Smart Routing of memento Requests

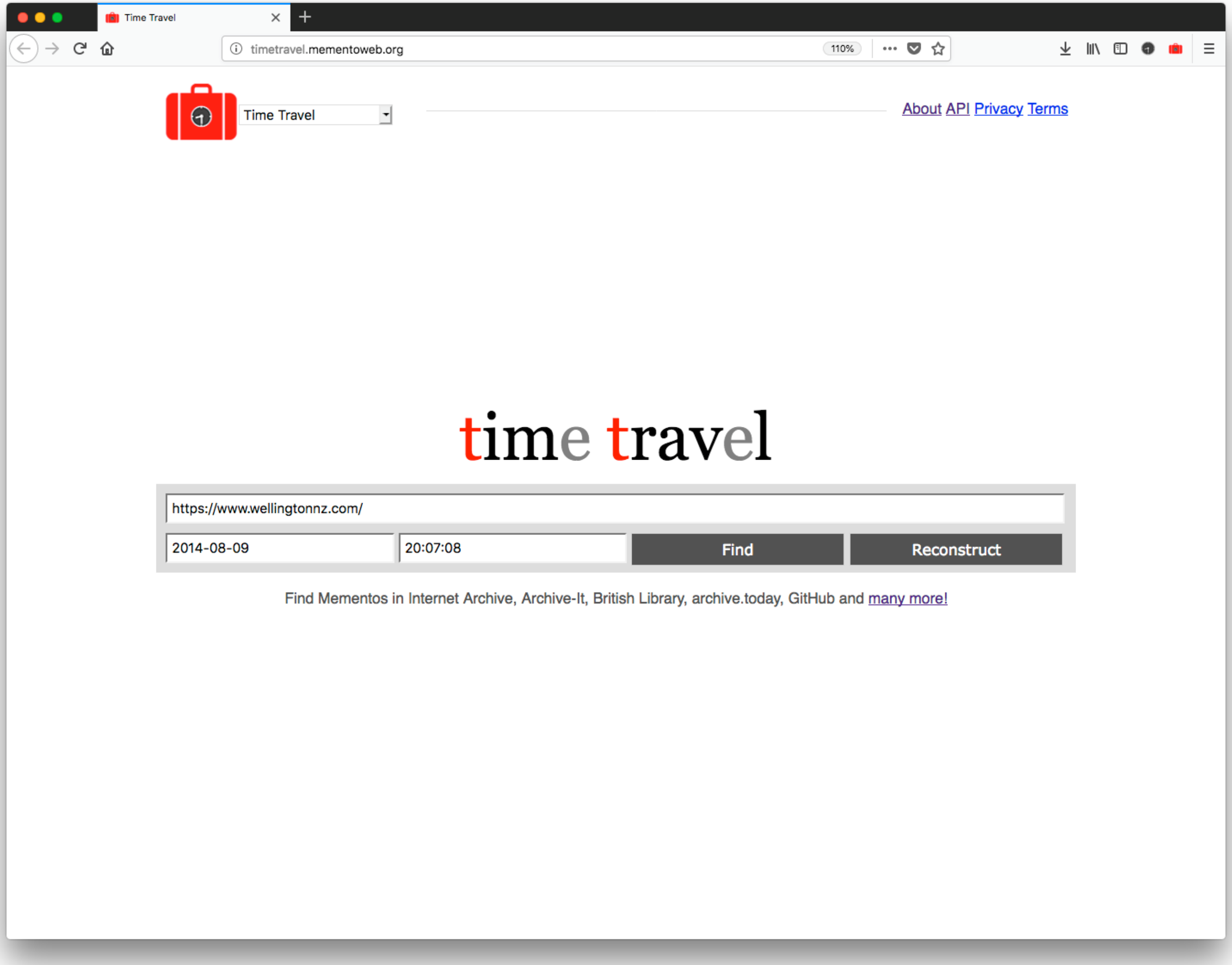
Martin Klein[1]
Lyudmila Balakireva[1]
Harihar Shankar[1]
James Powell[1]
Herbert Van de Sompel[2]

[1]Research Library
Los Alamos National Laboratory

[2]Data Archiving and Networked Services
The Netherlands

Los Alamos
NATIONAL LABORATORY

Memento



https://www.wellingtonnz.com/

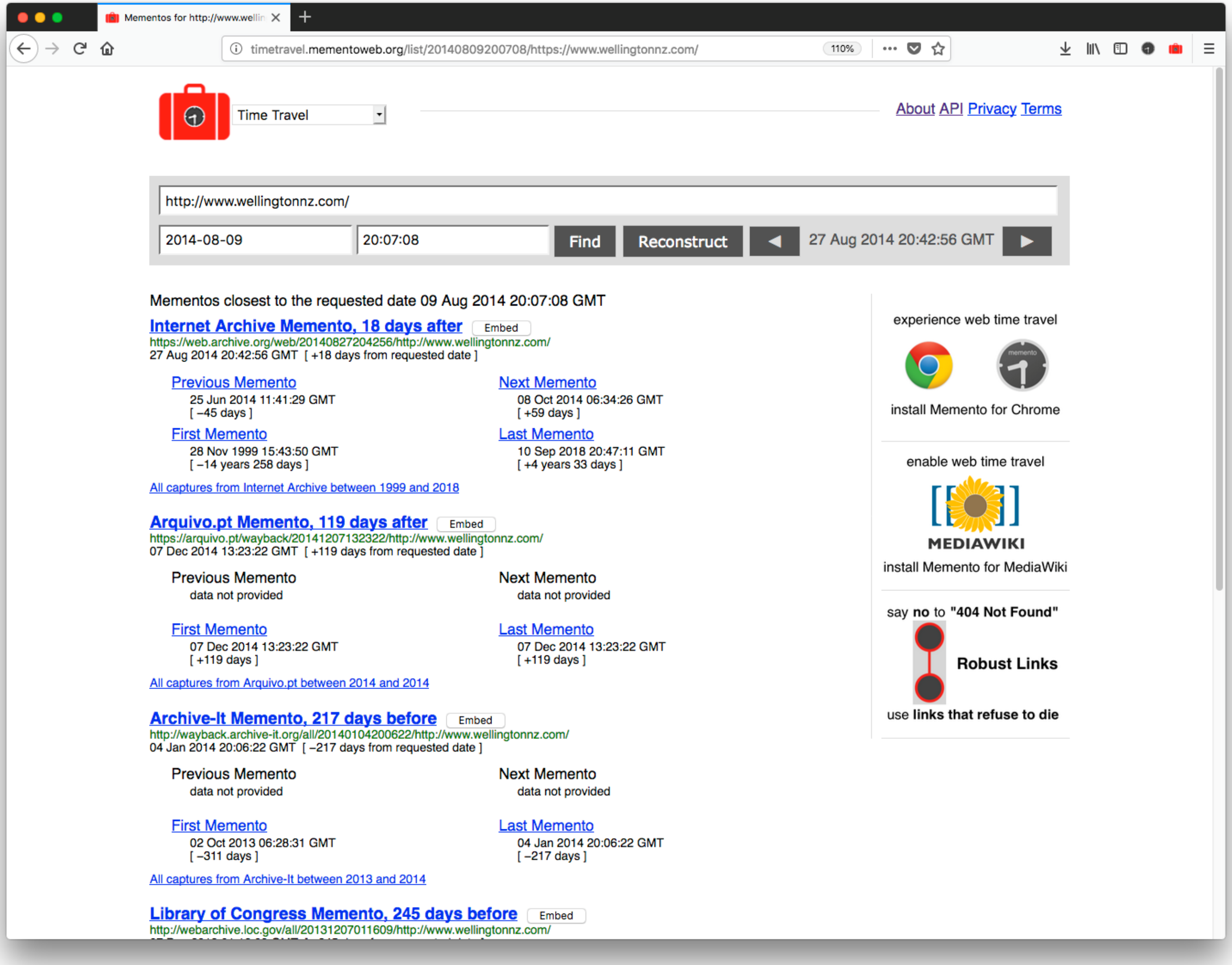
2014-08-09

20:07:08

Find Mementos in Internet Archive, Archive-It, British Library, archive.today, GitHub and many more!

http://timetravel.mementoweb.org/

Memento



Mementos for http://www.wellin
timetravel.mementoweb.org/list/20140809200708/https://www.wellingtonnz.com/
110%
Time Travel
About API Privacy Terms

http://www.wellingtonnz.com/

2014-08-09 | 20:07:08 | Find | Reconstruct | ◀ | 27 Aug 2014 20:42:56 GMT | ▶

Mementos closest to the requested date 09 Aug 2014 20:07:08 GMT

Internet Archive Memento, 18 days after Embed
https://web.archive.org/web/20140827204256/http://www.wellingtonnz.com/
27 Aug 2014 20:42:56 GMT [+18 days from requested date]

Previous Memento 25 Jun 2014 11:41:29 GMT [−45 days]	Next Memento 08 Oct 2014 06:34:26 GMT [+59 days]
First Memento 28 Nov 1999 15:43:50 GMT [−14 years 258 days]	Last Memento 10 Sep 2018 20:47:11 GMT [+4 years 33 days]

All captures from Internet Archive between 1999 and 2018

Arquivo.pt Memento, 119 days after Embed
https://arquivo.pt/wayback/20141207132322/http://www.wellingtonnz.com/
07 Dec 2014 13:23:22 GMT [+119 days from requested date]

Previous Memento data not provided	Next Memento data not provided
First Memento 07 Dec 2014 13:23:22 GMT [+119 days]	Last Memento 07 Dec 2014 13:23:22 GMT [+119 days]

All captures from Arquivo.pt between 2014 and 2014

Archive-It Memento, 217 days before Embed
http://wayback.archive-it.org/all/20140104200622/http://www.wellingtonnz.com/
04 Jan 2014 20:06:22 GMT [−217 days from requested date]

Previous Memento data not provided	Next Memento data not provided
First Memento 02 Oct 2013 06:28:31 GMT [−311 days]	Last Memento 04 Jan 2014 20:06:22 GMT [−217 days]

All captures from Archive-It between 2013 and 2014

Library of Congress Memento, 245 days before Embed
http://webarchive.loc.gov/all/20131207011609/http://www.wellingtonnz.com/

