

# Forward Experiment

## Investigation of content drift and link rot in Web-at-large references in a sample of two corpora of scholarly research papers

James Powell

### High level outline

- First part explores available metadata for web-at-large references
  - Link rot (HTTP status, content length, mime type)
  - Content drift (content length, mime type)
- Second part focuses on content similarity from interval to interval
  - Content length and content drift
  - Various similarity measures – Hash, Vsim, Graph Edit?
  - Exploring content drift in aggregate for a given publication, e.g. visualizing drift, calculating a “drift score” based on aggregate drift among web at large references for a paper

## High level outline ... cont'd

- Third part looks at Machine Learning algorithms applied to this data
  - Can machine learning algorithms learn to identify/anticipate either of these problems?
  - Are there associations among the data that other methods didn't detect?
  - What are possible uses for ML in this context?

## Question(s)?

- Is it the case that for this sample that both link rot and content drift increase with the passage of time, as has previously been noted?
- What do the link availability dynamics (changes in HTTP status and content availability) tell us about the potential for content drift for a given resource?
- Do status chains exhibit any predictable patterns over time?
- Is content change dramatic? Fast or slow? Is it ever prone to reversal?
- Does the information content of a cited resource increase, decrease or stay the same over time?
- Are some top level domains more volatile than others?

## Other similarity measures

- Similarity hash comparisons
- Similarity hash results as compared to preliminary feature vector comparisons
- Analysis of citations when status code changes during review period
- Status and file size changes
- What else?

## Describe experiment

- A sample of arXiv and PLOS papers were identified for this experiment
- For each collection for a period of time (whenever), all links that were not DOI/scholarly links were extracted
- These links were harvested on a regular basis
- At each harvest a copy was pushed to a Web archive, and saved locally.
- The initial HTTP status of a request along with size and mime (in fact, entire HTTP header) were stored

## Descriptive stats for experimental data

- ArXiv : #source papers
  - ☒ 17,158 citations with 14, 966 unique URLs
  - ☒ 97.82% were HTML documents
  - ☒ 74.06% of cited items changed size
  - ☒ 74.50% of originals that were available were “always” available
- PLOS : #source papers
  - ☒ 8,284 citations with 6,549 unique URLs,
  - ☒ 99.12% were HTML documents
  - ☒ 80.84% of cited items changed size
  - ☒ 69.00% of originals that were available were “always” available

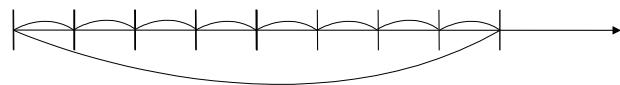
## Characterize Data after Filtering

- What filtered and why? Network outage days.
- All papers and references affected by the outage were discarded. Stats.
- Stats of final corpus. Including % of URLs changed according to status code, content length, content type

Metadata	Content
<b>Compare to original (examine all URIs)</b>	Similarity hash Vector-Cosine similarity
<b>Compare with last item retrieved (examine only URIs that exhibit change over lifespan)</b>	Byte count Status code

## Illustration of comparison strategies

Metadata: HTTP last status, previous vs current



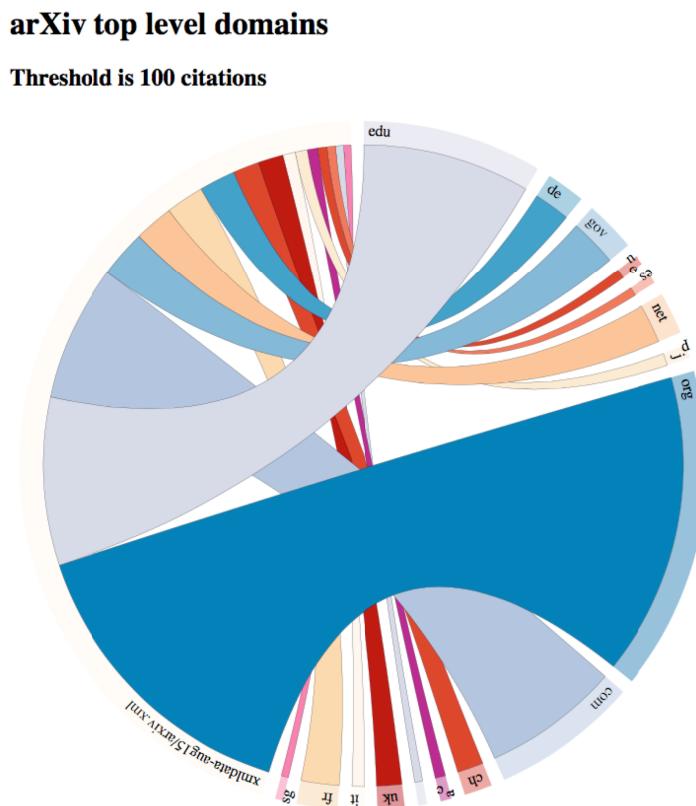
Metadata: Size, subsequent interval compared to initial

Content: Similarity, subsequent interval compared to initial

In sequence vs forward looking/predictive, e.g. previous vs current status, original size vs size of item retrieved at subsequent intervals, etc.

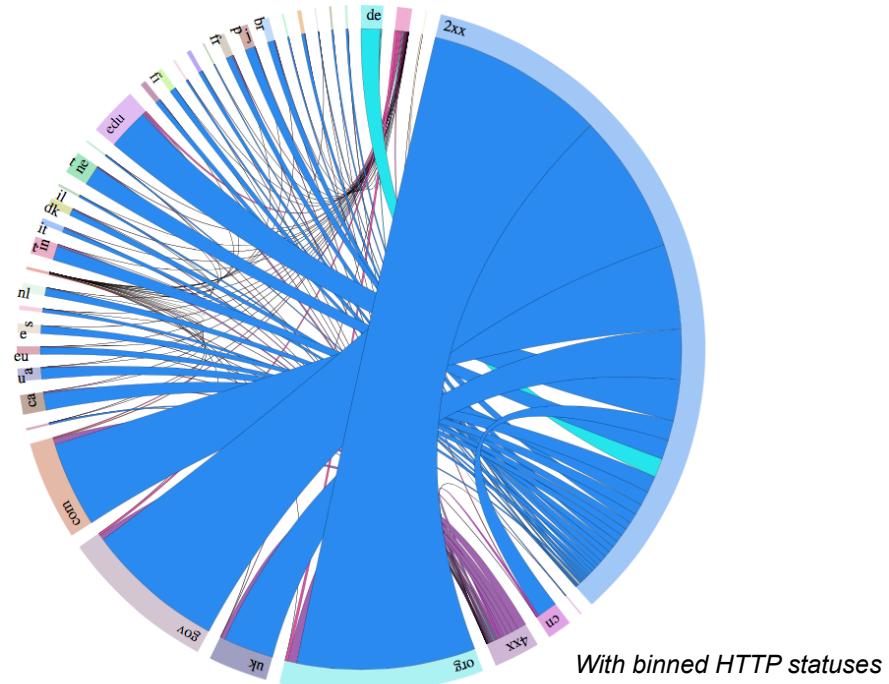
# Metadata Extraction

- Procedure description
  - HTTP headers: content-length, mime-type, status codes
  - Status inspection can be used to characterize link rot
    - Most common top level domains for a corpus
    - Most frequently occurring tlds and their associated statuses
    - Most common tlds yielding a 404



