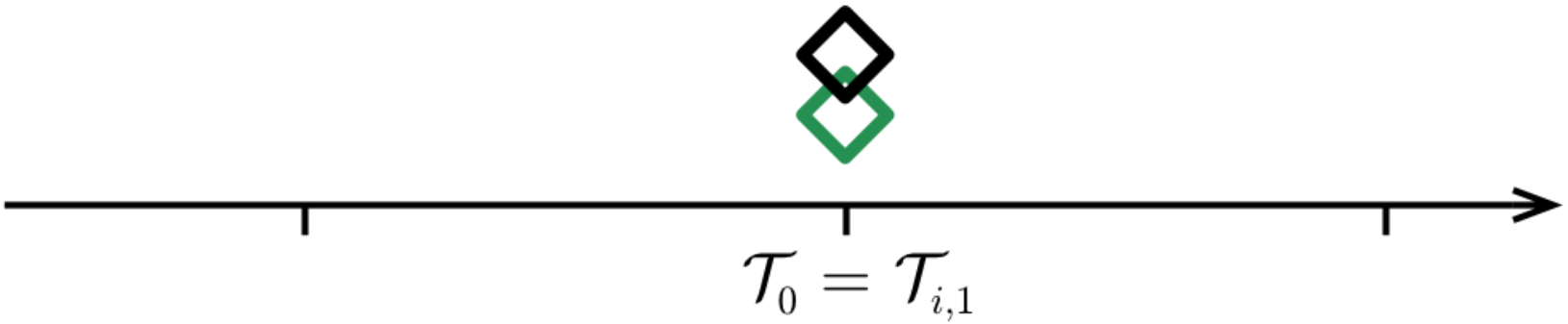
(yes, Last-Modified is sometimes wrong, but many of those cases can be detected)



PRIMA FACIE COHERENT

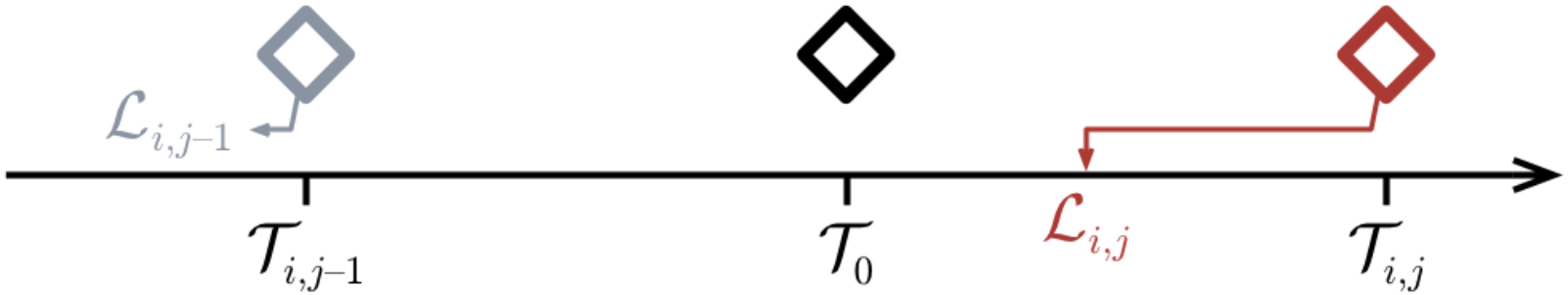
Equal Pattern: simultaneous capture

(with an optionally tunable “bubble of simultaneity”)



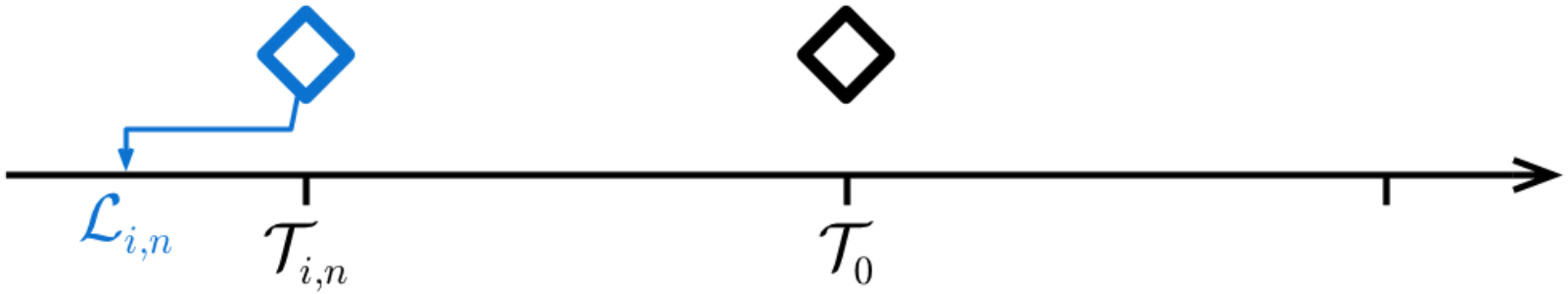
PRIMA FACIE VIOLATIVE

Closest memento created and acquired after the root was acquired



POSSIBLY COHERENT

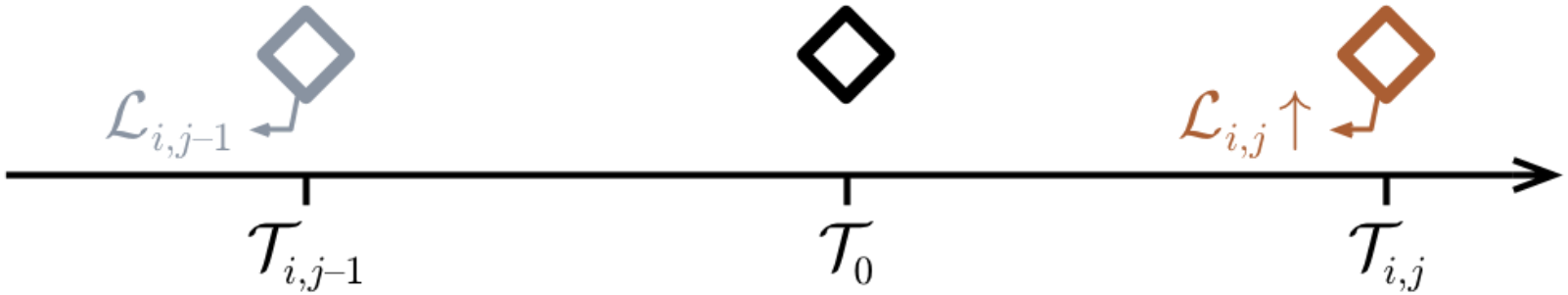
Closest (or only) memento captured before the root