experience web time travel

install Memento for Chrome

enable web time travel

MEDIAWIKI

install Memento for MediaWiki

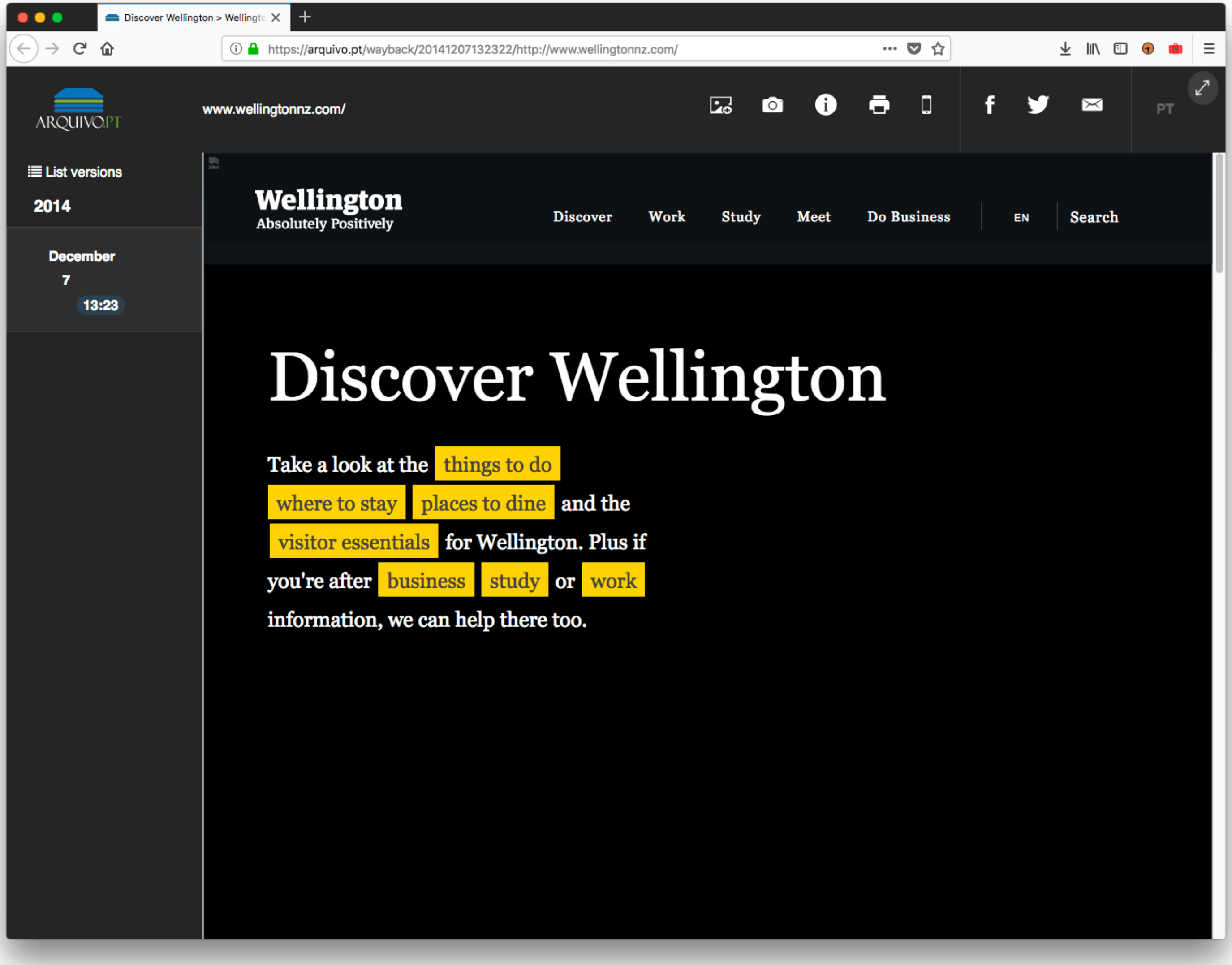
say **no** to **"404 Not Found"**

Robust Links

use **links that refuse to die**

http://timetravel.mementoweb.org/list/20140809200708/https://www.wellingtonnz.com/

Memento

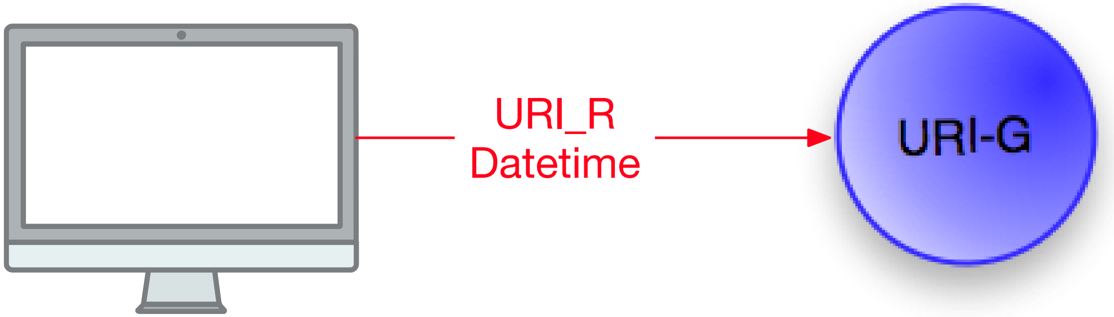


Discover Wellington

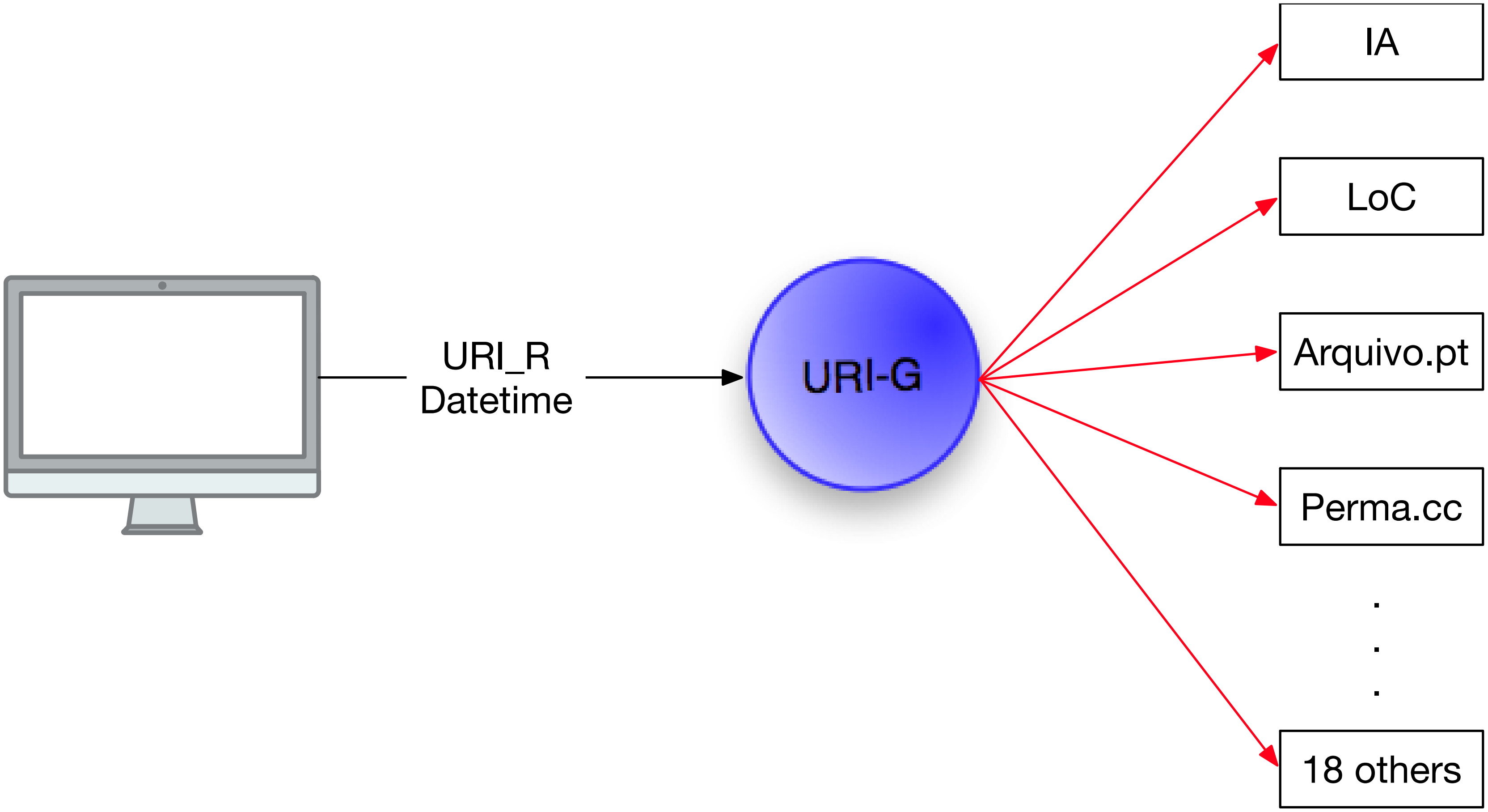
Take a look at the things to do where to stay places to dine and the visitor essentials for Wellington. Plus if you're after business study or work information, we can help there too.

https://arquivo.pt/wayback/20141207132322/http://www.wellingtonnz.com/

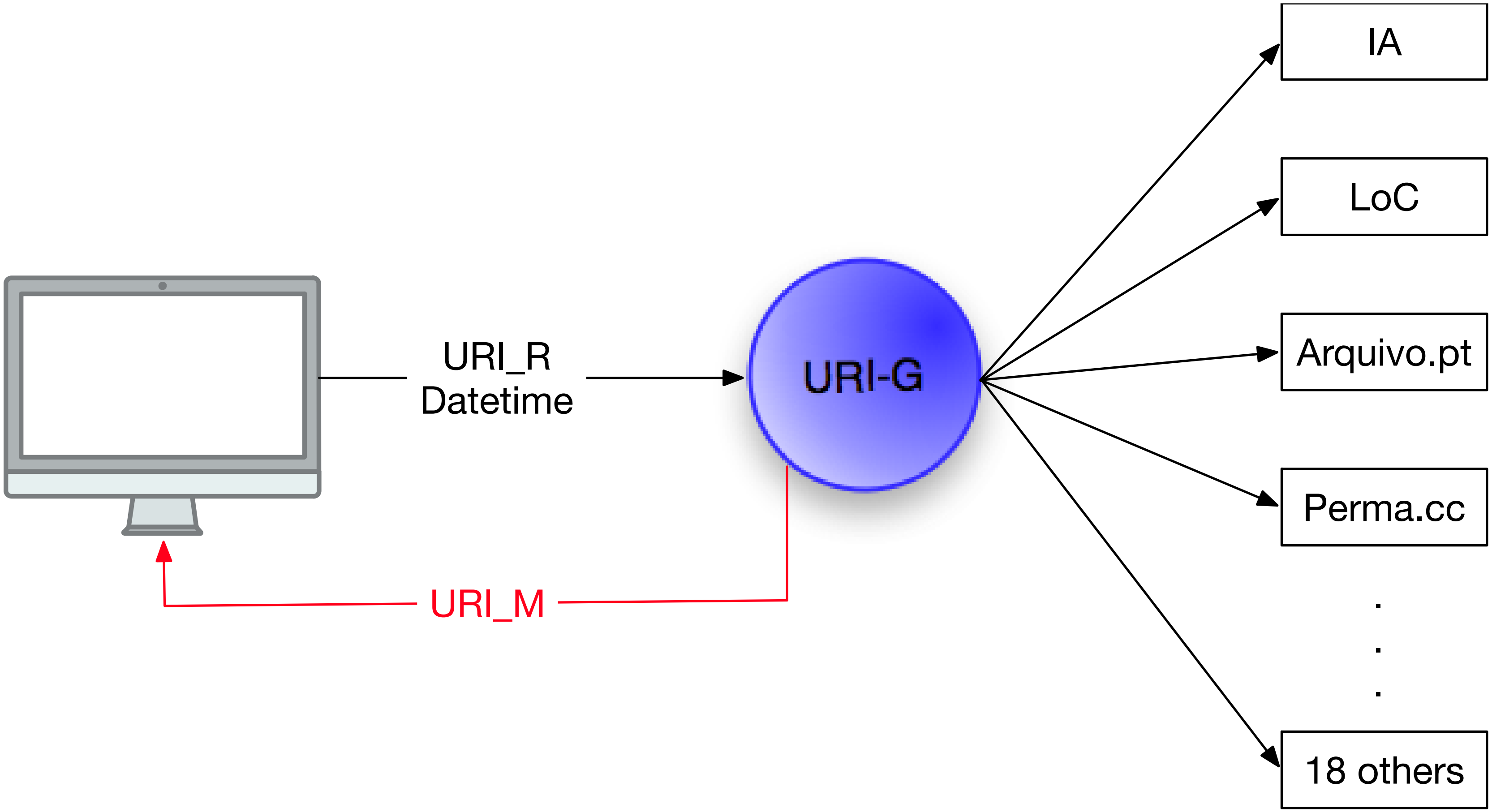
How does this work?
Memento Aggregator (very simplistic view)



Memento Aggregator



Memento Aggregator



LANL Memento Aggregator - Problem

- As the number of archives grows, sending requests to each archive for every incoming request is not feasible
  - Response times
  - Memento infrastructure load
  - Load on distributed archives

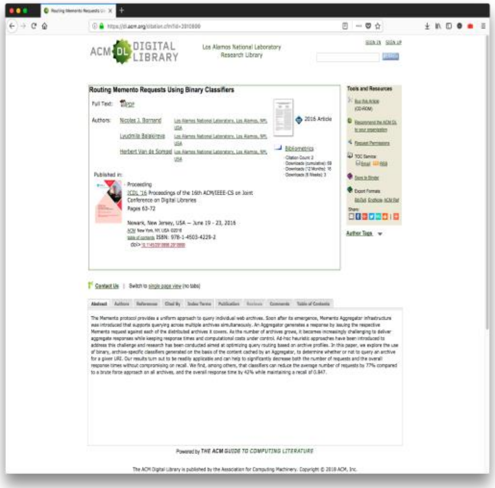
What if…

- We could predict, by merely looking at a URI-R, whether or not to issue a request to a specific archive?
  - A binary classifier per archive

- We could train the classifiers using cached data?