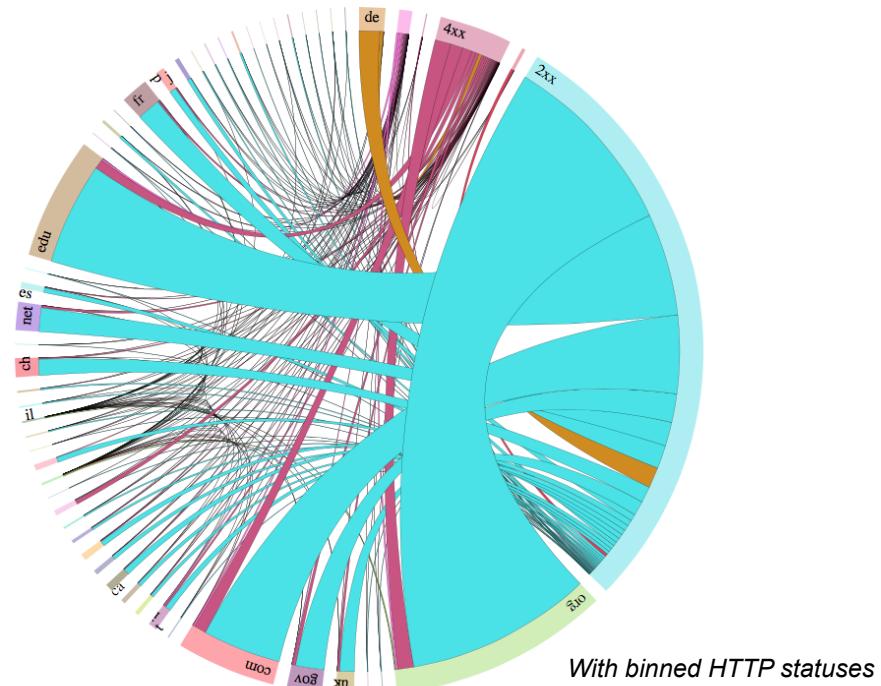
## PLOS top level domains

**Threshold is 100 occurrences**



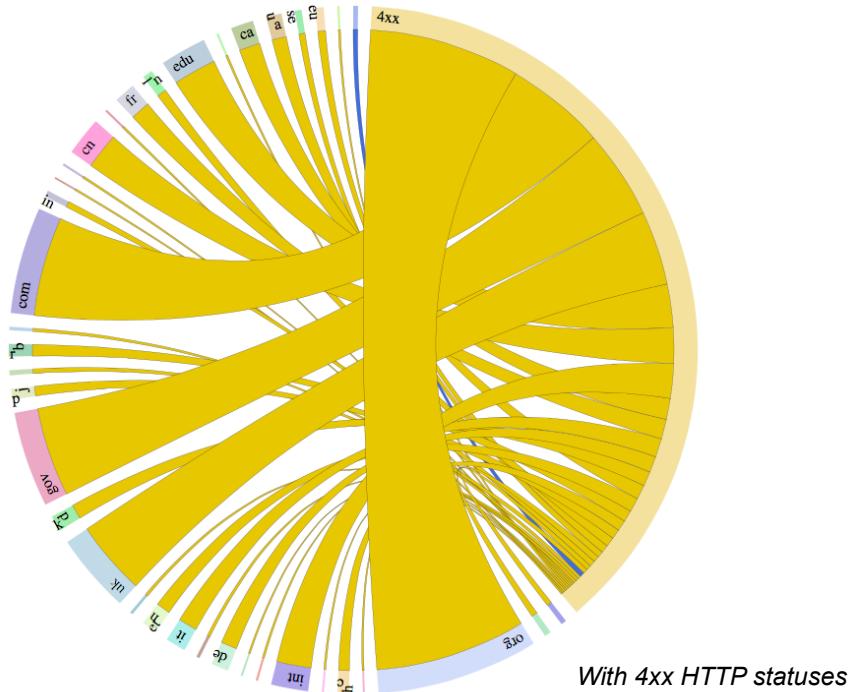
## arXiv top level domains

**Threshold is 100 occurrences**



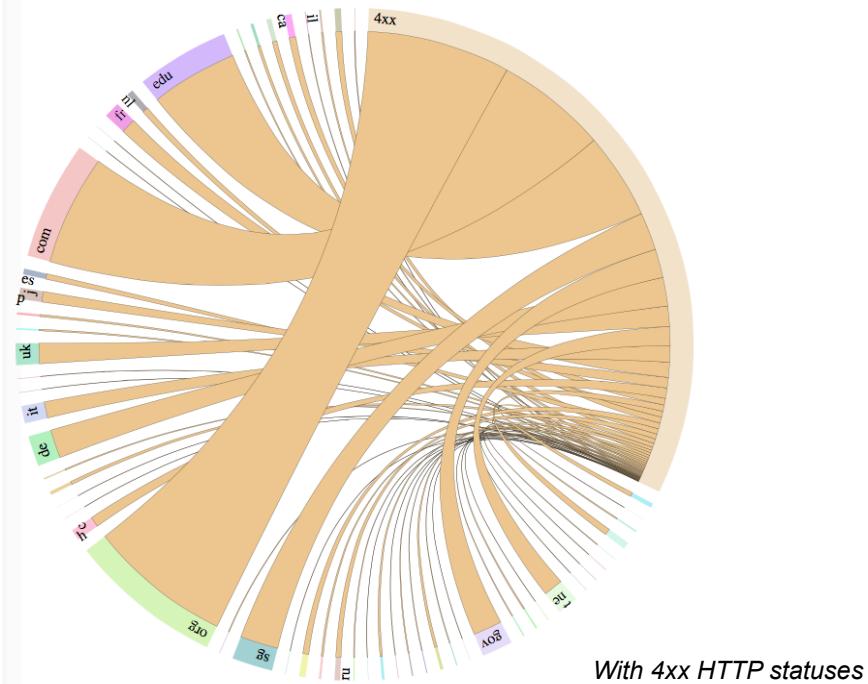
## PLOS top level domains

Threshold is 10 occurrences



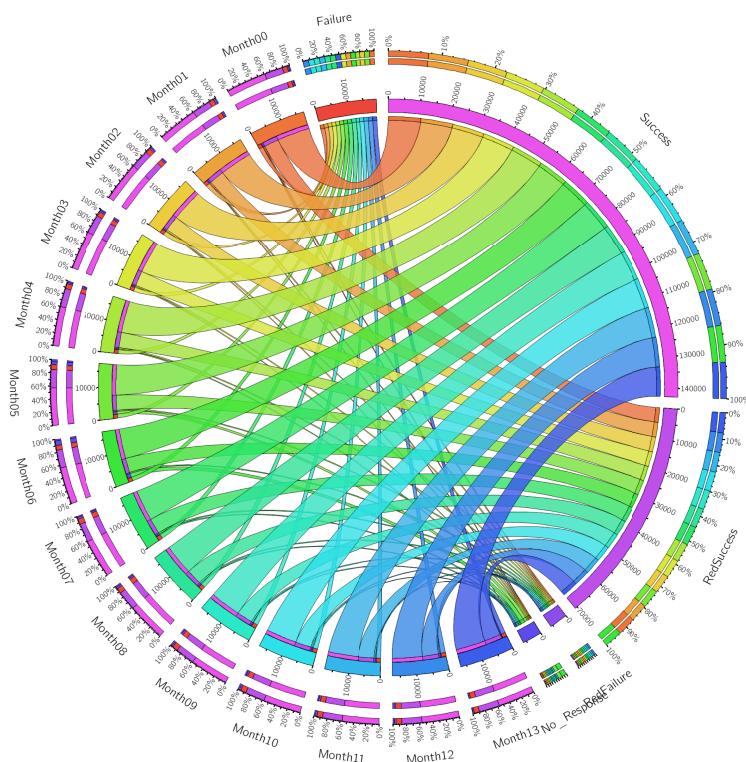
## arXiv top level domains

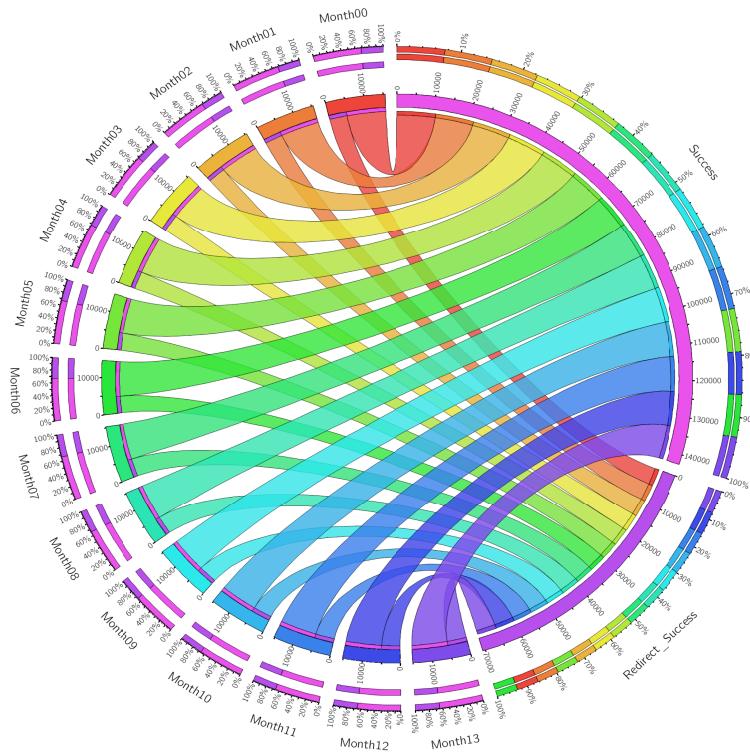
Threshold is 10 occurrences



# Circos diagram

- Circos diagrams show relationship of referenced URIs to four categories of status:
  - ☒ Success
  - ☒ Redirect – Success
  - ☒ Failure
  - ☒ Redirect – Failure
- Circos with only successful statuses





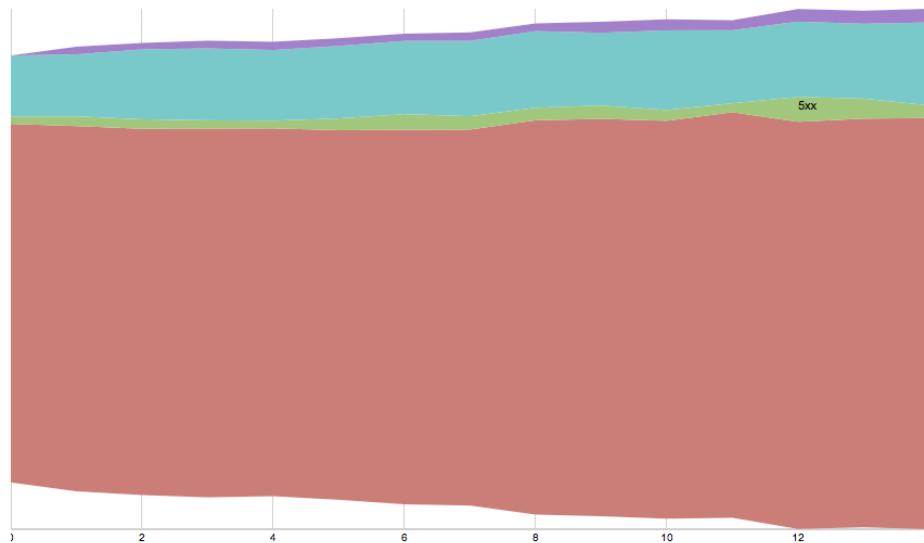
## Stream diagrams

- Stream diagrams of statuses by interval reveal trends with respect to the occurrence of statuses
  - All statuses of a given type are grouped together to form a band for a given month
  - The aggregate represents a complete band of reported statuses for the retrieval period
  - Subsequent bands flow left to right to show change over time
  - Again, status changes indicate the stability or volatility of a set of web-at-large links over time