PROBABLY VIOLATIVE

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

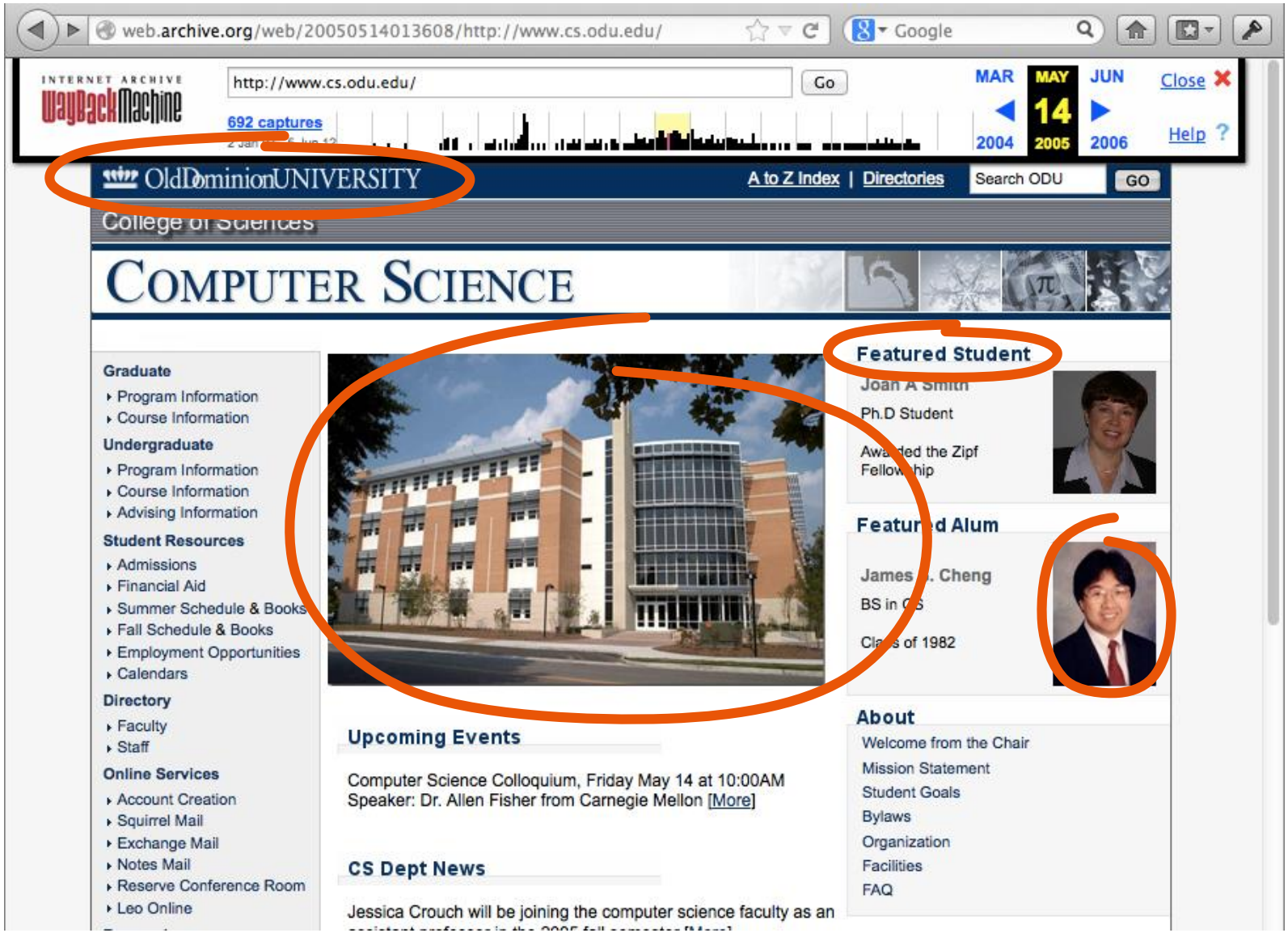
(for both PC & PV, you could do content comparison if there are 2 mementos that straddle the root page)