- That would be pretty neat, indeed:
  - Retrain classifiers as web archive collections evolve
  - Not dependent on external data
  - Querying classifiers probably way faster (msec) than polling archives (sec)

We can! Published @ JCDL 2016



Routing Memento Requests Using Binary Classifiers

Full Text: PDF

Authors: Nicolas J. Bornand, Los Alamos National Laboratory, Los Alamos, NM, USA; Lyudmila Balakireva, Los Alamos National Laboratory, Los Alamos, NM, USA; Herbert Van de Sompel, Los Alamos National Laboratory, Los Alamos, NM, USA

2016 Article

Bibliometrics
· Citation Count: 3
· Downloads (cumulative): 69
· Downloads (12 Months): 16
· Downloads (6 Weeks): 3

Published in:
Proceeding
JCDL '16 Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries
Pages 63-72

Newark, New Jersey, USA — June 19 - 23, 2016
ACM New York, NY, USA ©2016
table of contents ISBN: 978-1-4503-4229-2
doi>10.1145/2910896.2910899

The Memento protocol provides a uniform approach to query individual web archives. Soon after its emergence, Memento Aggregator infrastructure was introduced that supports querying across multiple archives simultaneously. An Aggregator generates a response by issuing the respective Memento request against each of the distributed archives it covers. As the number of archives grows, it becomes increasingly challenging to deliver aggregate responses while keeping response times and computational costs under control. Ad-hoc heuristic approaches have been introduced to address this challenge and research has been conducted aimed at optimizing query routing based on archive profiles. In this paper, we explore the use of binary, archive-specific classifiers generated on the basis of the content cached by an Aggregator, to determine whether or not to query an archive for a given URI. Our results turn out to be readily applicable and can help to significantly decrease both the number of requests and the overall response times without compromising on recall. We find, among others, that classifiers can reduce the average number of requests by 77% compared to a brute force approach on all archives, and the overall response time by 42% while maintaining a recall of 0.847.

Powered by THE ACM GUIDE TO COMPUTING LITERATURE

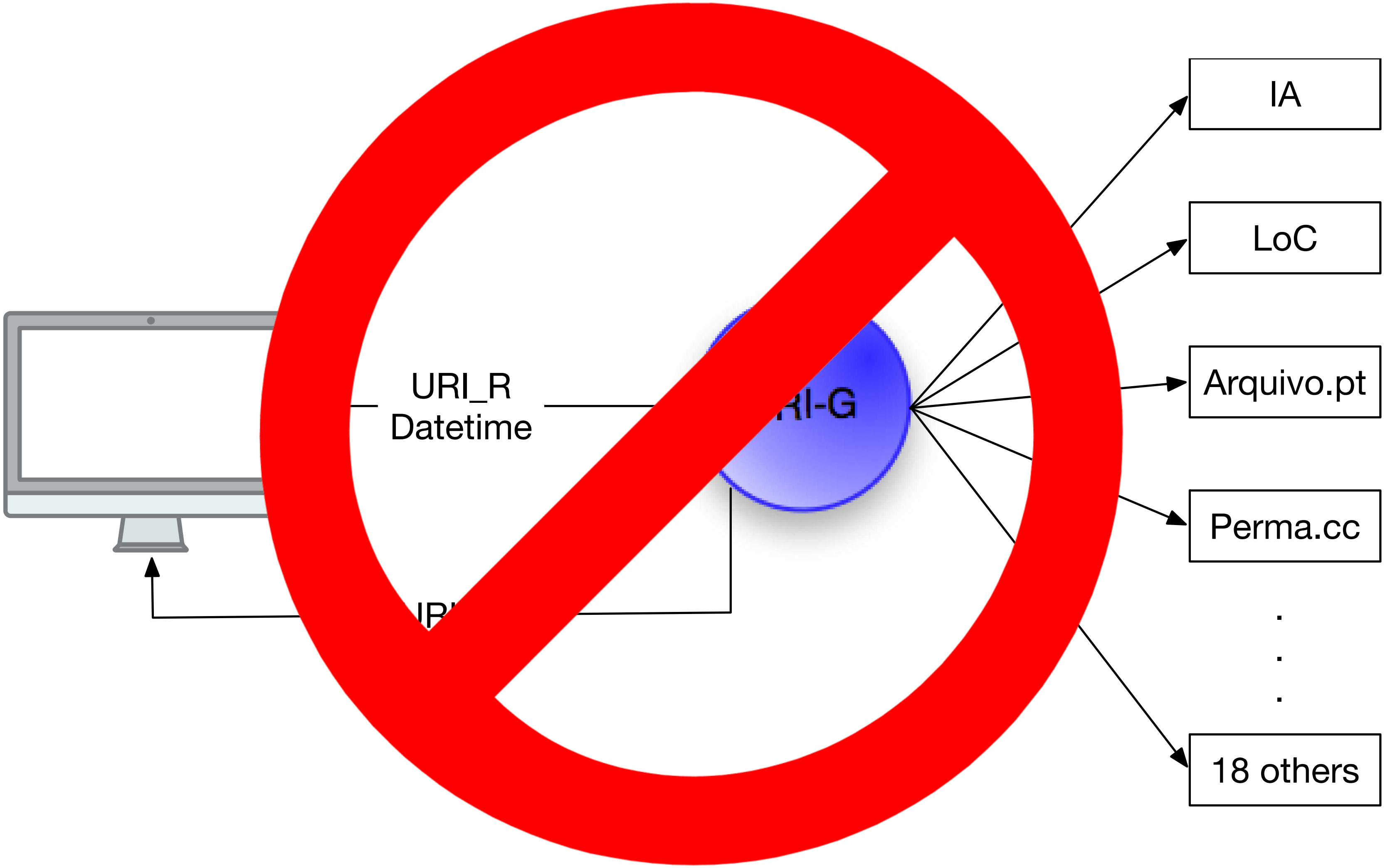
The ACM Digital Library is published by the Association for Computing Machinery. Copyright © 2018 ACM, Inc.

https://doi.org/10.1145/2910896.2910899

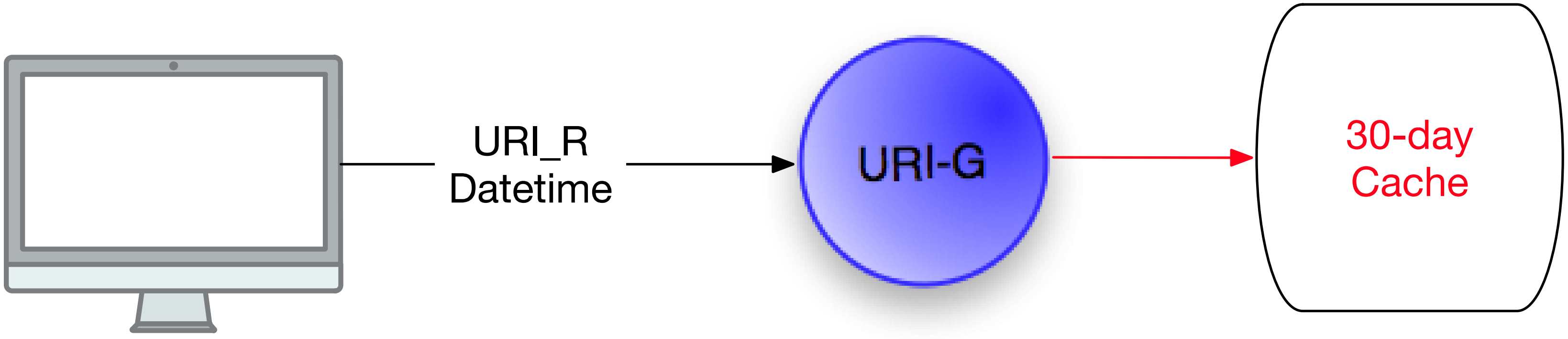
- ML models based on simple URI features
  - Character count, n-grams, domain
- Common ML algorithms used per archive
  - Logistic Regression, Multinomial Bayes, SVM
- Optimized for
  - Prediction time, not training time
  - Reduction of false positive rate

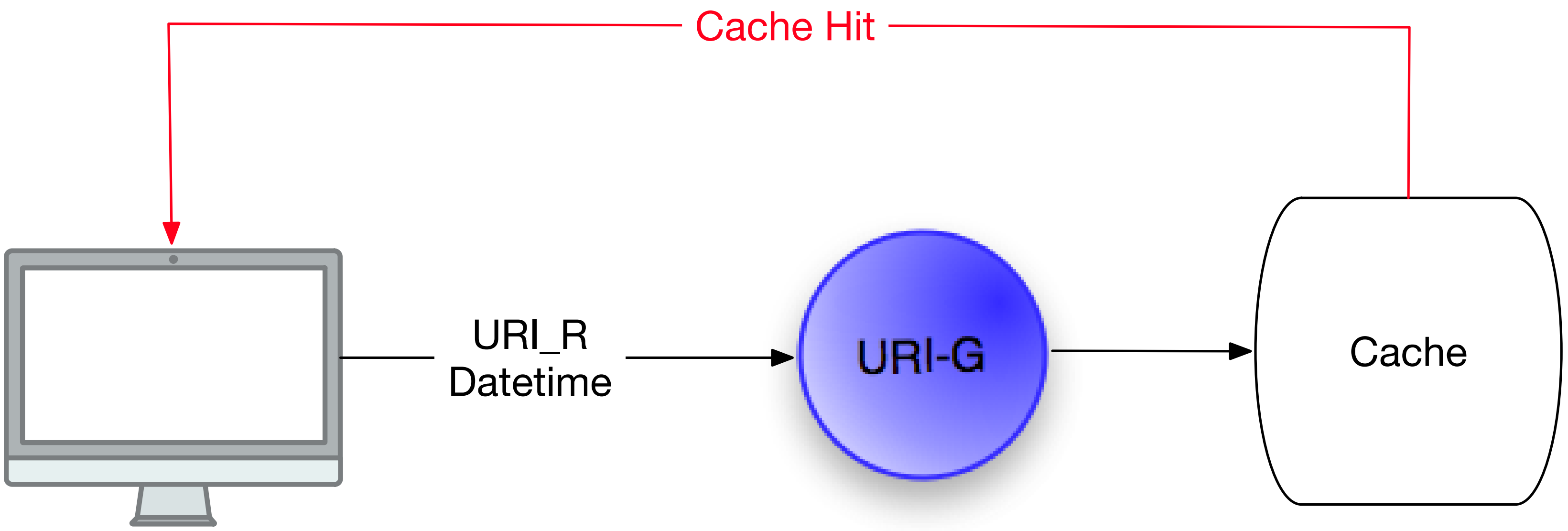
Results:

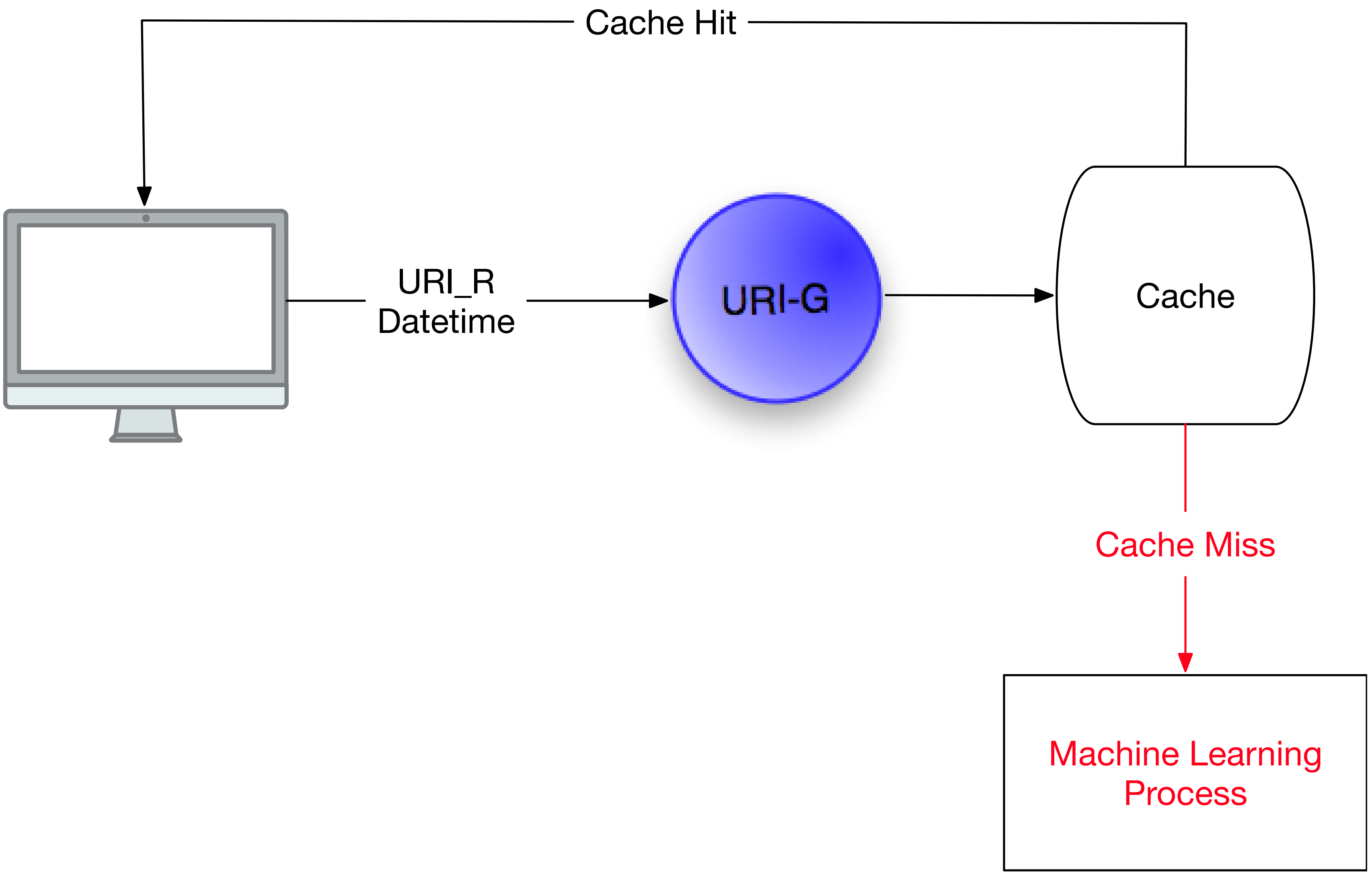
- Requests per URI-R: **3.96 vs 11**
- Response time:
  2.16s vs 3.08s
- Recall:
  84.7%

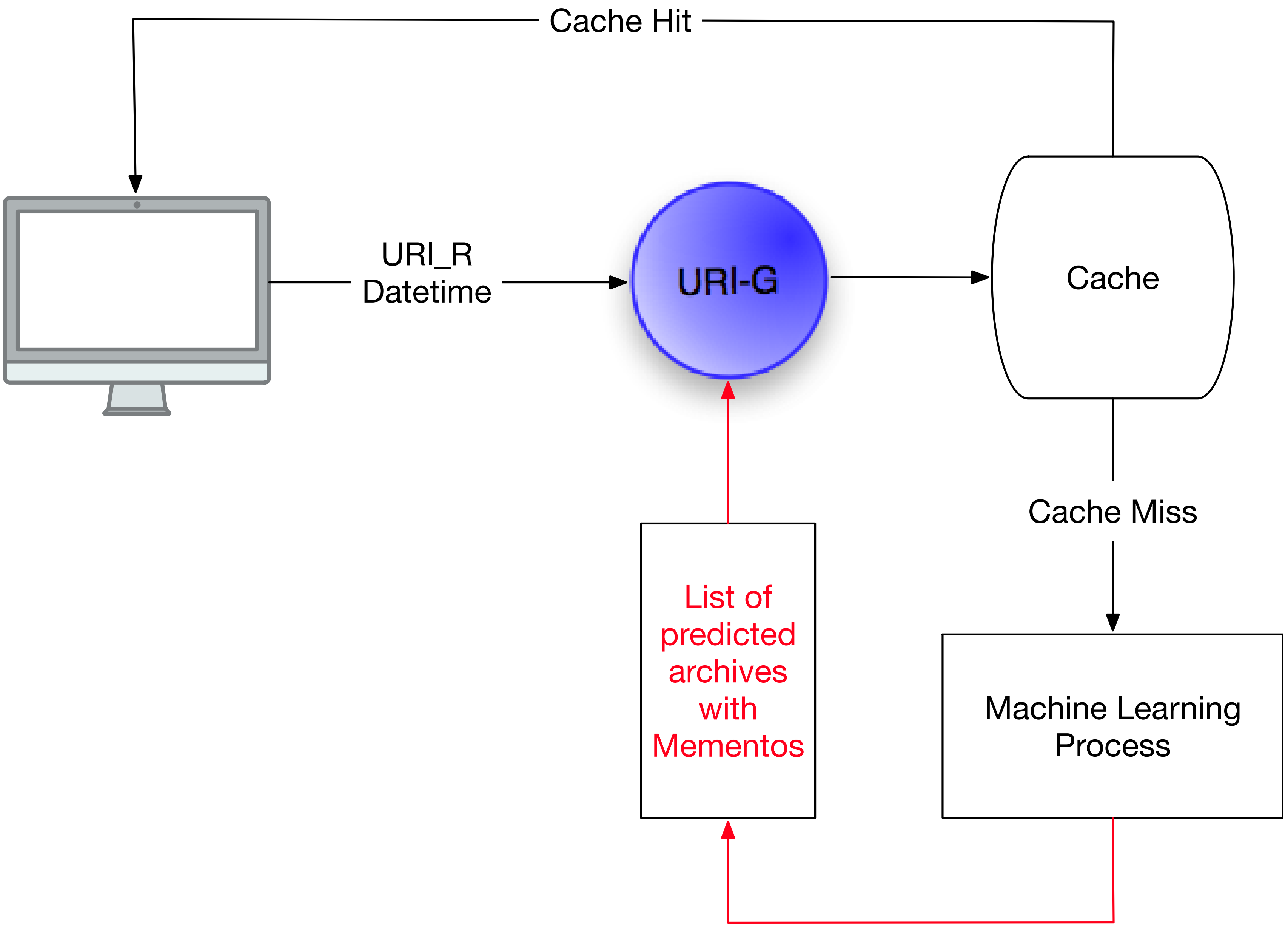

IA
LoC
URI_R
Datetime
RI-G
Arquivo.pt
Perma.cc
.
.
.
IR
18 others

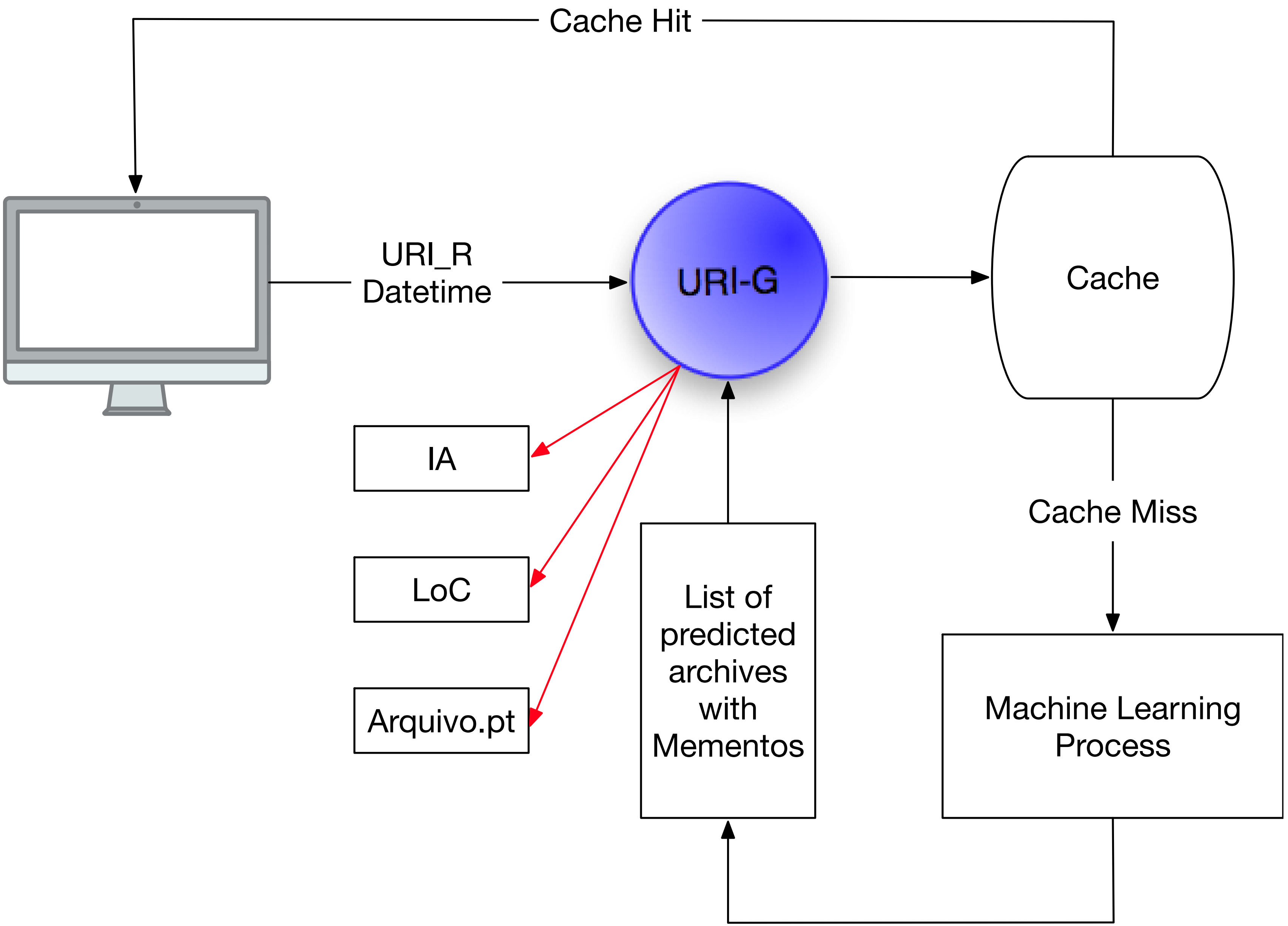
In Production…


URI_R
Datetime
URI-G
30-day
Cache


Cache Hit
URI_R
Datetime
URI-G
Cache


Cache Hit
URI_R
Datetime
URI-G
Cache
Cache Miss
Machine Learning Process


Cache Hit
URI_R
Datetime
URI-G
Cache
Cache Miss
List of predicted archives with Mementos
Machine Learning Process