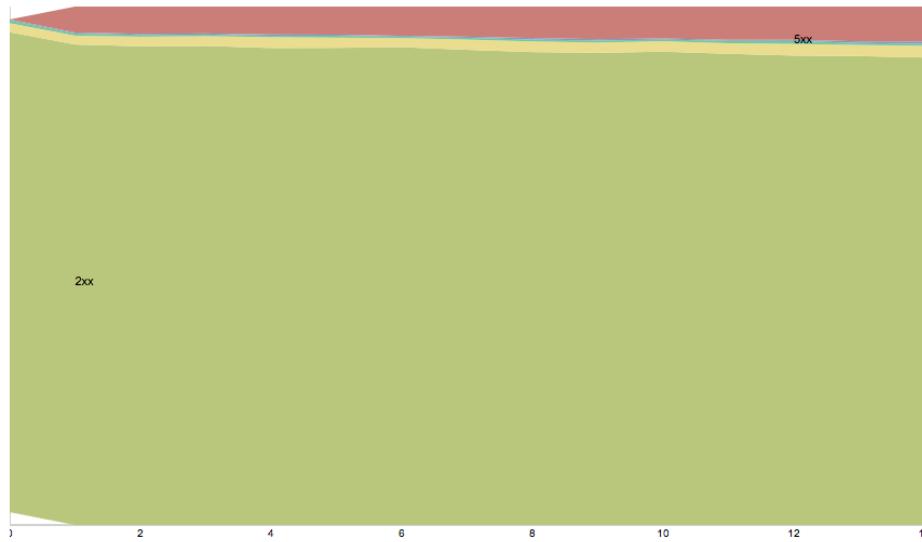
## arXiv, all statuses



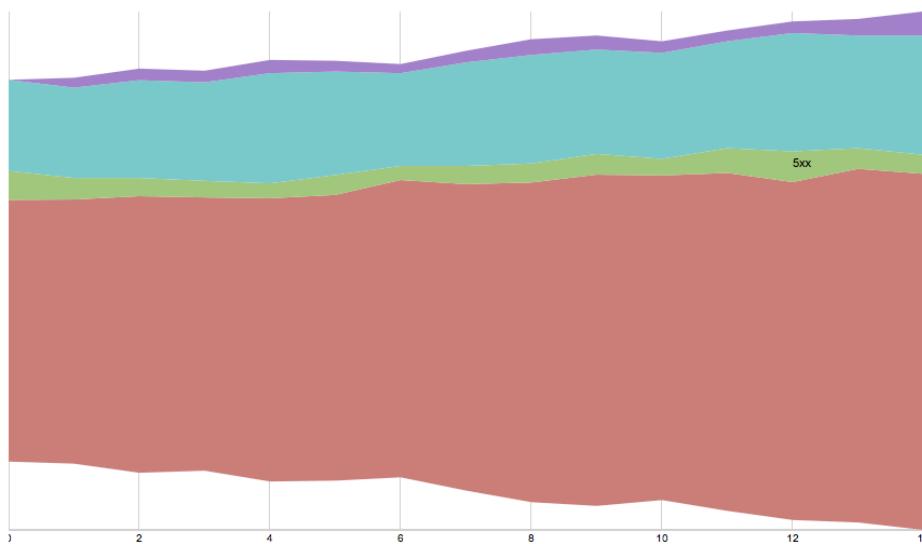
## arXiv, “bad” statuses



## PLOS, all statuses

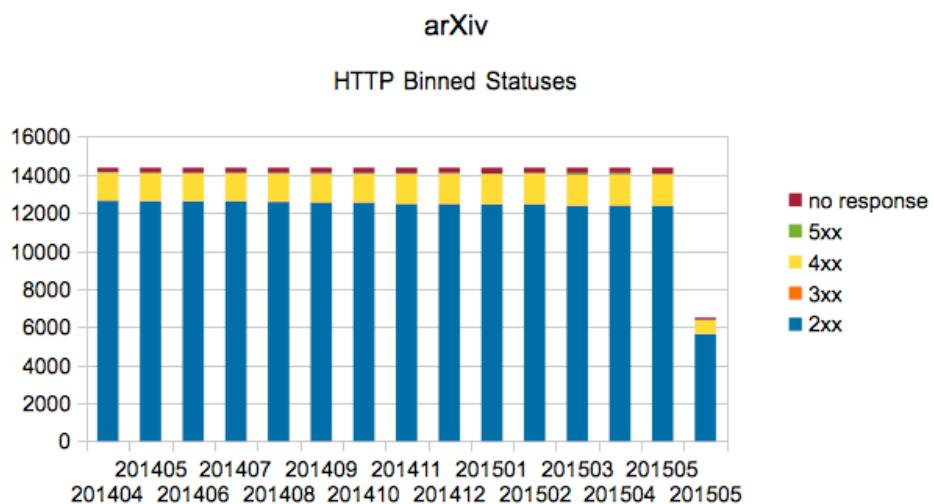


## PLOS, “bad” statuses



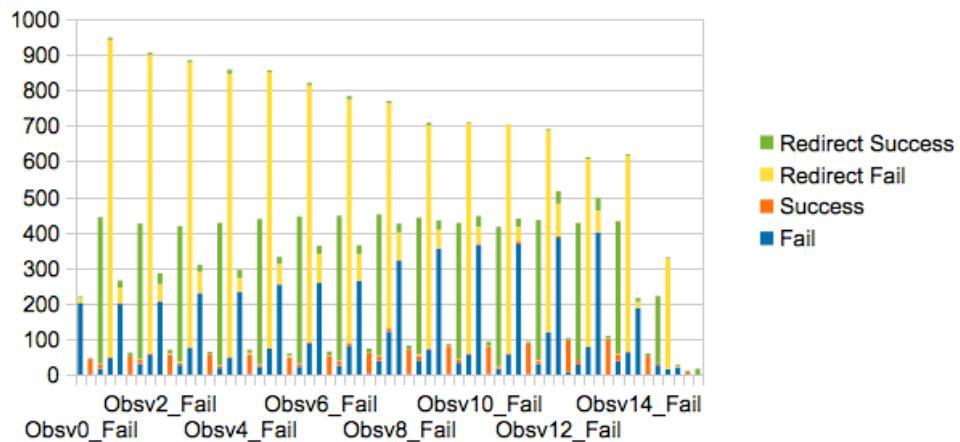
## Still more about HTTP statuses

- Sometimes a success or failed attempt to retrieve an object occurs only after one or more redirects
- It may be useful to
  - Examine status chain length and content
  - Longer status chains may hint that link rot will occur in the future even if the object was successfully retrieved this interval
  - Look at the first and last status in the chain to characterize whether this was a success or failure, indicative of link rot
  - Bar charts and sankey diagrams can help understand these possible trends