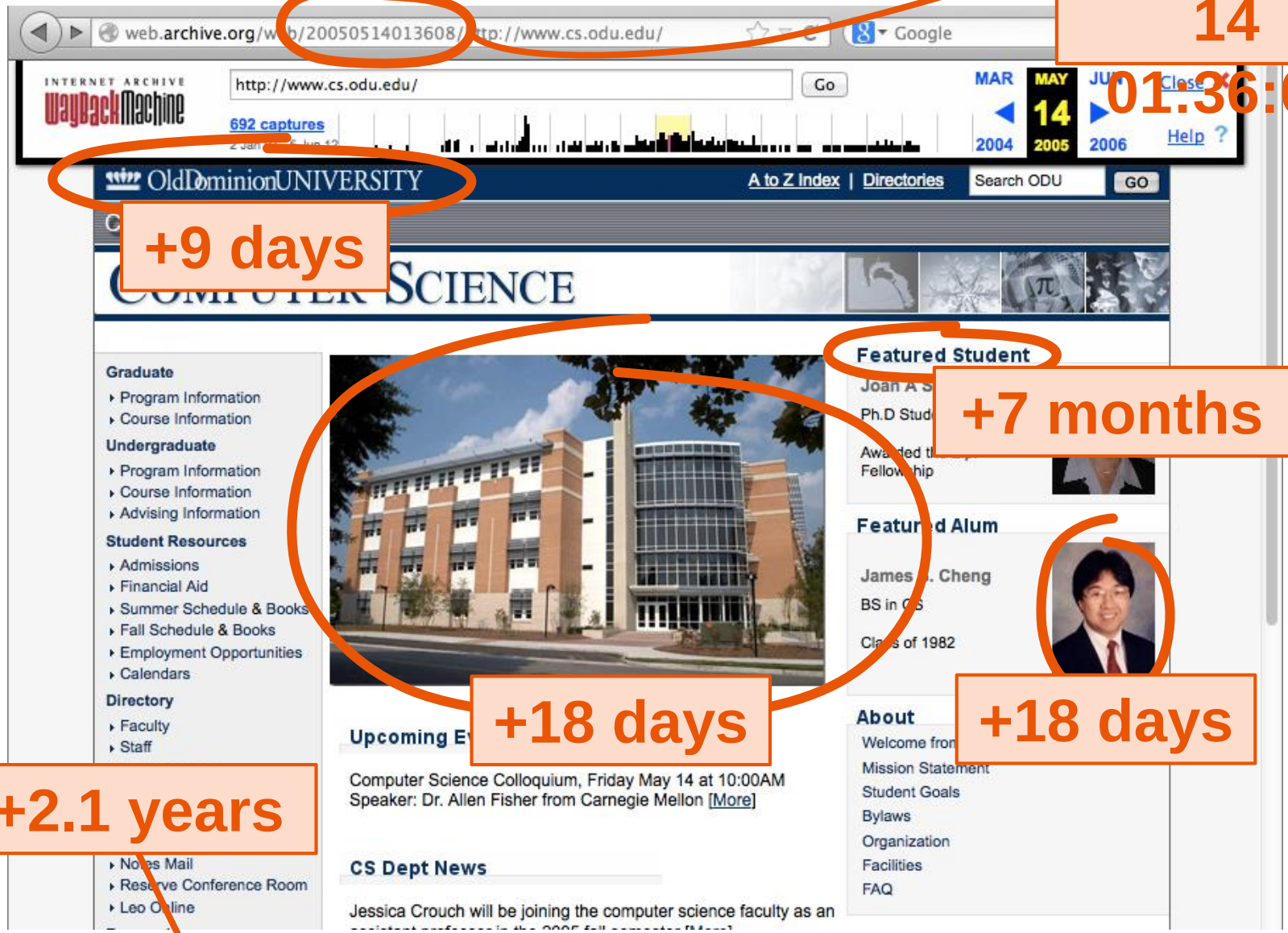
CONTENTS

- ☑ Motivation
- ☑ Composite Memento
- ☑ Coherence Framework
- ☐ Temporal Coherence
- ☐ Future Work
- ☐ Conclusion