Cache Hit

URI_R
Datetime

URI-G

Cache

IA

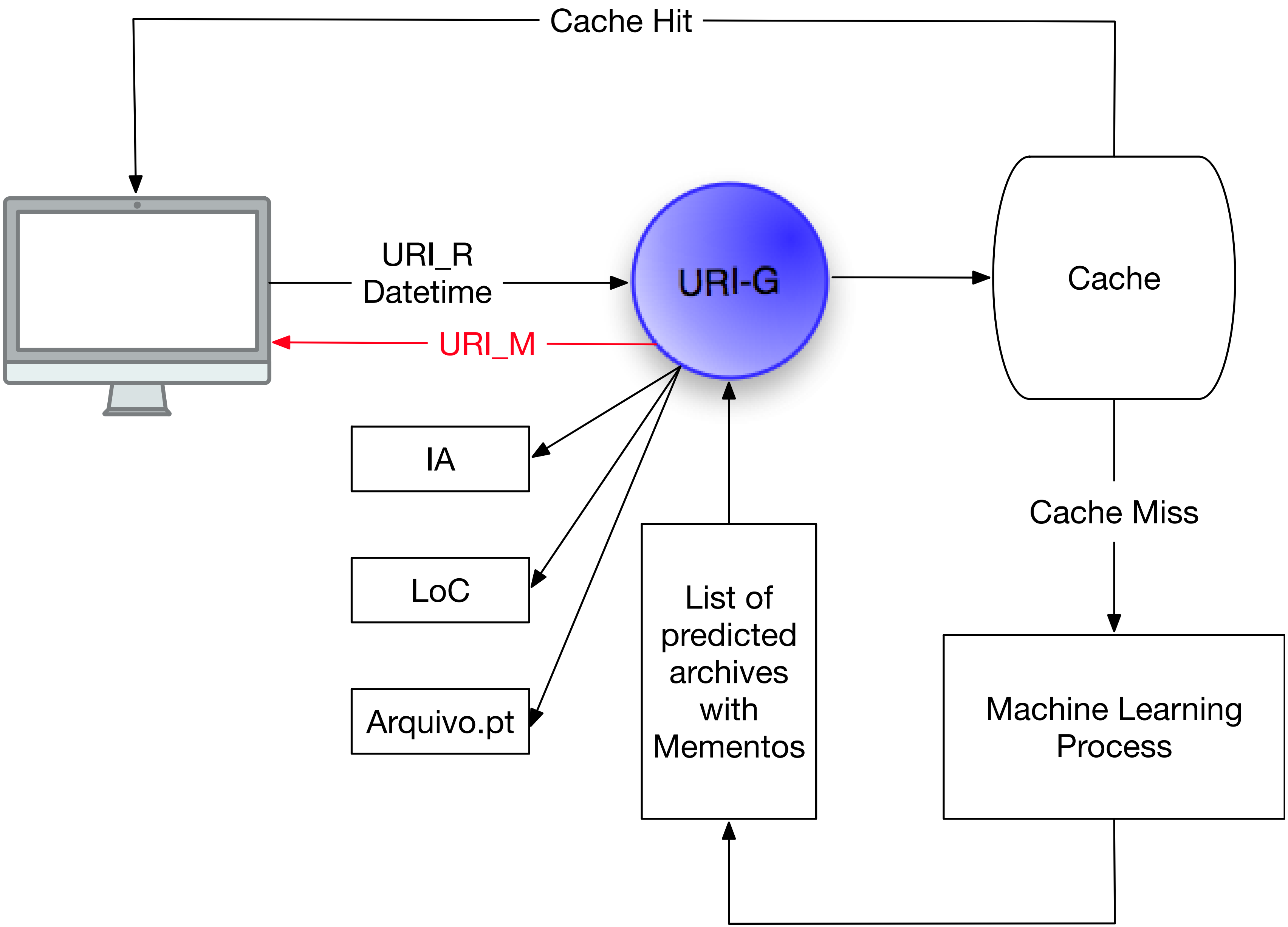
LoC

Arquivo.pt

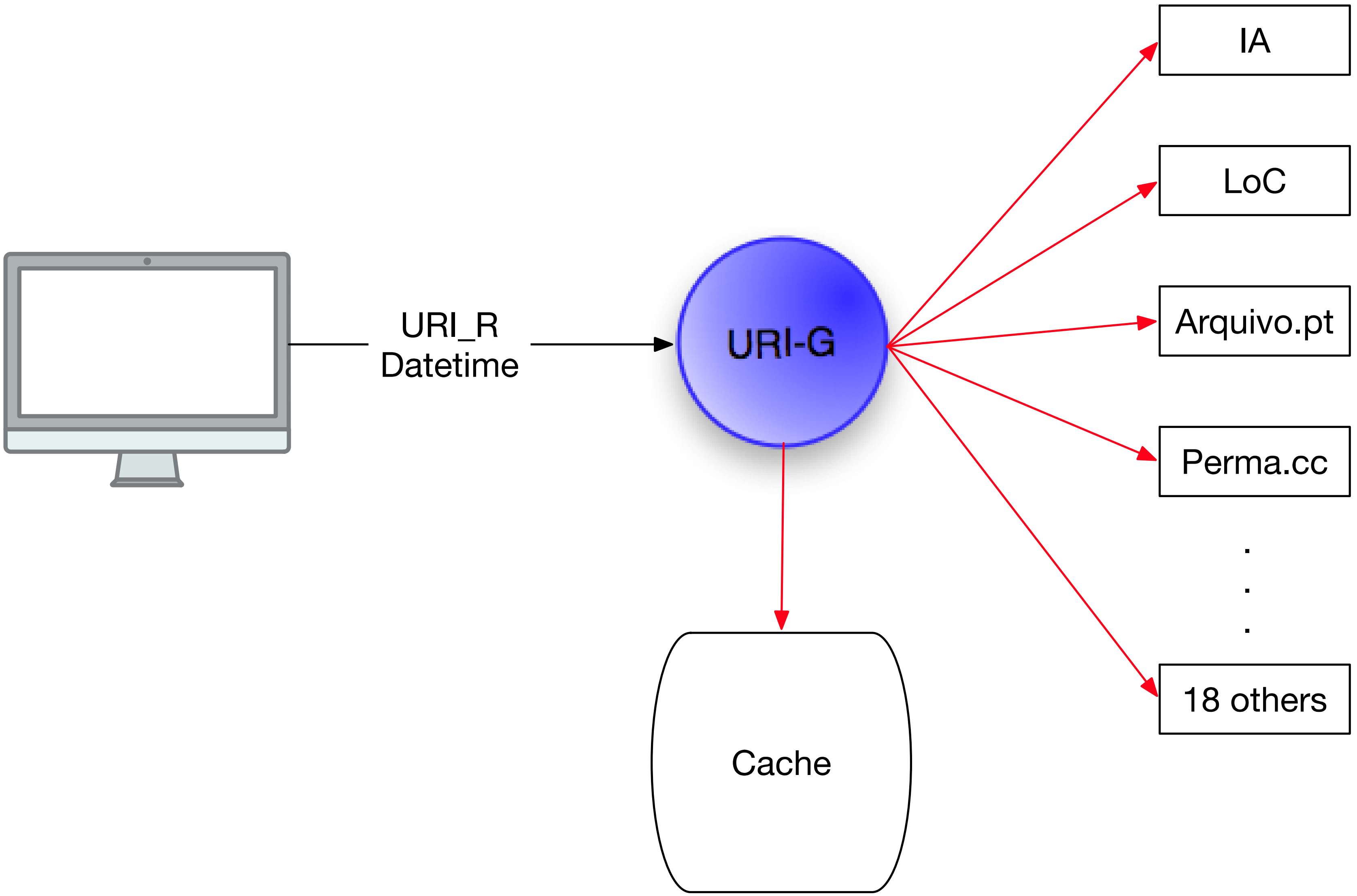
List of predicted archives with Mementos

Cache Miss

Machine Learning Process


Cache Hit
URI_R
Datetime
URI_M
URI-G
Cache
IA
LoC
Arquivo.pt
List of predicted archives with Mementos
Cache Miss
Machine Learning Process

Populating the Cache



Questions to Ask

- How effective is the cache?
  - What is the hit/miss ratio? Does it vary for different Memento services?
  - Is the cache freshness period appropriate?

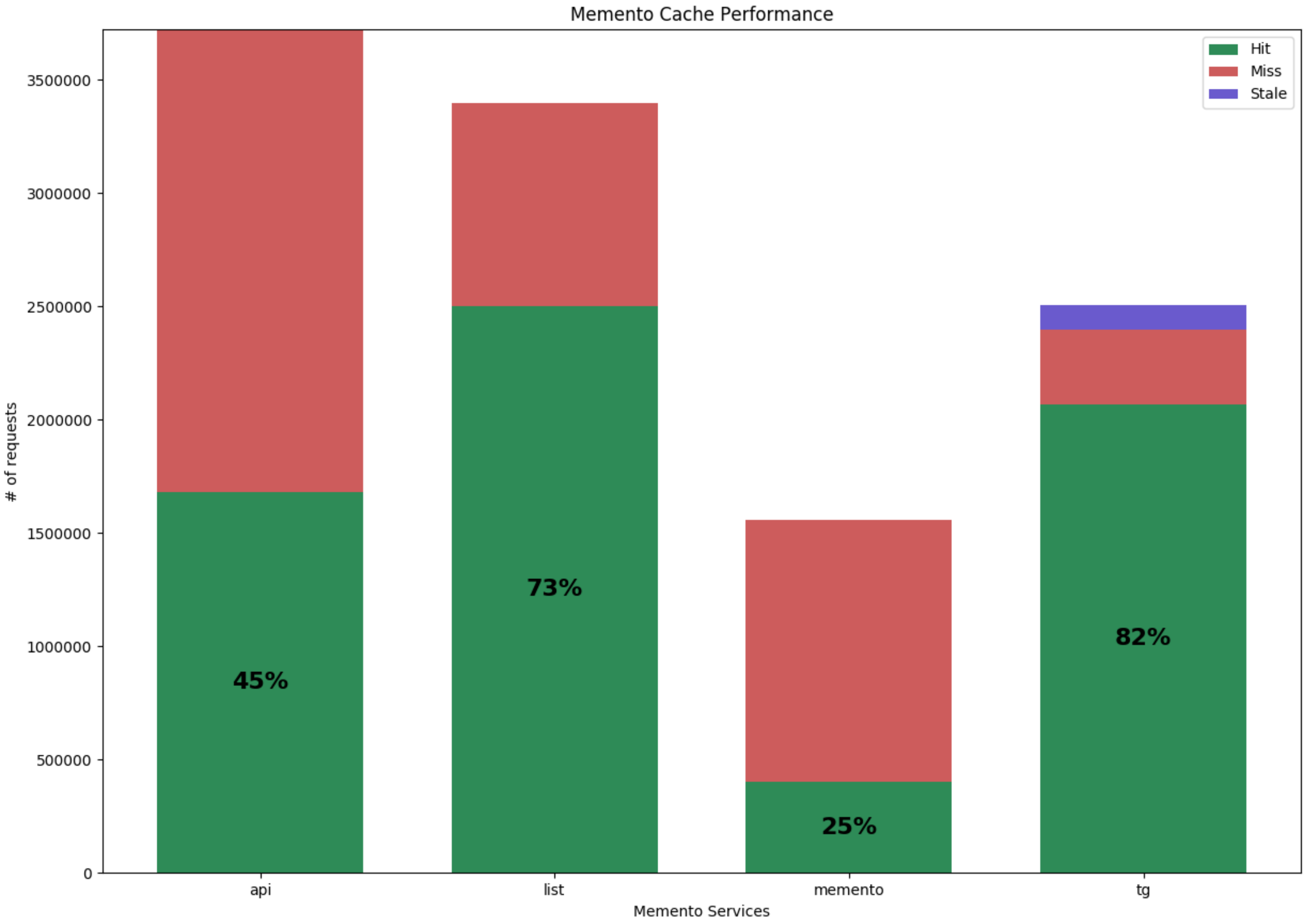
- How effective is the ML process?
  - What is the false negative and false positive rate?
  - Do we need to retrain the models? How often?