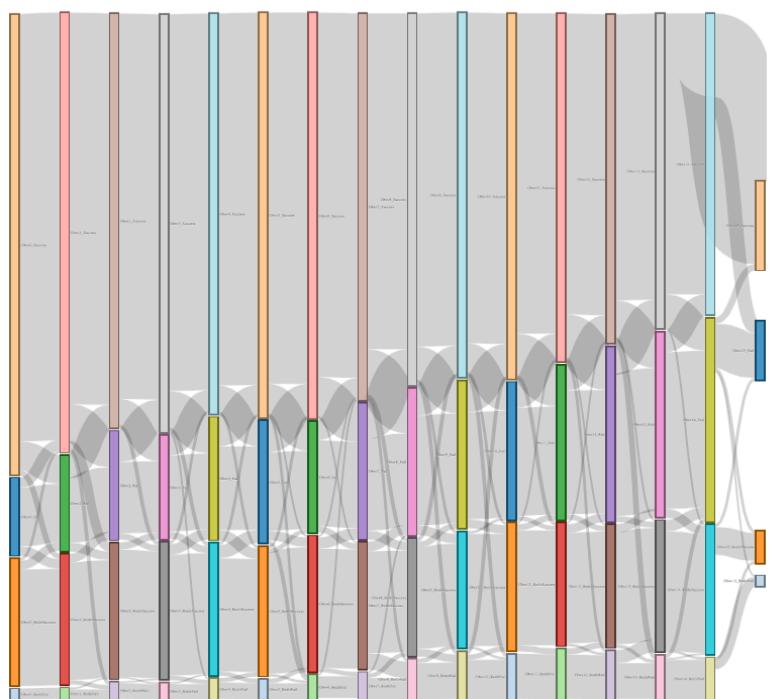


# Fail, Redirect-Fail, Success, Redirect-Success stacked by intervals

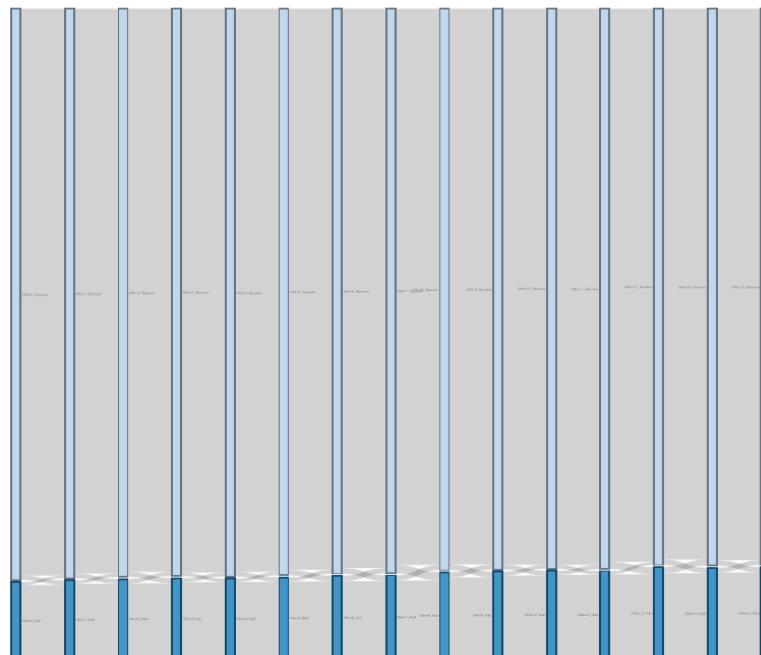
arXiv



## Successes, failures, with redirects



# Only success/fail statuses for arXiv



## Refs that change status at some interval

0	http://endo.endojournals.org/cgi/content/abstract/132/6/2614		200	301	302	302	200		text/html	45811	org		
1	http://endo.endojournals.org/cgi/content/abstract/132/6/2614	200		200	301	302	302	200	90.3796658767	77	text/html	46846	org
2	http://endo.endojournals.org/cgi/content/abstract/132/6/2614	200		200	301	302	302	200	90.284777831	71	text/html	46927	org
3	http://endo.endojournals.org/cgi/content/abstract/132/6/2614	200		200	301	302	302	200	90.1297140115	71	text/html	47024	org
4	http://endo.endojournals.org/cgi/content/abstract/132/6/2614	200		503	302	503		2.8024073079	incomparable	text/html	1397	org	
5	http://endo.endojournals.org/cgi/content/abstract/132/6/2614	503		200	301	302	302	200	89.791017923	68	text/html	47395	org
6	http://endo.endojournals.org/cgi/content/abstract/132/6/2614	200		200	301	302	302	200	81.1282894744	66	text/html	71890	org
7	http://endo.endojournals.org/cgi/content/abstract/132/6/2614	200		200	301	302	302	200	80.8311160244	66	text/html	71636	org
8	http://endo.endojournals.org/cgi/content/abstract/132/6/2614	200		200	301	302	302	200	80.5321845623	63	text/html	73495	org
9	http://endo.endojournals.org/cgi/content/abstract/132/6/2614	200		200	301	302	302	200	80.5321845623	63	text/html	73952	org
10	http://endo.endojournals.org/cgi/content/abstract/132/6/2614	200		200	301	302	302	200	80.4100148788	63	text/html	75110	org
11	http://endo.endojournals.org/cgi/content/abstract/132/6/2614	200		200	301	302	302	200	37.1635988601	61	text/html	72302	org
12	http://endo.endojournals.org/cgi/content/abstract/132/6/2614	200	none					n/a	failed	0	org		
13	http://endo.endojournals.org/cgi/content/abstract/132/6/2614			200	301	302	302	200	37.7072221217	65	text/html	71052	org
14	http://endo.endojournals.org/cgi/content/abstract/132/6/2614	200		200	301	302	302	200	78.908547846	65	text/html	76417	org

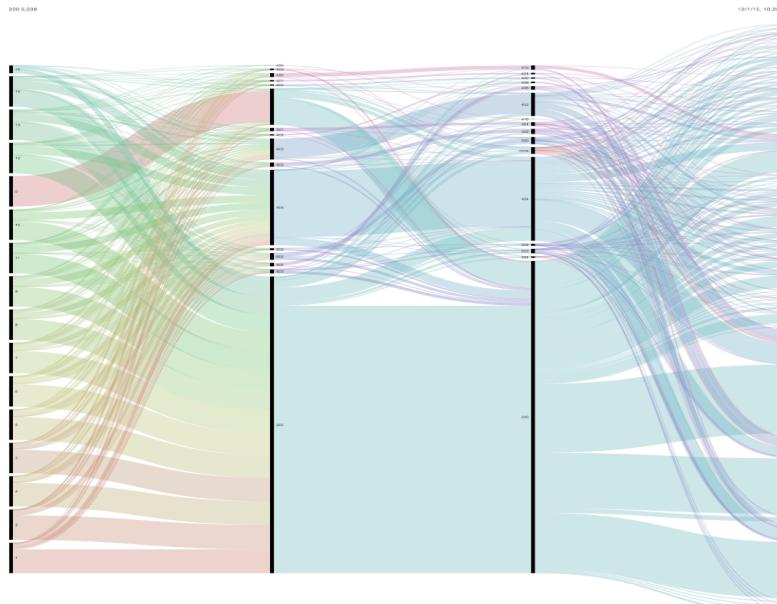
0	http://euils.ncbi.nlm.nih.gov/ventrez/euils/elinc_fcg?fbFrom=pubmed&id=1730180&retmode=rfr&cmd=prtnks	200	302 301 301 301 301 200	text/html	4110 gov
1	http://euils.ncbi.nlm.nih.gov/ventrez/euils/elinc_fcg?fbFrom=pubmed&id=1730180&retmode=rfr&cmd=prtnks	200	302 301 301 301 301 200	95.6512004697	96.xhtml
2	http://euils.ncbi.nlm.nih.gov/ventrez/euils/elinc_fcg?fbFrom=pubmed&id=1730180&retmode=rfr&cmd=prtnks	200	302 301 301 301 301 200	95.4551942199	96.xhtml
3	http://euils.ncbi.nlm.nih.gov/ventrez/euils/elinc_fcg?fbFrom=pubmed&id=1730180&retmode=rfr&cmd=prtnks	200	302 301 301 301 301 200	95.1081448696	96.xhtml
4	http://euils.ncbi.nlm.nih.gov/ventrez/euils/elinc_fcg?fbFrom=pubmed&id=1730180&retmode=rfr&cmd=prtnks	200	302 301 301 301 301 200	5.8739140255 incomparable	text/xhtml
5	http://euils.ncbi.nlm.nih.gov/ventrez/euils/elinc_fcg?fbFrom=pubmed&id=1730180&retmode=rfr&cmd=prtnks	200	302 301 301 301 301 200	80.8437653193	80.xhtml
6	http://euils.ncbi.nlm.nih.gov/ventrez/euils/elinc_fcg?fbFrom=pubmed&id=1730180&retmode=rfr&cmd=prtnks	200	302 301 301 301 301 200	80.8437653193	80.xhtml
7	http://euils.ncbi.nlm.nih.gov/ventrez/euils/elinc_fcg?fbFrom=pubmed&id=1730180&retmode=rfr&cmd=prtnks	200	302 301 301 301 301 200	77.5871293726	69.xhtml
8	http://euils.ncbi.nlm.nih.gov/ventrez/euils/elinc_fcg?fbFrom=pubmed&id=1730180&retmode=rfr&cmd=prtnks	200	302 301 301 301 301 200	77.5815113343	69.xhtml
9	http://euils.ncbi.nlm.nih.gov/ventrez/euils/elinc_fcg?fbFrom=pubmed&id=1730180&retmode=rfr&cmd=prtnks	200	302 301 301 301 301 200	77.5523042744	66.xhtml
10	http://euils.ncbi.nlm.nih.gov/ventrez/euils/elinc_fcg?fbFrom=pubmed&id=1730180&retmode=rfr&cmd=prtnks	200	302 301 301 301 301 200	76.2236407524	61.xhtml
11	http://euils.ncbi.nlm.nih.gov/ventrez/euils/elinc_fcg?fbFrom=pubmed&id=1730180&retmode=rfr&cmd=prtnks	200	302 301 301 301 301 200	72.1601626171	58.xhtml
12	http://euils.ncbi.nlm.nih.gov/ventrez/euils/elinc_fcg?fbFrom=pubmed&id=1730180&retmode=rfr&cmd=prtnks	200	302 301 301 301 301 200	72.1198027825	54.xhtml
13	http://euils.ncbi.nlm.nih.gov/ventrez/euils/elinc_fcg?fbFrom=pubmed&id=1730180&retmode=rfr&cmd=prtnks	200	404 302 301 301 404	5.7940534260 incomparable	text/xhtml
14	http://euils.ncbi.nlm.nih.gov/ventrez/euils/elinc_fcg?fbFrom=pubmed&id=1730180&retmode=rfr&cmd=prtnks	404	200 302 301 301 301 200	71.8744148764	0.xhtml

## Refs that change status at some interval

0 http://dbpartners.stanford.edu/RegaSubtyping/		200	200			text/html	7826.edu
1 http://dbpartners.stanford.edu/RegaSubtyping/	200	200	200	100	100	text/html	7826.edu
2 http://dbpartners.stanford.edu/RegaSubtyping/	200	200	200	100	100	text/html	7826.edu
3 http://dbpartners.stanford.edu/RegaSubtyping/	200	200	200	100	100	text/html	7826.edu
4 http://dbpartners.stanford.edu/RegaSubtyping/	200	403	403	0 incomparable	0 incomparable	text/html	315.edu
5 http://dbpartners.stanford.edu/RegaSubtyping/	403	403	403	0 incomparable	0 incomparable	text/html	315.edu
6 http://dbpartners.stanford.edu/RegaSubtyping/	403	403	403	0 incomparable	0 incomparable	text/html	315.edu
7 http://dbpartners.stanford.edu/RegaSubtyping/	403	403	403	0 incomparable	0 incomparable	text/html	315.edu
8 http://dbpartners.stanford.edu/RegaSubtyping/	403	200	200	7.6037510794	incomparable	text/html	1690.edu
9 http://dbpartners.stanford.edu/RegaSubtyping/	200	200	200	7.6037510794	incomparable	text/html	1690.edu
10 http://dbpartners.stanford.edu/RegaSubtyping/	200	200	200	7.6037510794	incomparable	text/html	1690.edu
11 http://dbpartners.stanford.edu/RegaSubtyping/	200	200	200	7.6037510794	incomparable	text/html	1690.edu
12 http://dbpartners.stanford.edu/RegaSubtyping/	200	200	200	6.1813916982	incomparable	text/html	1309.edu
13 http://dbpartners.stanford.edu/RegaSubtyping/	200	200	200	6.1813916982	incomparable	text/html	1309.edu
14 http://dbpartners.stanford.edu/RegaSubtyping/	200	200	200	6.1813916982	incomparable	text/html	1309.edu