TEMPORAL COHERENCE



TEMPORAL COHERENCE



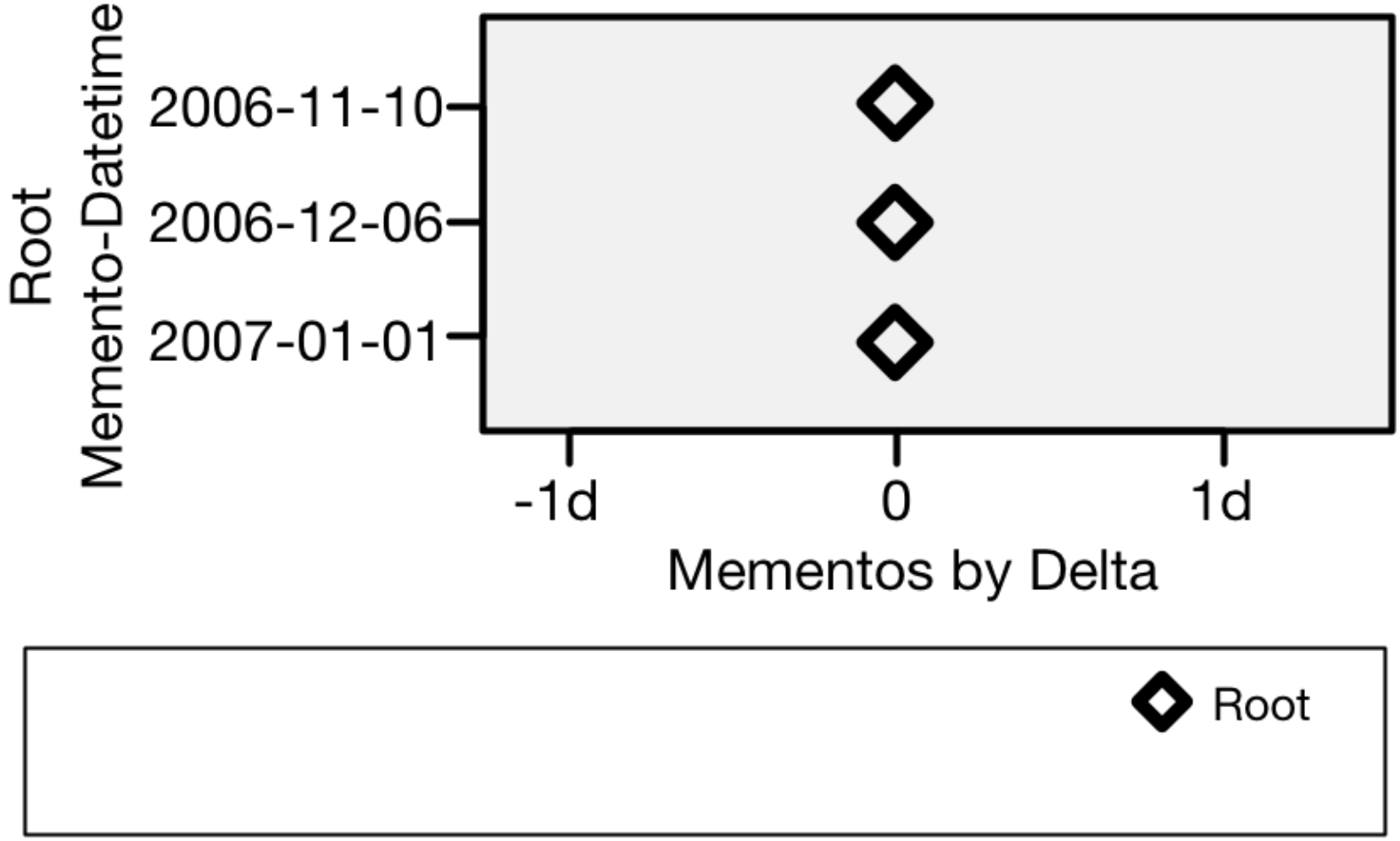
EMBEDDED RESOURCES

Resource	Memento-Datetime	Delta
http://www.cs.odu.edu	2005-05-14 01:36:08	
mm_menu.js	2005-05-23 02:39:12	9.0 d
style.css		
gfx-logo-odu-crown.gif		
ddmenu_ddown.js		
university.js		
rmenu_1st_about.png		
rmenu_bottom_229.gif		
shadow-bl.gif		
ecsbdg.jpg		
shadow-br.gif		
gfx-btn-go-dblue.gif		
shadow-tr.gif		
header-right1.gif		

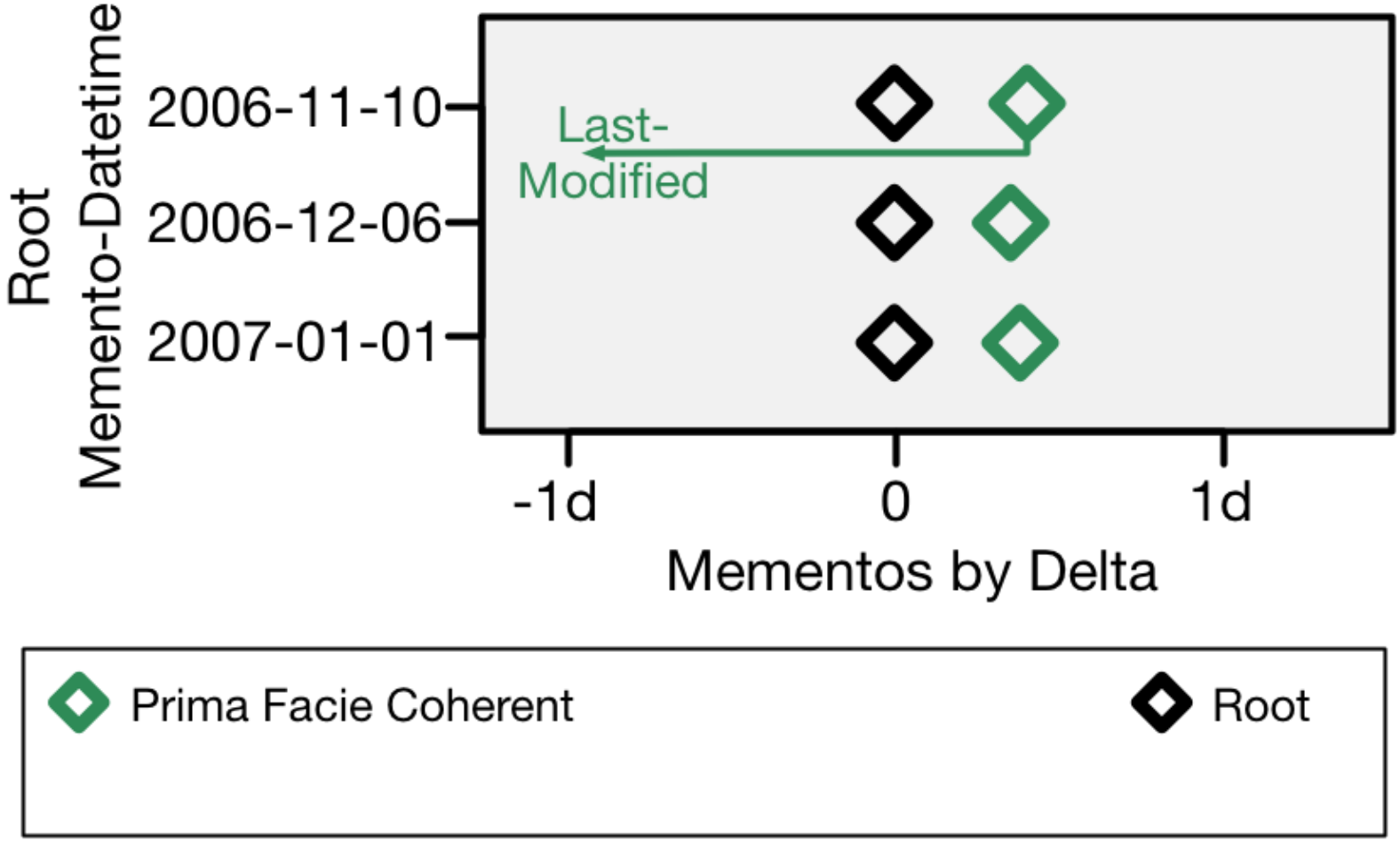
Resource	Memento-Datetime	Delta
spacer.gif	2005-06-01 16:23:10	18.6 d
jimcheng.gif	2005-06-01 16:37:39	18.6 d
	06-01 16:58:50	18.6 d
	06-01 21:21:45	18.8 d
	12-21 20:14:25	7.3 mo
	12-21 20:15:14	7.3 mo
	12-21 21:01:12	7.3 mo
	12-28 17:47:41	7.5 mo
	12-28 19:43:48	7.5 mo
	12-28 19:54:29	7.5 mo
	06-12 02:36:07	2.1 years
	06-21 02:35:17	2.1 years
	Not Found	