Evaluation

- Memento Aggregator currently covers
  - 23 web archives
  - 17 with native memento support
  - 6 with by-proxy memento support

- Analysis of log files
  - recorded between July 4th 2017 and October 17th 2018
  - > 11m requests in total
  - Approx. 2.6m requests against machine learning process
    - Results in 2.6m lookups to populate cache
      - Used as “truth” to assess ML prediction

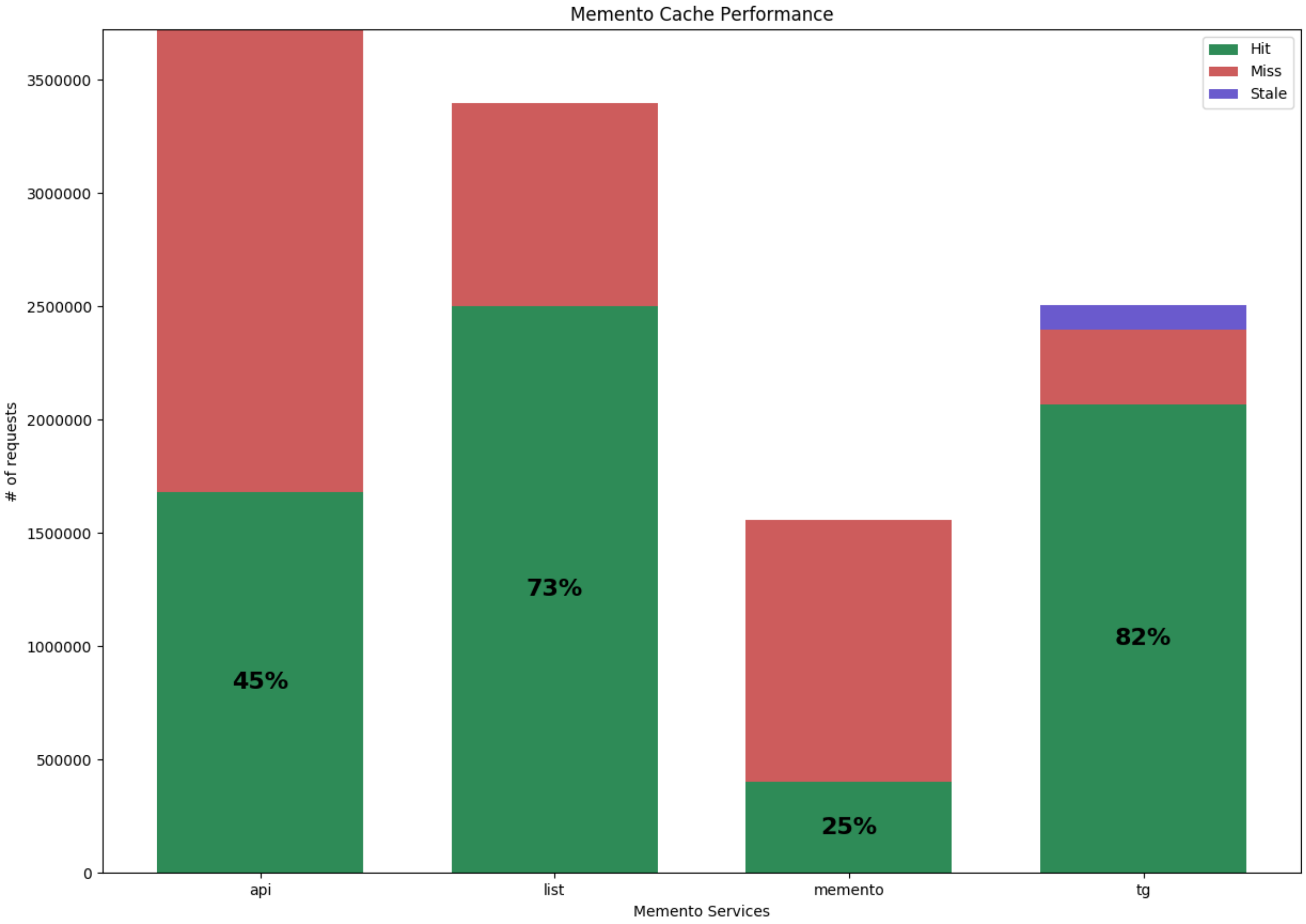
Cache Hit/Miss Rate



Smart Routing of Memento Requests
@mart1nkle1n
IIPC WAC 2018, 11/15/2018, Wellington, NZ