0 http://scott.sherrillmix.com/R-GMT_HexPlot.php		200	200			text/html	3774.com
1 http://scott.sherrillmix.com/R-GMT_HexPlot.php	200	406	406	8.3440653028	incomparable	text/html	387.com
2 http://scott.sherrillmix.com/R-GMT_HexPlot.php	406	406	406	8.3440653028	incomparable	text/html	387.com
3 http://scott.sherrillmix.com/R-GMT_HexPlot.php	406	200	200	99.9999986584	100	text/html	3774.com
4 http://scott.sherrillmix.com/R-GMT_HexPlot.php	200	200	200	99.9999986584	100	text/html	3774.com
5 http://scott.sherrillmix.com/R-GMT_HexPlot.php	200	200	200	99.9999986584	100	text/html	3774.com
6 http://scott.sherrillmix.com/R-GMT_HexPlot.php	200	200	200	99.9999986584	100	text/html	3774.com
7 http://scott.sherrillmix.com/R-GMT_HexPlot.php	200	200	200	99.9999986584	100	text/html	3774.com
8 http://scott.sherrillmix.com/R-GMT_HexPlot.php	200	200	200	99.9999986584	100	text/html	3774.com
9 http://scott.sherrillmix.com/R-GMT_HexPlot.php	200	200	200	99.9999986584	100	text/html	3774.com
10 http://scott.sherrillmix.com/R-GMT_HexPlot.php	200	200	200	99.9999986584	100	text/html	3774.com
11 http://scott.sherrillmix.com/R-GMT_HexPlot.php	200	200	200	99.9999986584	100	text/html	3774.com
12 http://scott.sherrillmix.com/R-GMT_HexPlot.php	200	200	200	99.9999986584	100	text/html	3774.com
13 http://scott.sherrillmix.com/R-GMT_HexPlot.php	200	200	200	99.9999986584	100	text/html	3774.com
14 http://scott.sherrillmix.com/R-GMT_HexPlot.php	200	200	200	99.9999986584	100	text/html	3774.com
15 http://scott.sherrillmix.com/R-GMT_HexPlot.php	200	200	200	99.9999986584	100	text/html	3774.com

## Intervals, previous status, current status, and top level domain (PLOS)

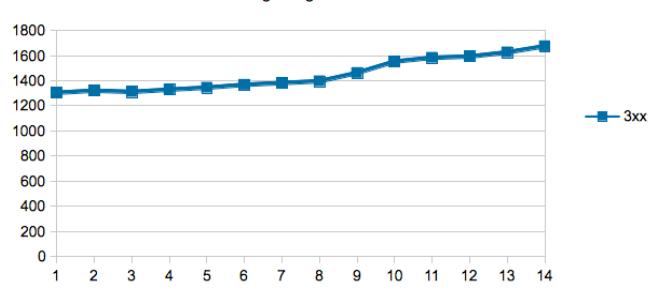


# HTTP statuses and link rot

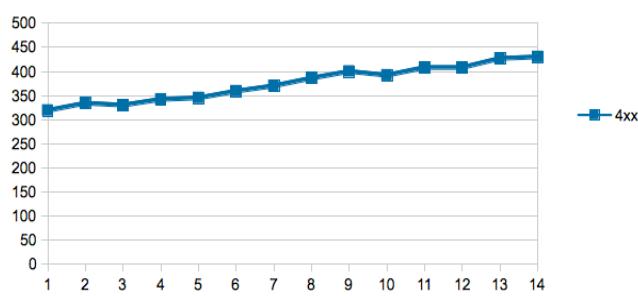
- Is it the case that for this sample that link rot increases with the passage of time?
  - During the course of this experiment, the plos sample experienced a 34% increase in 4xx final statuses and a 28% increase in status chains that began with 3xx (indicating redirection).
  - arXiv exhibited a 12% increase in 4xx final statuses, and an 11% increase in status chains that began with 3xx.
- So the answer would appear to be **yes**.