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

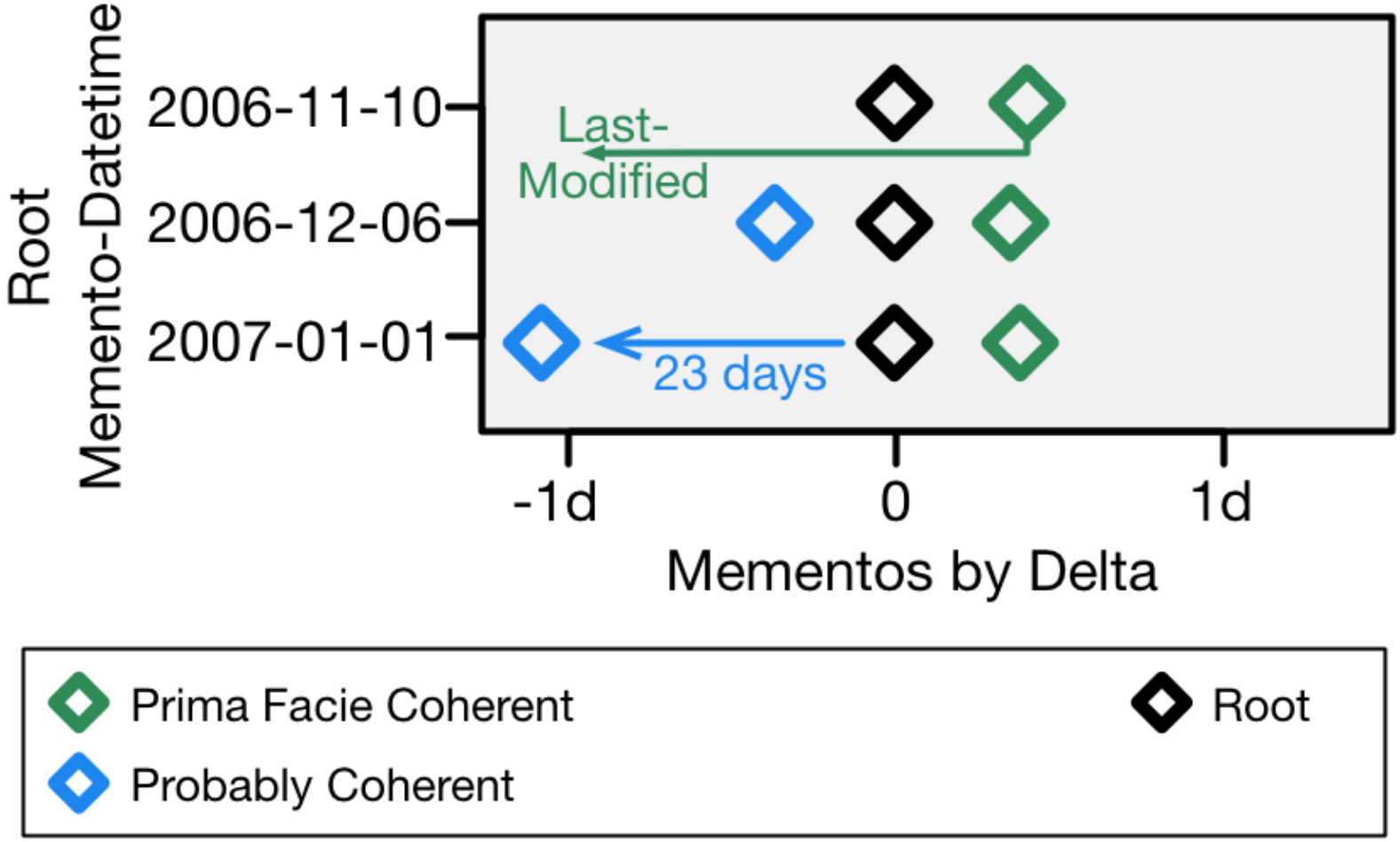
REPRESENTING COHERENCE



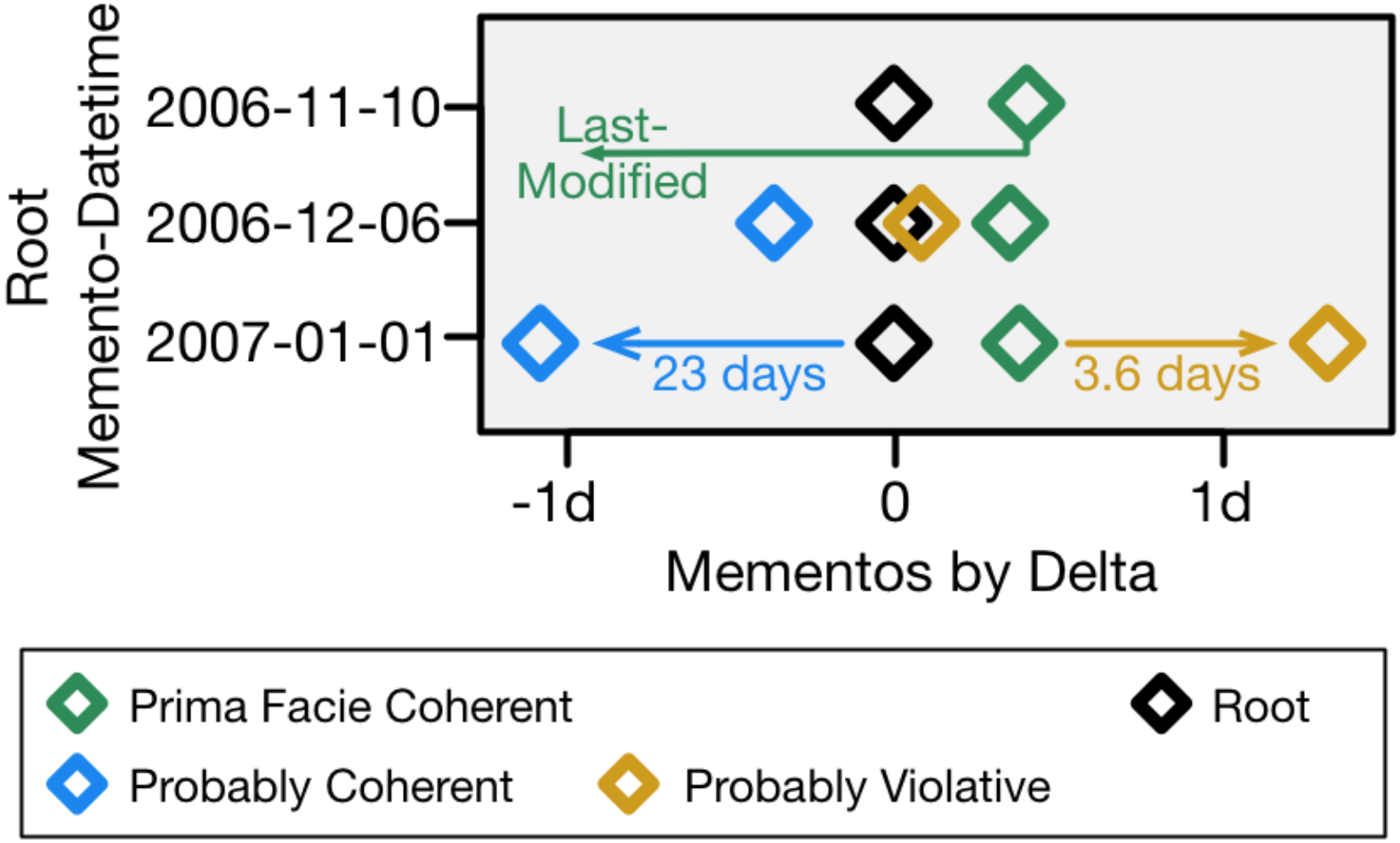