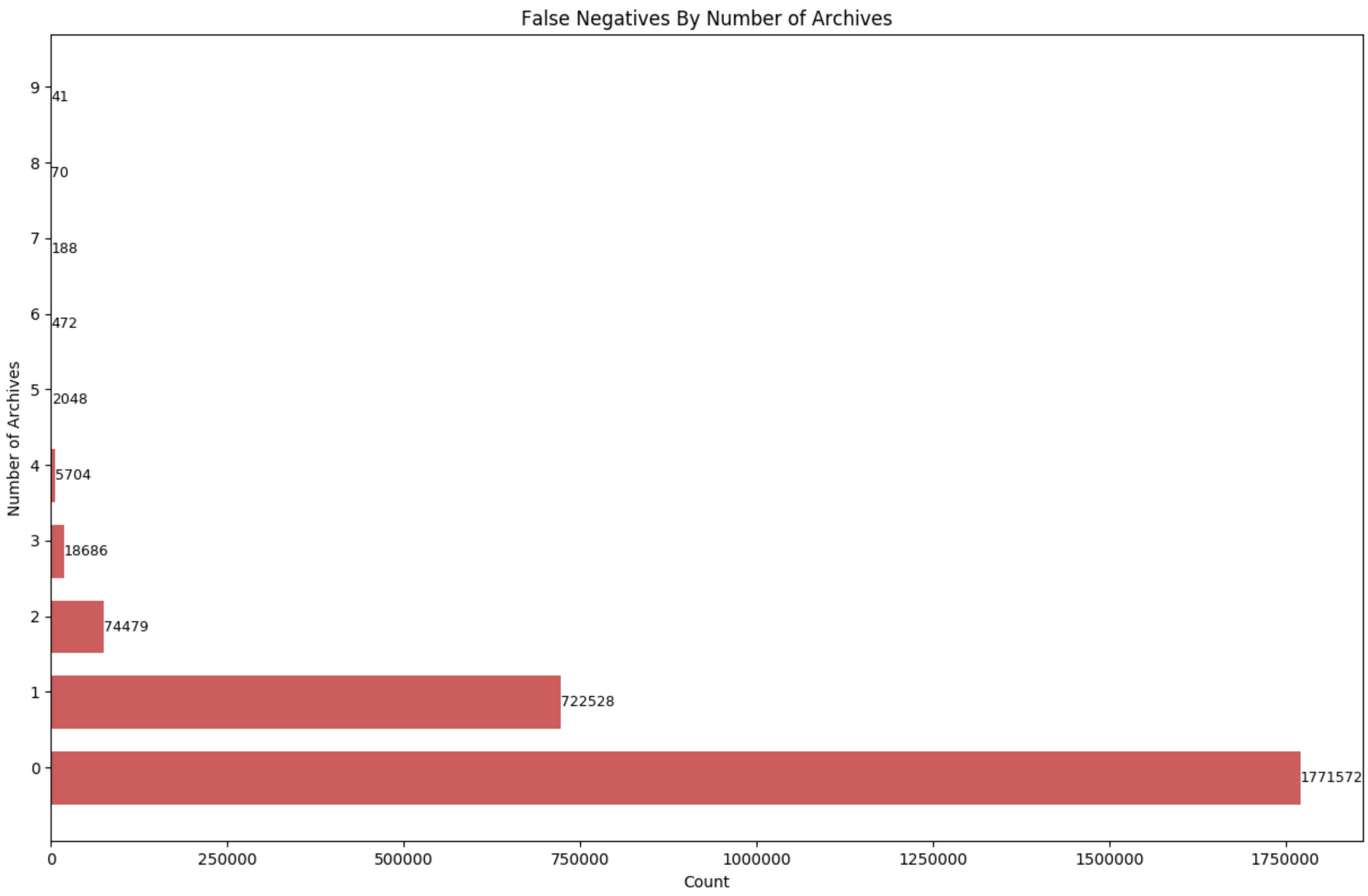
Los Alamos
NATIONAL LABORATORY

Cache Hit/Miss Rate

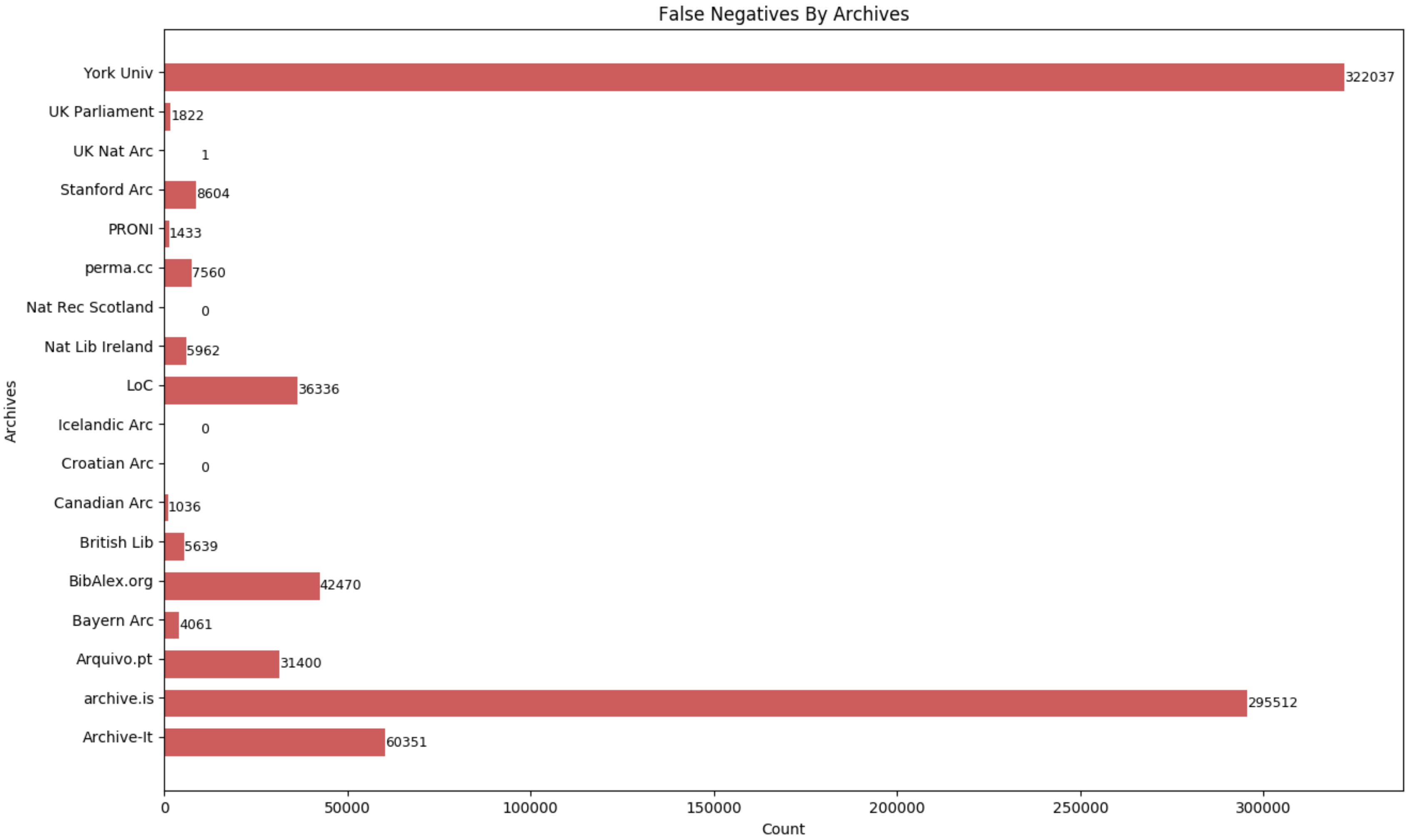


Mostly driven by machines humans machines humans

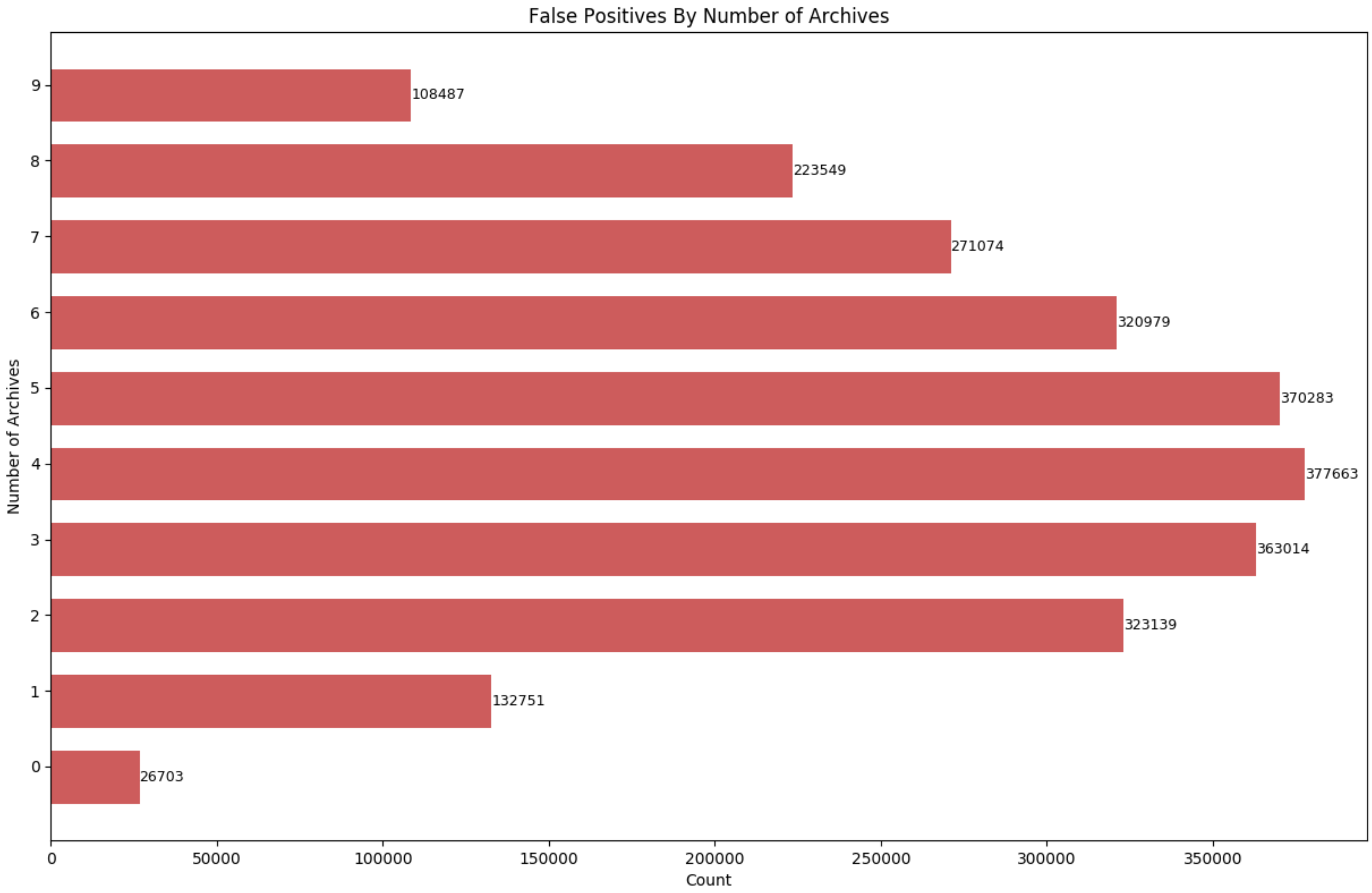
False Negatives by Number of Archives



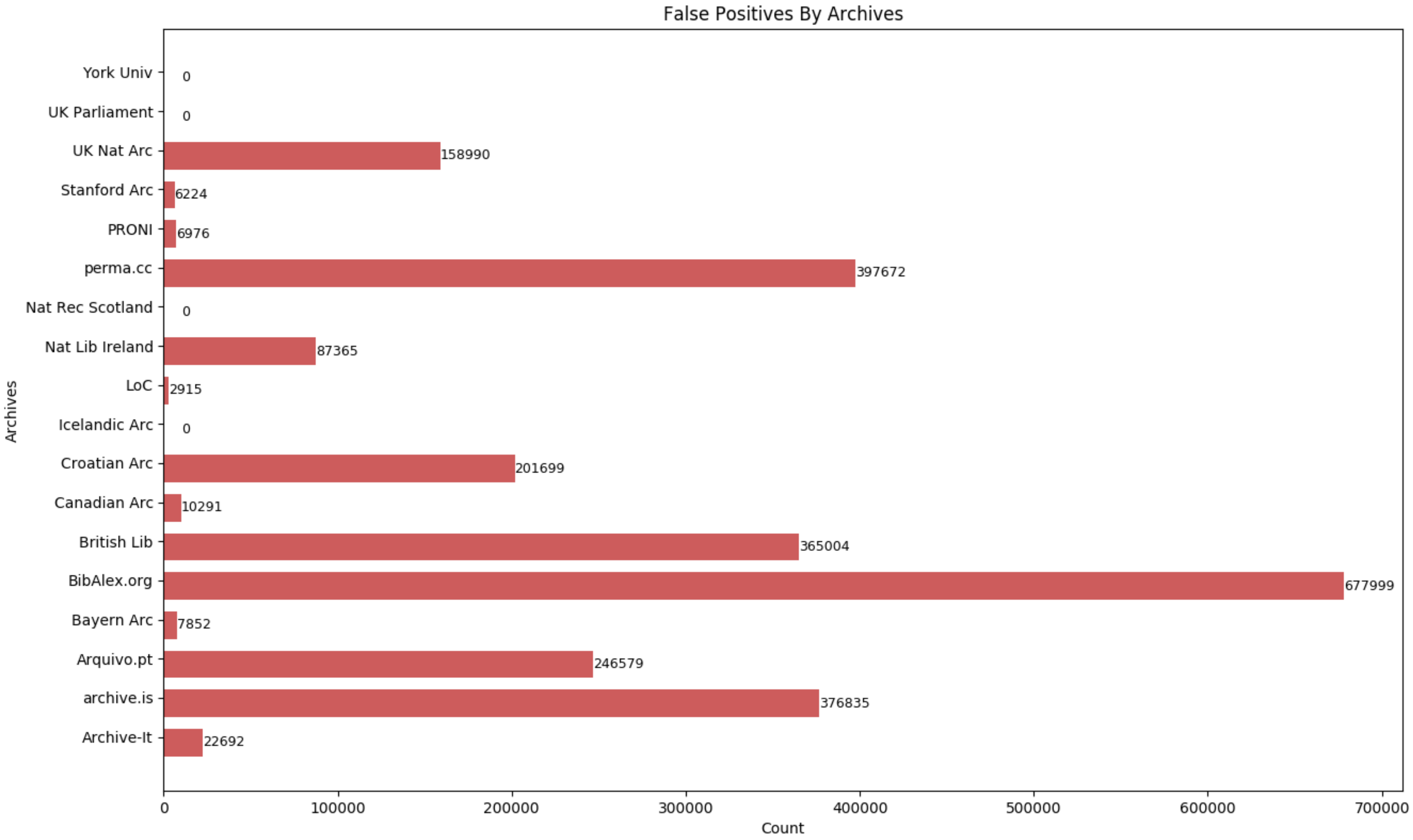
False Negatives by Archive



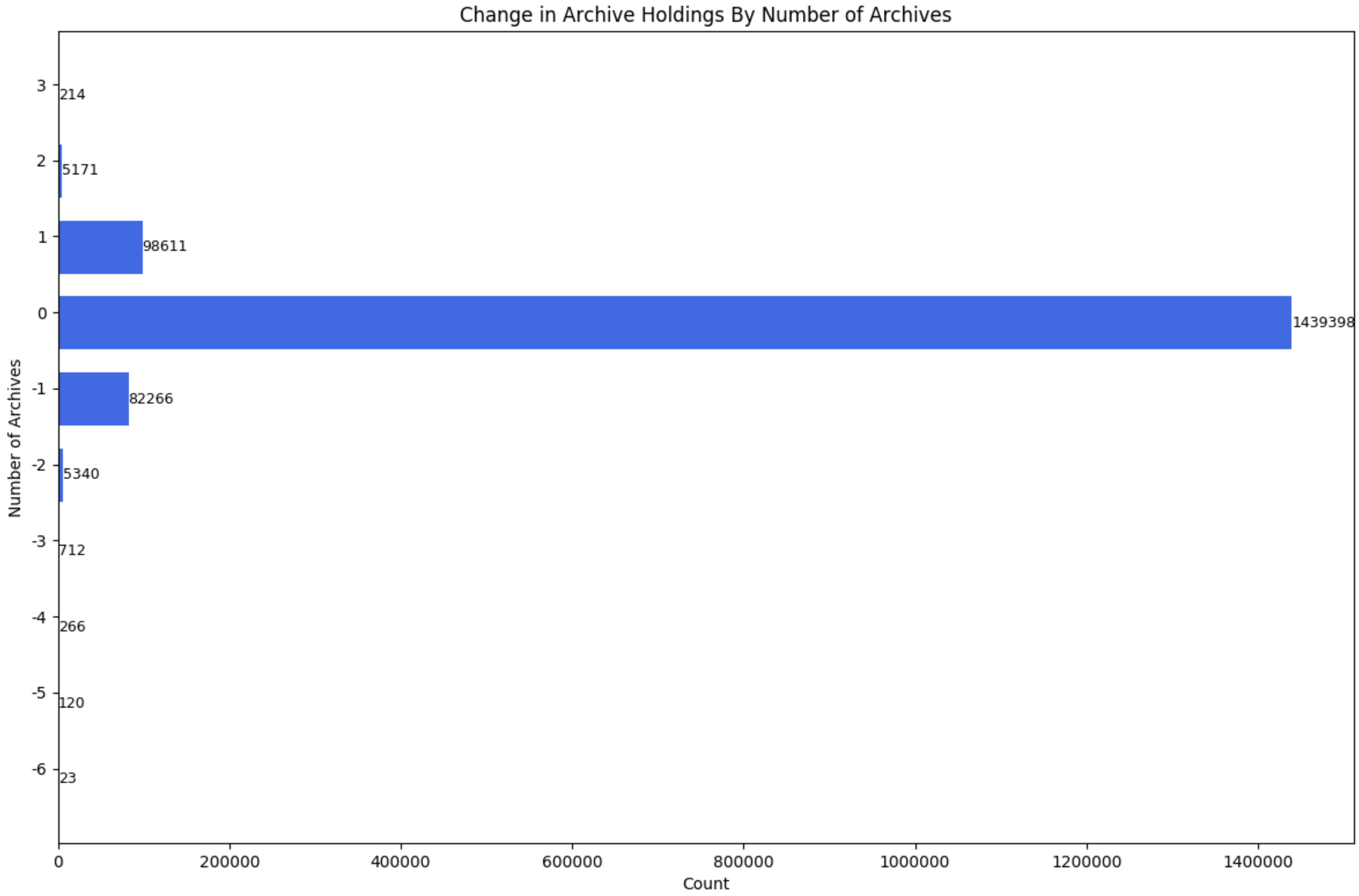
False Positives by Number of Archives



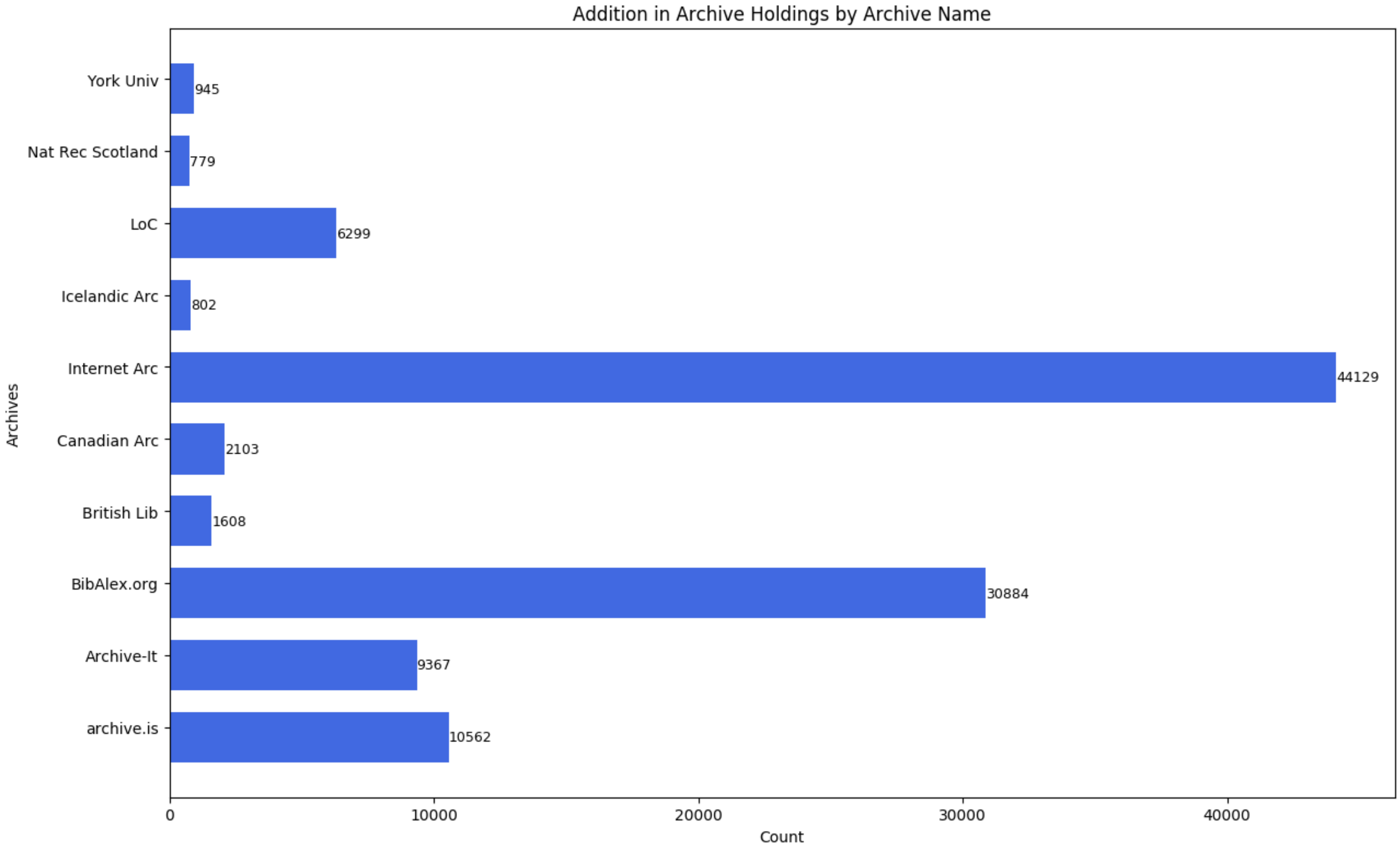
False Positives by Archive



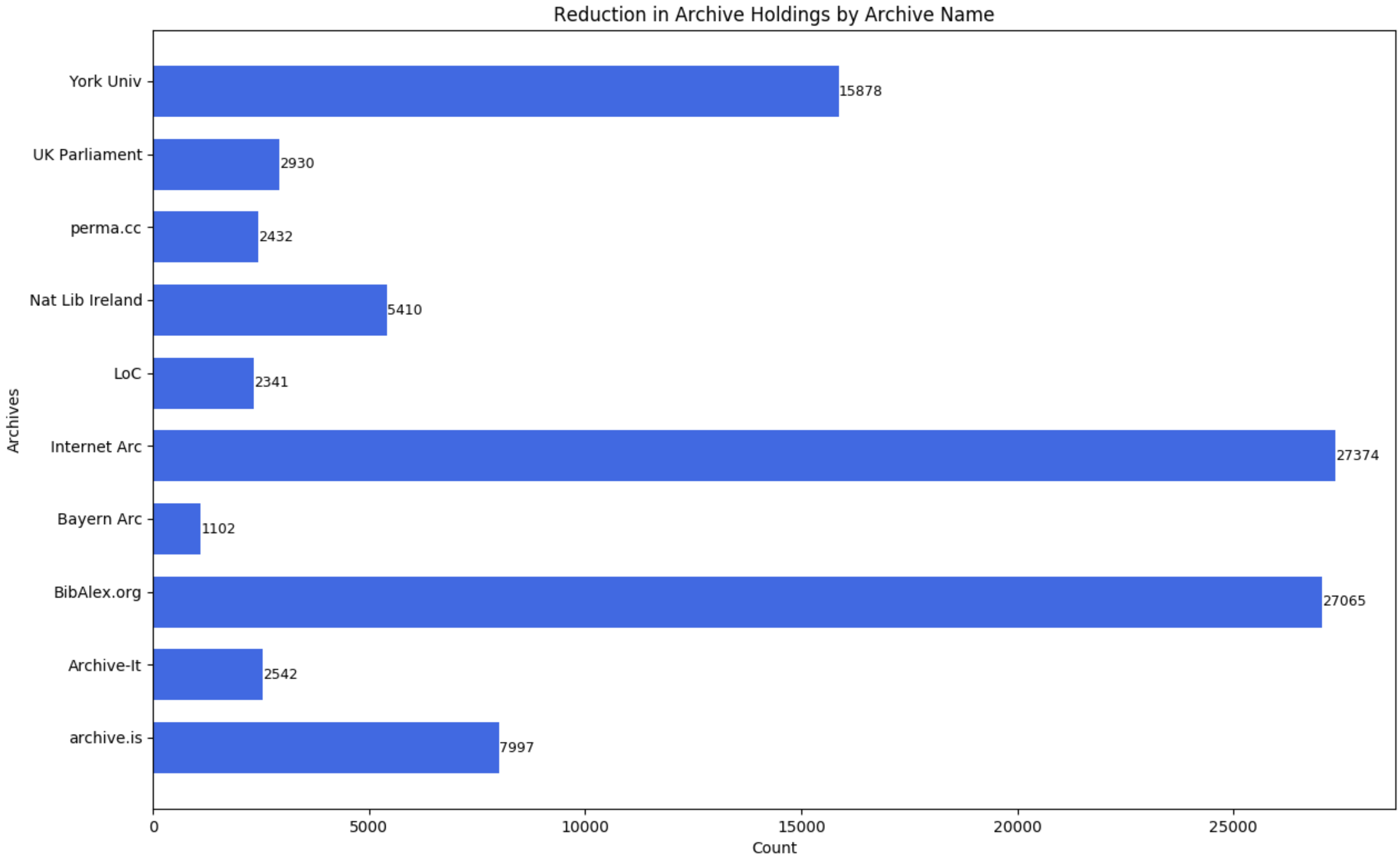
Changes in Archive Holdings



Archives Added



Archives Removed



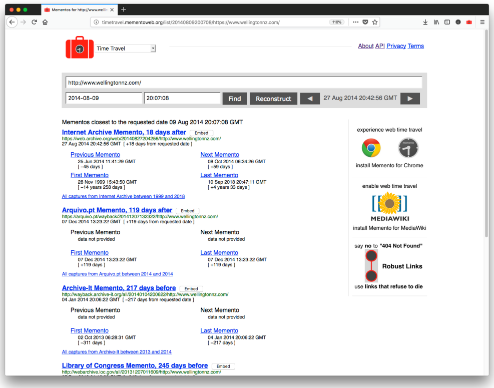
Takeaways

- Memento Aggregator cache is very effective
  - Especially for human-driven services
- Machine learning process saves!
  - Requests & time while at acceptable recall level
  - FPR: 0.33 (std dev: 0.16)
- Re-training seems necessary, frequency TBD

Optimization

- ML model trained on archival holdings, not usage logs/cache
  - Beneficial for new archives
- Neural network classifier, based on simple URI features, show promising results

Smart Routing of Memento Requests



Mementos for http://www.wellin

timetravel.mementoweb.org/list/20140809200708/https://www.wellingtonnz.com/

110%

Time Travel

About API Privacy Terms

http://www.wellingtonnz.com/

2014-08-09 | 20:07:08 | Find | Reconstruct | ◀ | 27 Aug 2014 20:42:56 GMT | ▶

Mementos closest to the requested date 09 Aug 2014 20:07:08 GMT

Internet Archive Memento, 18 days after Embed
https://web.archive.org/web/20140827204256/http://www.wellingtonnz.com/
27 Aug 2014 20:42:56 GMT [+18 days from requested date]

Previous Memento
25 Jun 2014 11:41:29 GMT
[−45 days]

Next Memento
08 Oct 2014 06:34:26 GMT