## PLOS

PLOS  
3xx beginning of status chain



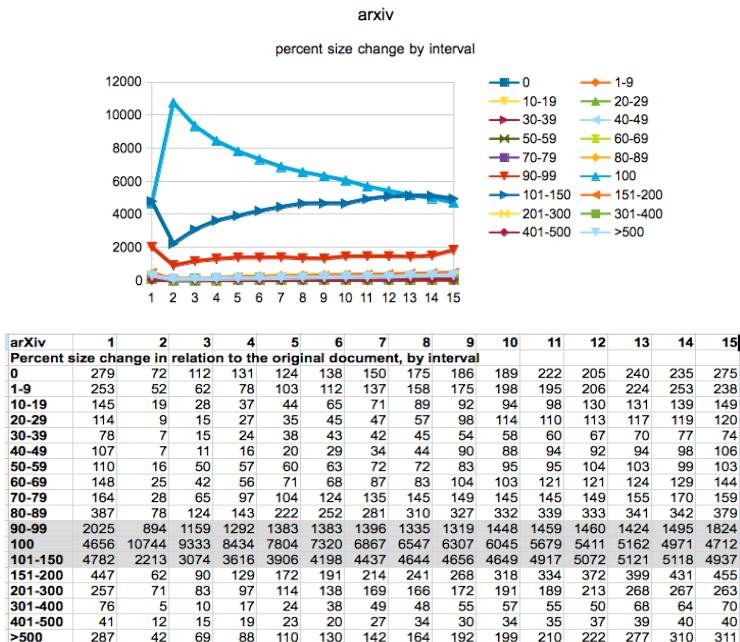
PLOS  
4xx last status by interval

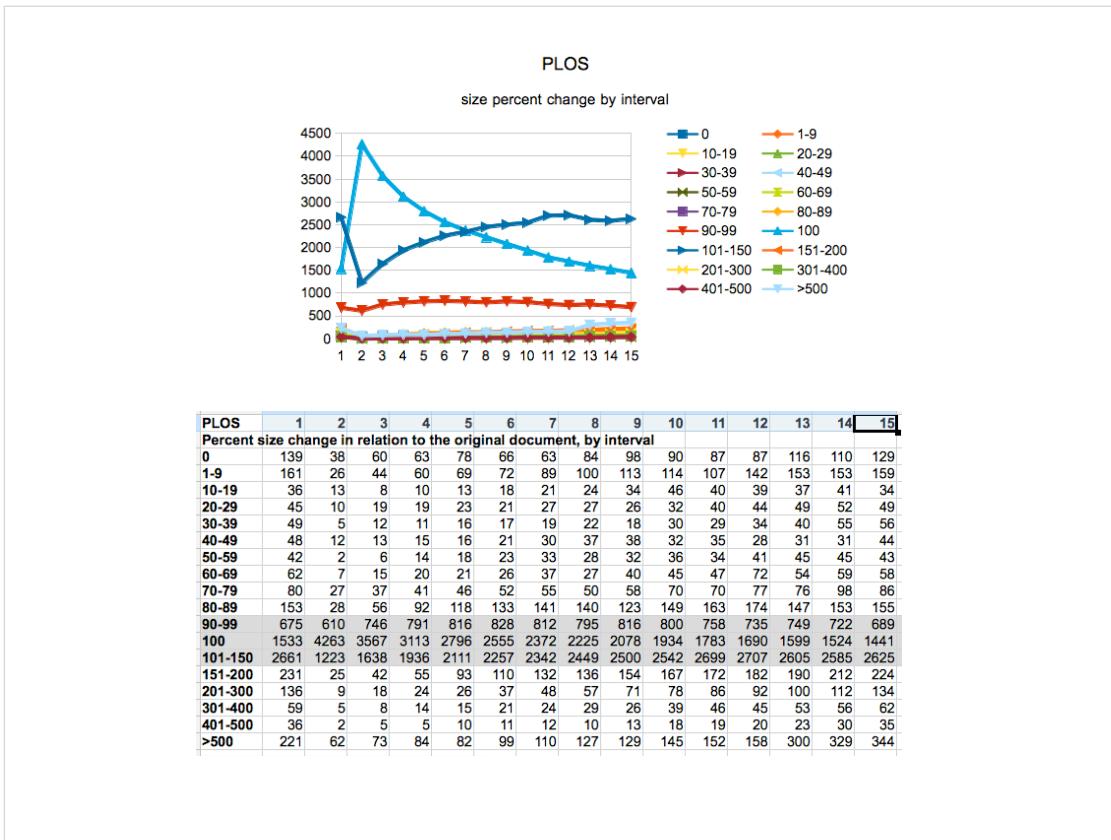


# Size (content length) changes

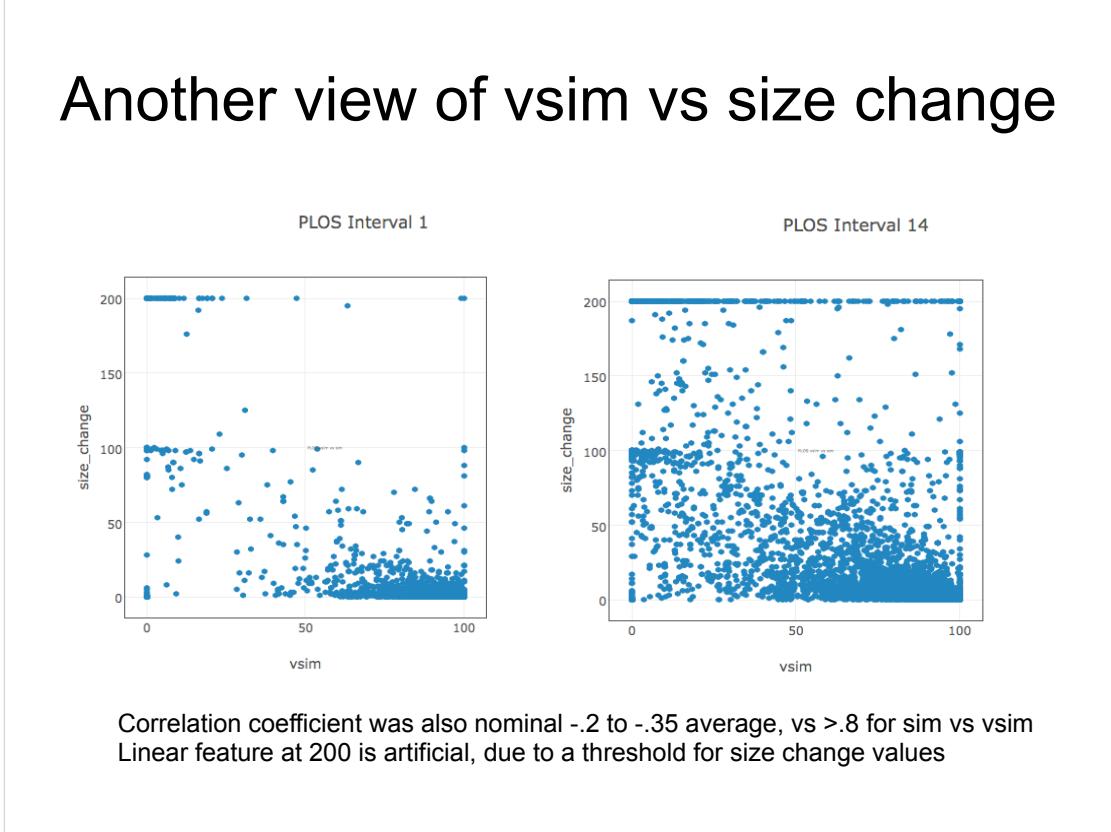
- Size changes hint that content drift may have occurred
  - Do document sizes vary over time?
  - If so, by how much?
  - To what extent do they vary?
  - What might this mean?

A bit confusing... 100% actually indicates that a copy is exactly the same size as the original





## Another view of vsim vs size change



# Content size discussion

Size changes mean something...

- Quantitatively, we want to be able to say that a document changed by some meaningful fraction (half as large as it originally was, doubled in size, etc)
- We might also investigate whether size change correlates with a change in similarity (meaning) as compared to the original document

## Size changes and content drift

- Is it the case that for this sample that content drift increases with the passage of time?
  - Based only on size change, it appears that it does
  - The most common change was for content size to increase, with content growth most commonly falling into the 1 – 50% increase range
  - Also the number of documents that were exactly the same size as the original plummeted over time
    - PLOS by 66%
    - arXiv by 56%

## Content Similarity

- One approach: Binary hash of content, compare subsequent interval binary hash to original harvested web-at-large resource using python library ssdeep
  - Byte level comparison
  - Works for any mime type
  - Sometimes small changes to a file can cause hashes to appear more different than they really are – especially true for textual content

## Feature vectors

- Between 97 and 99% of the objects harvested and monitored for this experiment were of mime type text/html
- Textual content can be represented as feature vectors and compared using various measures (e.g. Cosine, Euclidean distance)
- Feature vectors are content based so they are less vulnerable to changes in content formatting or minor rearrangement of content

## Feature Vector construction

- First, all HTML, CSS and Javascript content is stripped from document using the python library *BeautifulSoup* with the lxml parser
- Next a stop word list is consulted, so frequently occurring words are omitted
- Next, Porter stemmer is employed, thus reducing some English words to a common root
- Next, a vector is created where each word represents a dimension and the frequency of the word is a point on that axis

## Comparing feature vectors

- Comparison is always between the original content found at the beginning of the experiment for the Web at large reference, and a subsequent interval's instance of that content
- The representation for each of these two documents is normalized (same number of slots in the matrix even if one document has count zero for a given term), so they have the same dimensions
- We then use cosine similarity to compare the two documents – which looks at the angle between the two vectors rather than the distance between n-dimensional endpoints
- This normalizes all terms, thus frequency of term occurrence is irrelevant (could argue less precise)

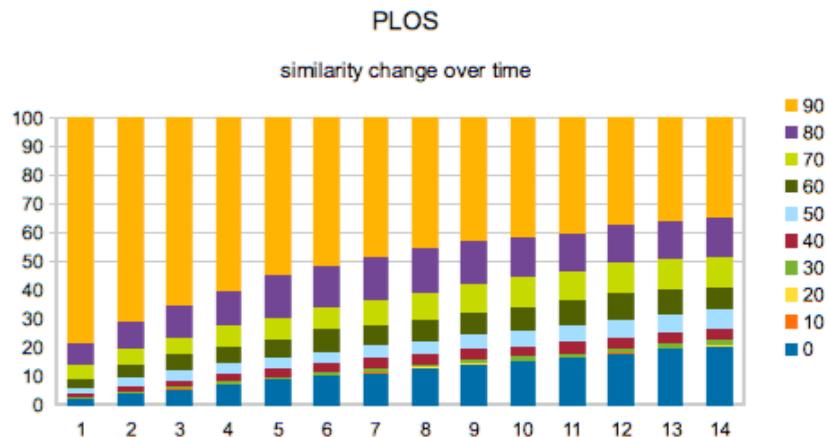
## Content similarity issues

- Reformatting and rearranging content can affect the similarity hash but not necessarily the information conveyed in the document
- Rearranging the same content might effectively change the relative information content as compared to a previous copy, or to the original
- Rearranging the content or reformatting it may not change what the referenced document is about

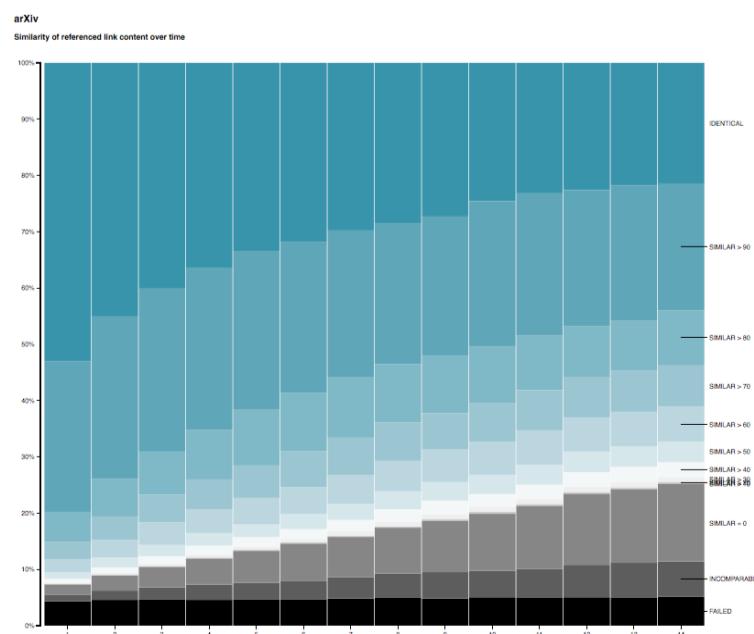
## Stacked similarity plots

- Stacked bar charts that compare month over month similarity hash scores for items
- These charts show that content drift is clearly increasing as time passes