REPRESENTING COHERENCE



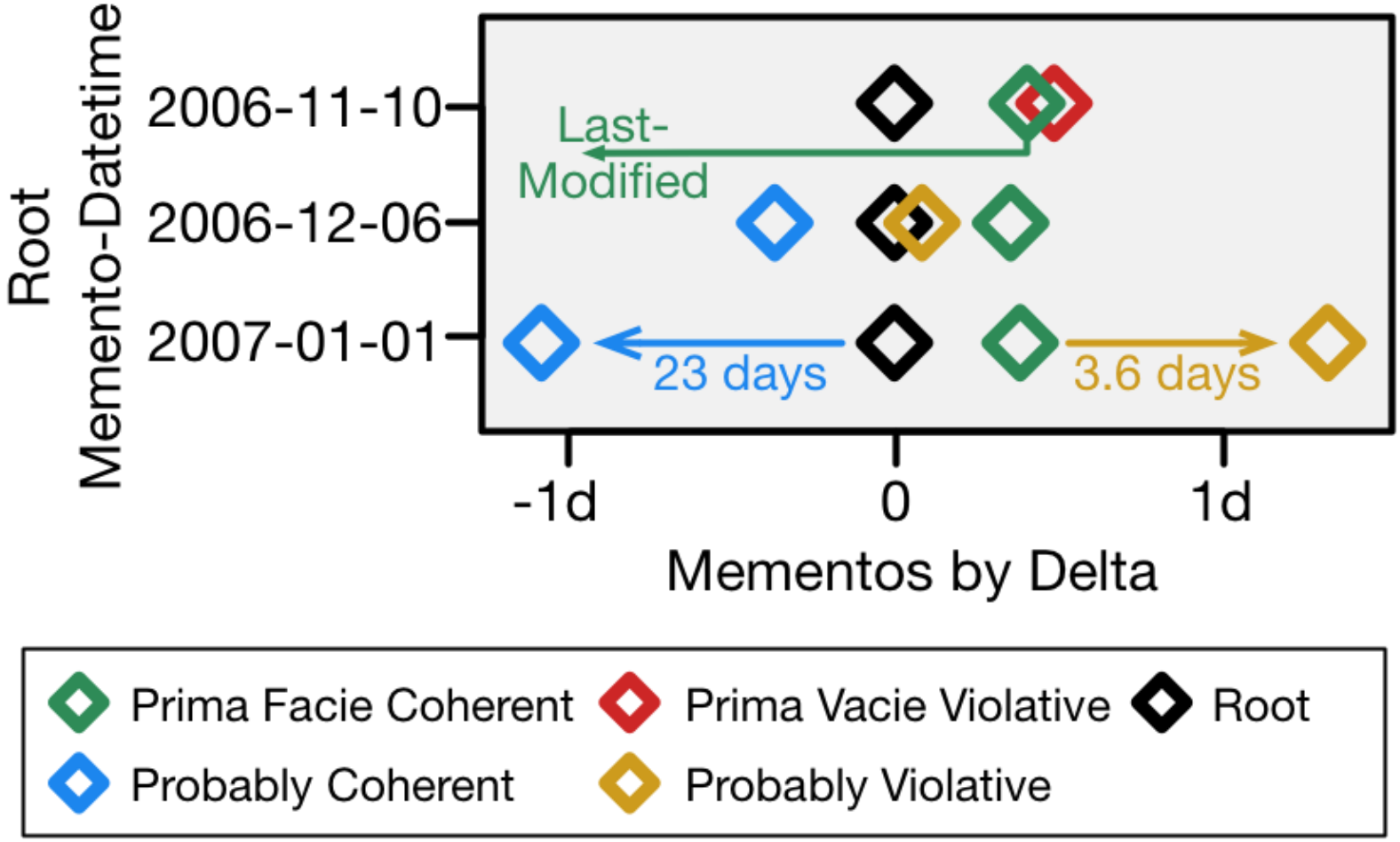
REPRESENTING COHERENCE



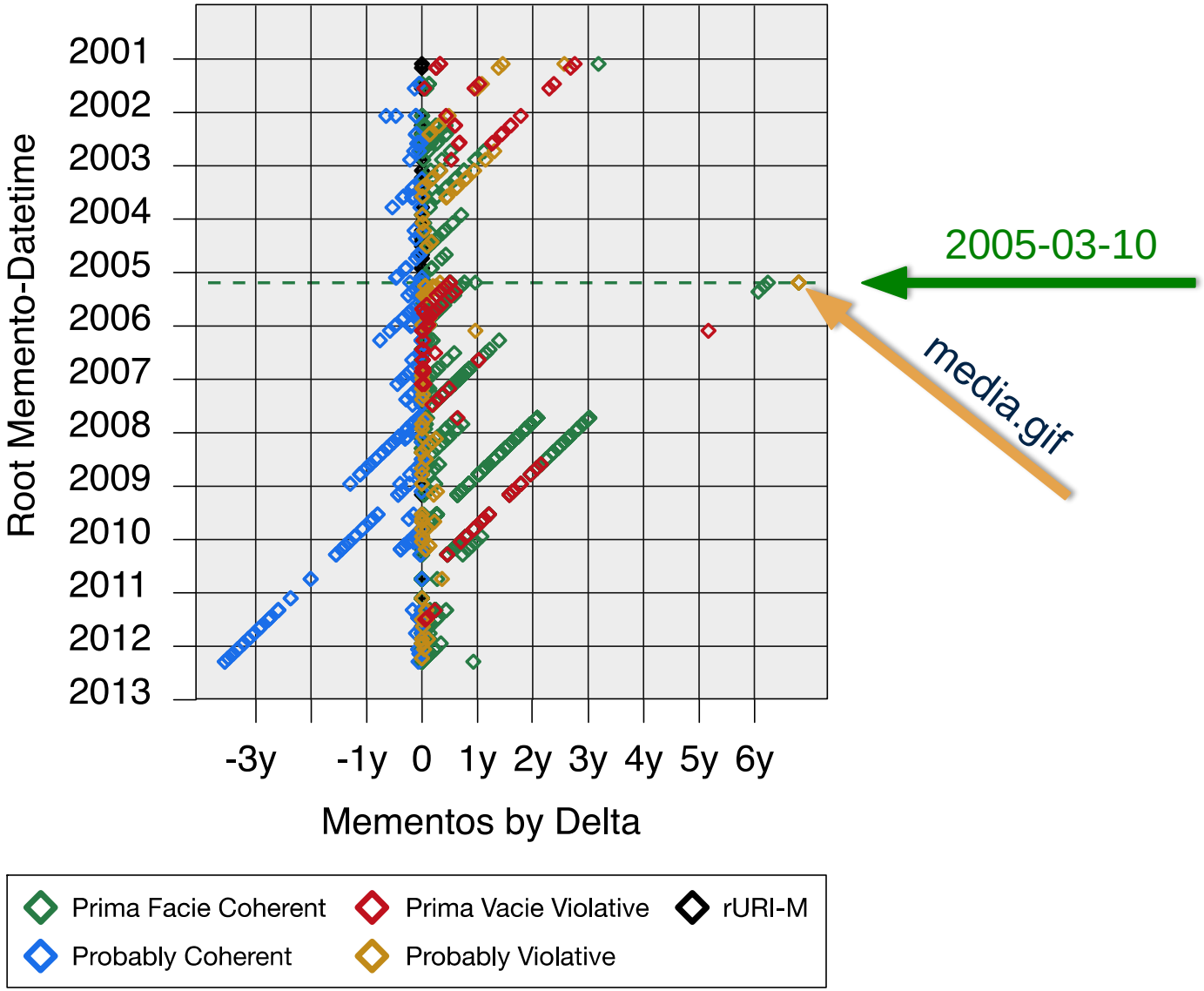
REPRESENTING COHERENCE