[+59 days]

First Memento
28 Nov 1999 15:43:50 GMT
[−14 years 258 days]

Last Memento
10 Sep 2018 20:47:11 GMT
[+4 years 33 days]

All captures from Internet Archive between 1999 and 2018

Arquivo.pt Memento, 119 days after Embed
https://arquivo.pt/wayback/20141207132322/http://www.wellingtonnz.com/
07 Dec 2014 13:23:22 GMT [+119 days from requested date]

Previous Memento
data not provided

Next Memento
data not provided

First Memento
07 Dec 2014 13:23:22 GMT
[+119 days]

Last Memento
07 Dec 2014 13:23:22 GMT
[+119 days]

All captures from Arquivo.pt between 2014 and 2014

Archive-It Memento, 217 days before Embed
http://wayback.archive-it.org/all/20140104200622/http://www.wellingtonnz.com/
04 Jan 2014 20:06:22 GMT [−217 days from requested date]

Previous Memento
data not provided

Next Memento
data not provided

First Memento
02 Oct 2013 06:28:31 GMT
[−311 days]

Last Memento
04 Jan 2014 20:06:22 GMT
[−217 days]

All captures from Archive-It between 2013 and 2014

Library of Congress Memento, 245 days before Embed
http://webarchive.loc.gov/all/20131207011609/http://www.wellingtonnz.com/

experience web time travel

install Memento for Chrome

enable web time travel

MEDIAWIKI

install Memento for MediaWiki

say **no** to **"404 Not Found"**

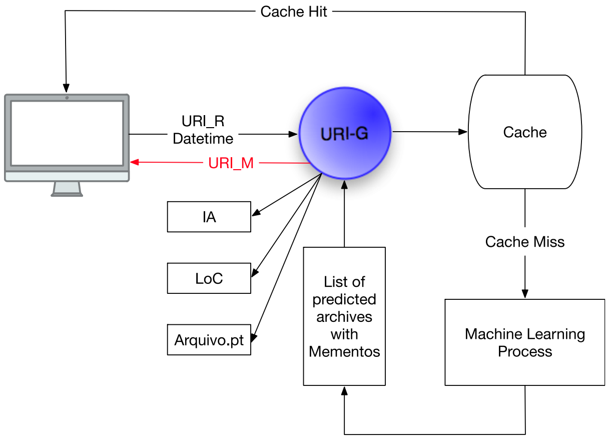
Robust Links

use **links that refuse to die**

Martin Klein[1]
Lyudmila Balakireva[1]
Harihar Shankar[1]
James Powell[1]
Herbert Van de Sompel[2]

[1]Research Library
Los Alamos National Laboratory

[2]Data Archiving and Networked Services
The Netherlands



Cache Hit
URI_R
Datetime
URI_M
URI-G
Cache
IA
LoC
Arquivo.pt
List of predicted archives with Mementos
Cache Miss
Machine Learning Process