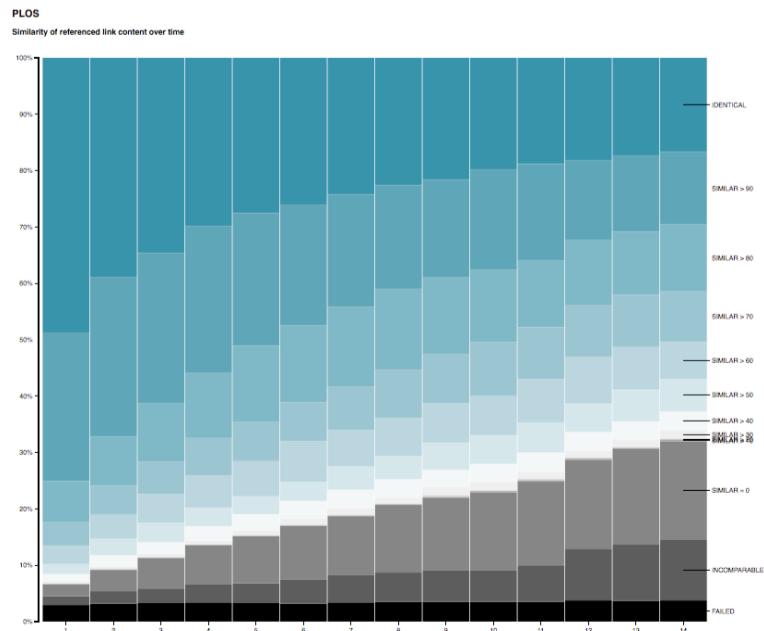
## PLOS stacked bar chart



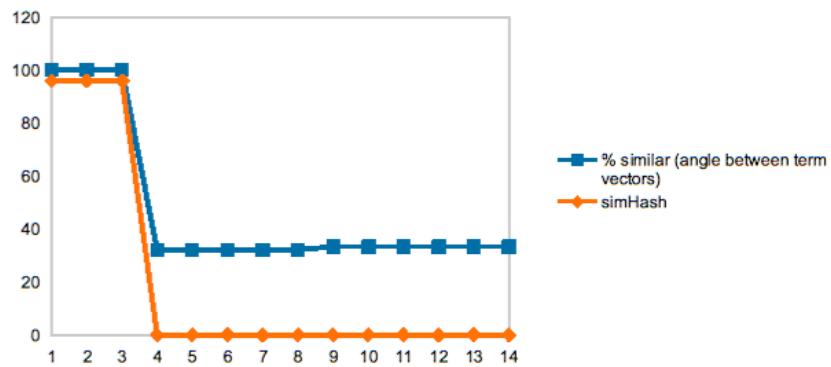
arXiv



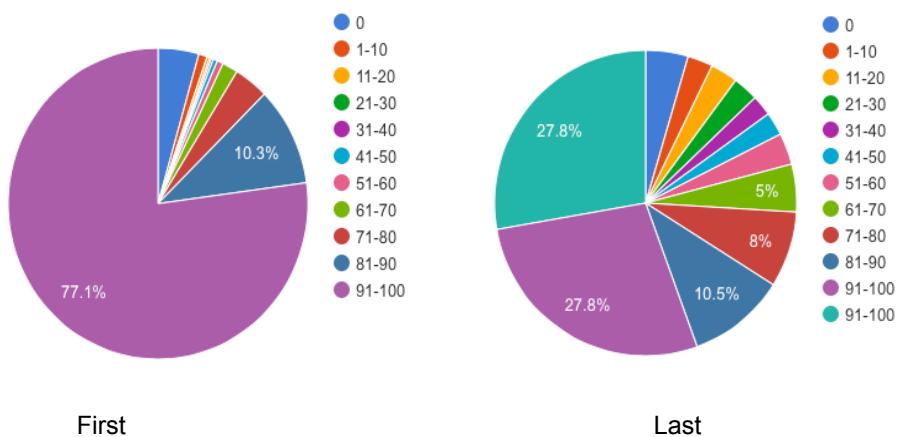
# PLOS



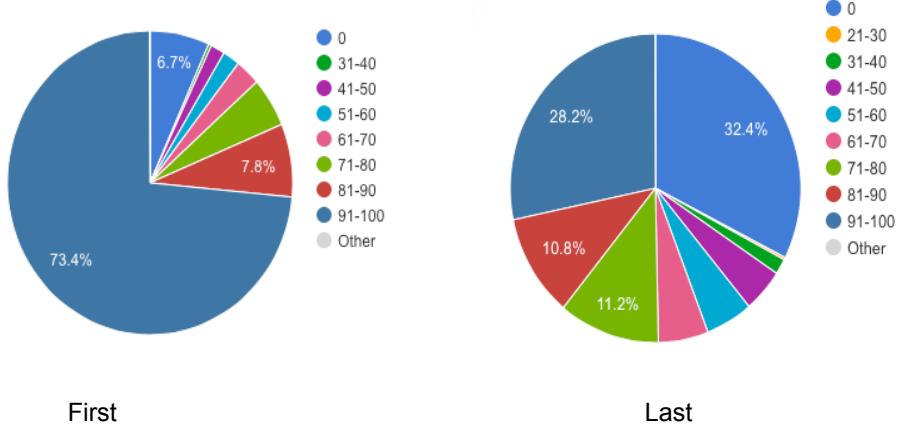
Similarity hash is far less forgiving,  
which can be deceiving with text

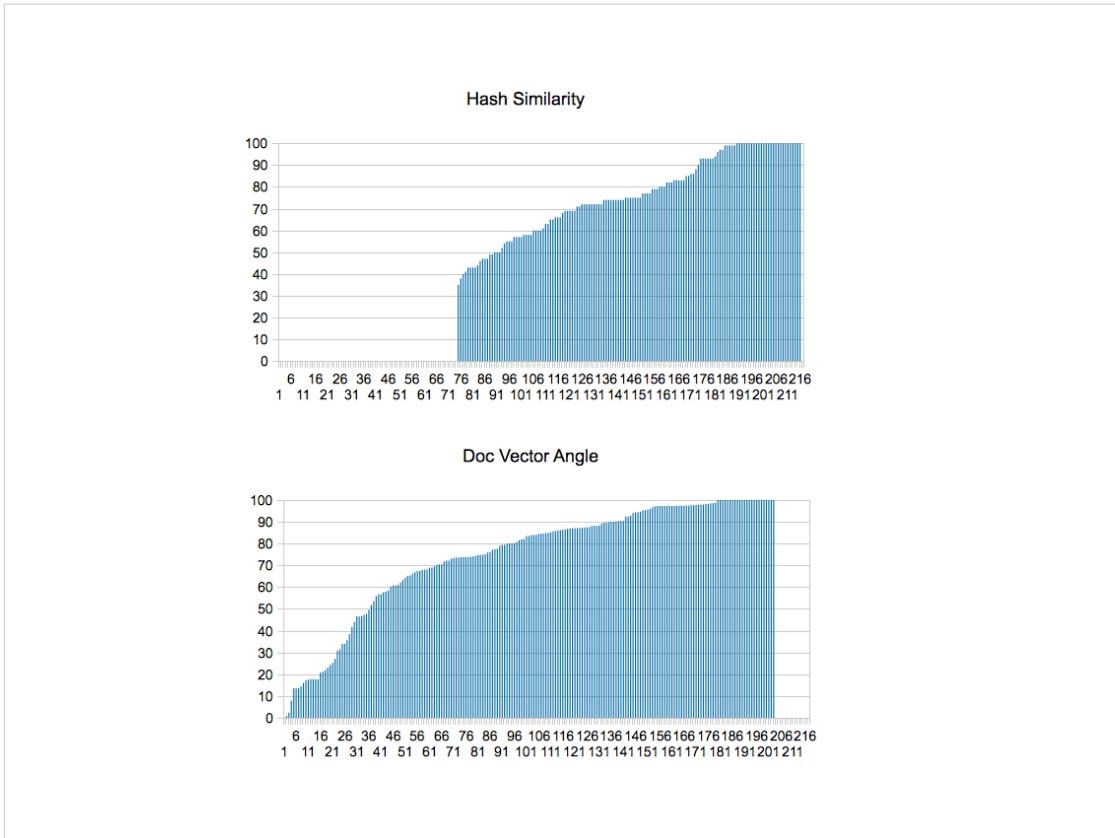


## VSIM PLOS first vs last interval



## Hash similarity PLOS first vs last interval





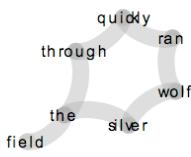
## Comparing similarity measures

- Hashes are sensitive to content rearrangement
- Feature vectors are not
- Hashes are fairly unforgiving with respect to content change, large or small
- Feature vectors can illustrate more subtle content drift but won't notice content rearrangement
- Using graph document models and graph matching to compare content would reveal more subtle changes in textual documents, but this would be computationally intensive

# Graph Model, Graph Edit Distance

- Really small GXL document graph model example

```
<!DOCTYPE gxl SYSTEM "http://www.gupro.de/GXL/gxl-1.0.dtd">
<gxl xmlns:xlink="http://www.w3.org/1999/xlink">
<graph id="SEN_004" edgeids="false" edgemode="directed">
<node id="_0"><attr name="word"><string>the</string></attr></node>
<node id="_1"><attr name="word"><string>silver</string></attr></node>
<node id="_2"><attr name="word"><string>wolf</string></attr></node>
<node id="_3"><attr name="word"><string>ran</string></attr></node>
<node id="_4"><attr name="word"><string>quickly</string></attr></node>
<node id="_5"><attr name="word"><string>through</string></attr></node>
<node id="_6"><attr name="word"><string>field</string></attr></node>
<edge from="_0" to="_1"/>
<edge from="_1" to="_2"/>
<edge from="_2" to="_3"/>
<edge from="_3" to="_4"/>
<edge from="_4" to="_5"/>
<edge from="_5" to="_0"/>
<edge from="_0" to="_6"/>
</graph></gxl>
```



- Output of a Graph Edit Distance run

Graph edit distance procedure: Beam

Edge mode: directed  
Cost for node deletion/insertion: 3.0  
Cost for edge deletion/insertion: 3.0

Alpha weighting factor between node and edge costs: 0.8

Node attribute 0: word; Cost function: discrete; mu = 0  
nu = 2; Soft factor: 1.0

No attributes for edges defined

Individual node costs are added  
Individual edge costs are added  
(Combined node cost)^(1/1.0)  
(Combined edge cost)^(1/1.0)

\*\*\* The distance matrix \*\*\*

1.20000,3.40000,2.80000,3.40000	3.40000,1.20000,5.00000,1.20000	2.80000,5.00000,1.20000,5.00000	3.40000,1.20000,5.00000,1.20000
---------------------------------	---------------------------------	---------------------------------	---------------------------------

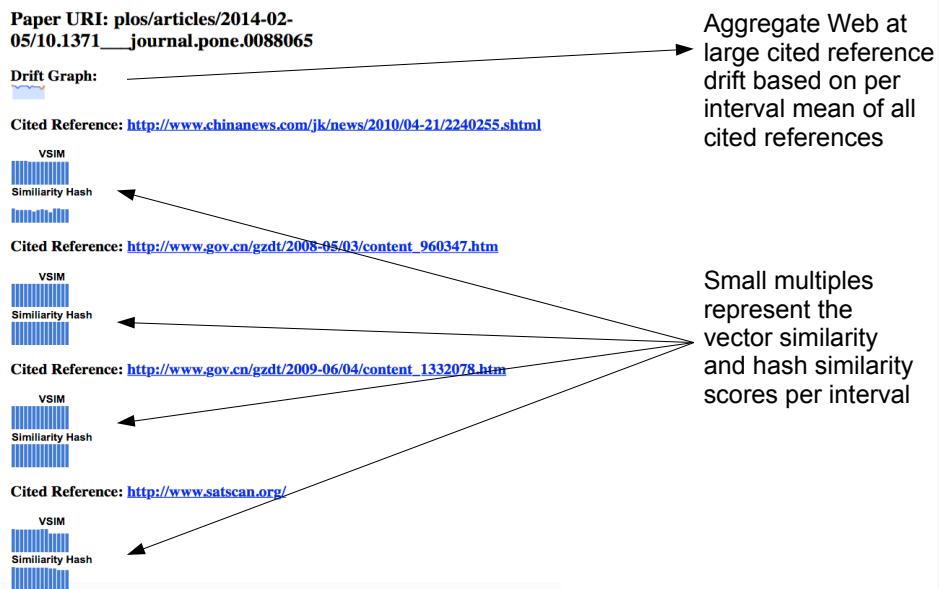
## Use entropy to compare original to harvested copies

- Entropy as defined in information theory: the amount of surprise, the probability of the occurrence of a given bit, byte, character, token, etc.
- Use the original document as a model
- Use English string tokenizer, as with feature vectors, to create a granular pair of data streams to compare
- Low entropy, little change, high entropy, significant content changes

## Another way to look at content drift

- The web-at-large links associated with a given paper represent a portion of its context
- So, what is happening to the aggregate context for a given paper?
- Can we visualize the aggregate content drift?
- Can we calculate some kind of meaningful aggregate drift score for a paper?