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%

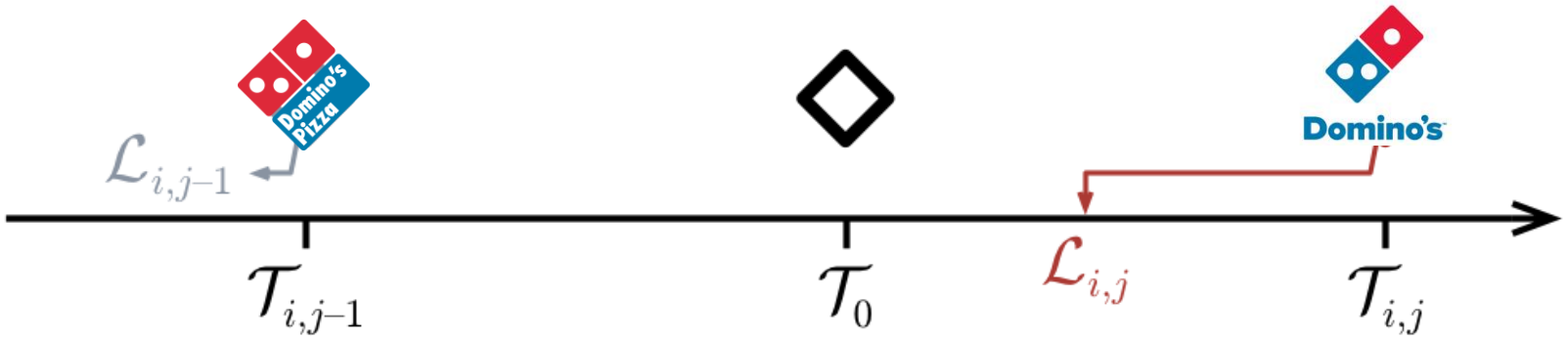
At least 7% of embedded resources are used violatively!

CONTENTS

- ☑ Motivation
- ☑ Related work
- ☑ Preliminary work
- ☑ Temporal Coherence
- ☐ Future work
- ☐ Conclusion

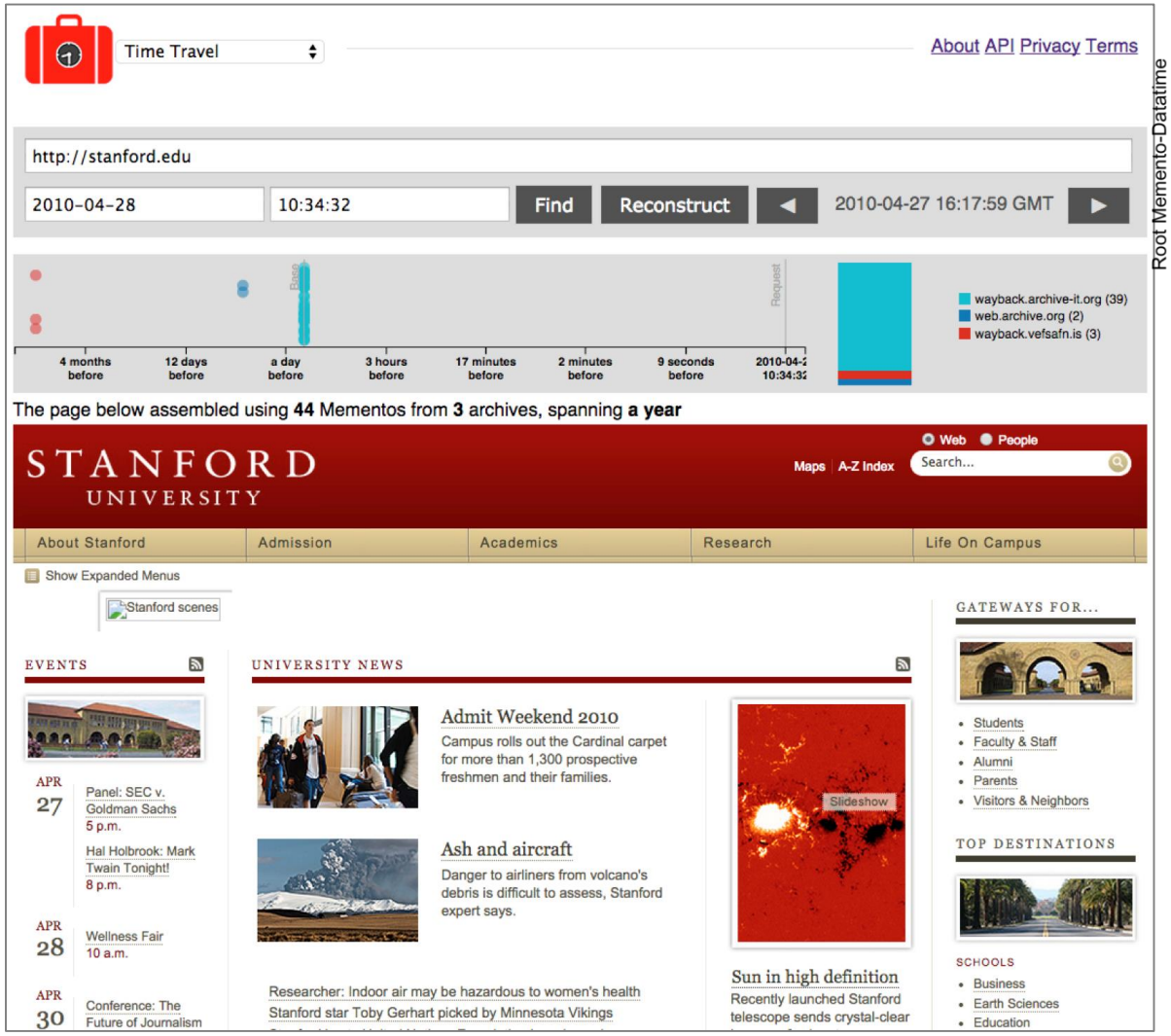
MINOR OR MAJOR VIOLATIONS?

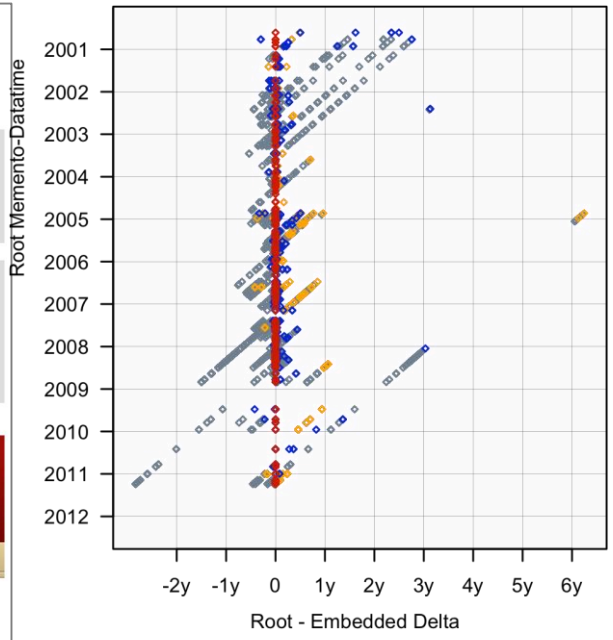
- **This is a temporal violation. But is it meaningful?**



- **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?





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**