## Citing paper exploring aggregate drift



# A paper that saw aggregate drift

Paper URI: plos/articles/2014-02-04/10.1371/journal.pone.0087534

Drift Graph:



Cited Reference: <http://crocodilian.com/cnbc/cnbc.html>



Cited Reference: <http://www.lirmm.fr/~caraux/Bioinformatics/NegativeMultinomial/>



Cited Reference: <http://www.ncbi.nlm.nih.gov>



Cited Reference: <http://www.r-project.org>



Drift Graph:



## NCBI Interval 1 and Interval 15

# R-Project Interval 1 and 15

## The R Project for Statistical Computing

### Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and Mac OS. To [download R](#), please choose your preferred CRAN mirror;

If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

### News

- R 3.2.1 (*World-Famous Astronaut*) pre-release versions will appear starting June 8. Final release is scheduled for 2015-06-18.
- R version 3.2.0 (Full of Ingredients) has been released on 2015-04-16.
- R version 3.1.3 (Smooth Sidewalk) has been released on 2015-03-09.
- The R Journal Volume 6(2) is available.

• userR! 2015, will take place at the University of Aalborg, Denmark, June 30 - July 3, 2015.

• userR! 2014, took place at the University of California, Los Angeles, USA June 30 - July 3, 2014.

### Documentation

#### R Foundation

#### R Project

#### Documentation

#### Links

# COSMIC Interval 1 and 14

# Another example

Paper URI: [plos/articles/2014-02-05/10.1371/journal.pone.0088388](https://doi.org/10.1371/journal.pone.0088388)

Drift Graph:



Cited Reference: <http://www.ars.usda.gov/ba/fsrg>



Cited Reference: <http://www.bhf.org.uk>



Cited Reference: <http://www.cdc.gov/nchs/nhanes.htm>



Cited Reference: [http://www.cdc.gov/nchs/nhanes2003-2004/analytical\\_guidelines.htm](http://www.cdc.gov/nchs/nhanes2003-2004/analytical_guidelines.htm)

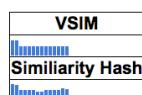


# Drifting reference

URI <http://www.cdc.gov/nchs/nhanes.htm>

Interval	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
VSIM	80.5495328035	78.688475321	48.1812683986	48.4617977811	46.6842573389	47.7940530164	47.1557947414	45.721880893	47.996968254	46.399786048	48.7327838301	47.133548909	46.6950464289	47.1437659677	
Similarity Hash	82	79	54	54	54	50	44	44	54	50	50	50	60	60	57

Visualizations of...



# But not necessarily

The left screenshot shows the main page of the National Health and Nutrition Examination Survey (NHANES) at [www.cdc.gov/nhanes/](http://www.cdc.gov/nhanes/). It features a navigation bar with links like 'CDC Home', 'About NHANES', 'What's New', 'Questionnaires, Datasets, and Related Documentation', 'Tutorials', 'Protocol Guidelines', 'Survey Results and Products', and 'Listserv'. Below the navigation is a large section titled 'National Health and Nutrition Examination Survey' with a sub-section 'What's New'. This section includes a 'Selected Participants' box asking if you've been selected to participate in the survey, and a 'Information for Health Professionals' box with a link to the 'NHANES Data and Documentation' page.

The right screenshot shows the same page but with different content. The 'What's New' section has been replaced by a 'Proposed Guidelines' box. The 'Information for Health Professionals' box has been replaced by a 'Proposed Guidelines' box, which includes a link to the 'NHANES Data and Documentation' page.

# An arXiv example

Paper URI: [arxiv/articles/2014-02-10/1402.1693.pdf](http://arxiv.org/abs/2014-02-10/1402.1693.pdf)

Drift Graph:



Cited Reference: <http://en.wikipedia.org/wiki/Radio-frequency>



Cited Reference: <http://www.jctjournal.org>



Cited Reference: <http://www.rfidreader.info>



Cited Reference: <http://www.webopedia.com/TERM/R/RFID.html>

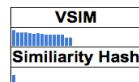


# Drift over intervals

URI <http://www.webopedia.com/TERM/R/RFID.html>

Interval	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
VSIM	94.6808587881	81.4394182743	60.9194831356	80.9752024027	74.3773192548	72.2005775768	72.7874750684	72.6705682689	88.397267836	87.9518357817	87.5099360714	66.6890930887	88.3990611449	48.5046277853	48.4777986223	
Similarity Hash	60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Visualizations of...



## A definition of RFID rewritten, expanded

The image displays two versions of the same web page from webopedia.com. The left version is the original, and the right version is a modified version with expanded text and additional sections.

**Original (Left):**

- Related Terms:** RFID reader, RFID tag, passive tag, active tag.
- Expert Articles:** RFID, RFID Connection to CRM, RFID ROI Calculator.
- WE RECOMMEND:** Top Tech Trends for 2014, Enterprise Mobility to Dominate IT Agenda.
- THE DIFFERENCE BETWEEN RFID AND BAR CODES:** One of the key differences between RFID and bar code technology is that RFID eliminates the need for line-of-sight reading that bar coding depends on. Also, RFID scanning can be done at greater distances than barcode scanning. High-frequency RFID systems (800 MHz to 950 MHz and 2.4 GHz to 2.5 GHz) offer transmission ranges of more than 90 feet, although wavelengths in the 2.4 GHz range are absorbed by water (the human body) and therefore have limitations.
- TECH RESOURCES FROM OUR PARTNERS:** A section featuring links to various tech resources from partners like Microsoft, Dell, and others.

**Modified (Right):**

- Related Terms:** (pronounced as separate letters) Short for radio frequency identification, a technology similar in theory to bar code identification, but using radio waves instead of lines of light or patterns of ink to transmit signals. An RFID system consists of an antenna and a transponder, or tag, which is attached to a product. An RF system uses the RF portion of the electromagnetic spectrum to transmit signals.
- RFID Systems:** RFID systems consist of an antenna and a transceiver, which read the radio frequency and transfer the information to a processing device, and a transponder, or tag, which is an integrated circuit containing the RF circuitry and information to be transmitted.
- Difference Between RFID and Bar Codes:** One of the key differences between RFID and bar code technology is that RFID eliminates the need for line-of-sight reading that bar coding depends on. Also, RFID scanning can be done at greater distances than barcode scanning. High-frequency RFID systems (800 MHz to 950 MHz and 2.4 GHz to 2.5 GHz) offer transmission ranges of more than 90 feet, although wavelengths in the 2.4 GHz range are absorbed by water (the human body) and therefore have limitations.
- Apple Pay Promises to Strengthen Payment Security:** Despite recent security issues, Apple Pay and other competitive payment systems will be more secure than cards, even cards equipped with EMV chips. [Read More](#)
- The Great Data Storage Debate: Is Tape Dead?** Tape clearly is on the decline. But remember, tape can last a long time. [Read More](#)

# On the other hand...

[African Economic Outlook](#)

Home | Economic Outlook | In Depth | Statistical Annex | Country Notes | News & Events | Resources | About us

**Benin**

Share Print PDF

Benin's economy is slowly recovering after experiencing a difficult period in 2009 and 2010, growth is estimated at 3.5% in 2011 and 4.6% in 2012 and is projected to consolidate in 2013 and 2014.

To reach its growth targets, the country will have to step up reforms of the port of Cotonou as well as its public financial management and continue to improve its infrastructure and the improvement of the business climate to further develop the private sector.

Benin will also have to remove constraints weighing on the exploitation of its agricultural and mining sectors. The government has already taken some steps to address these issues, notably the country's deficiencies in the infrastructure and services needed for exploiting these resources.

**Overview**

Benin's economic activity seems to have begun to recover since 2011, after having come under severe pressure in 2009 and 2010 from the combined effects of the global economic crisis and the floods that hit the country. The growth rate of the real economy increased from 3.5% in 2011 to 3.8% in 2012 and is projected to reach 4.6% in 2013. The government has decided to invest more in its ports to revive agriculture and repair the infrastructure after the floods of 2010. The country has also implemented a fiscal policy that includes a reduction in the budget deficit and a significant increase of a share increase in January 2012 in the price of subsidized petrol called "kerosene". The economic output of Benin is expected to grow at 4.6% in 2012 and 4.9% in 2013, which is slightly above the results from the 2012/13 cotton season, but recovery in port activities.

As important growth factor will, nonetheless, be the maintenance of macroeconomic stability by enhancing the efficiency of the public sector and by an administrative modernization in 2013 and 2014. Benin is facing here a threefold objective: to further modularize its domestic resources, to improve its business climate and to diversify its economy by helping to improve the country's business climate in order to help develop the private sector. The government, which has decided to implement a fiscal policy that includes a reduction in the budget deficit and some corrective measures to offset the impact of the short term rise in prices likely to result from this policy, will have to continue to implement its fiscal policy and its structural reform plan. The government needs to maintain its efforts through its 2011-15 growth and poverty-reduction strategy (GPRS), which aims to reduce poverty by 2015 and to achieve the United Nations' Millennium Development Goals (MDGs) by 2015. More than 36% of Beninese population are still living below the poverty line.

Benin has strong agricultural potential, an opening to the sea and a small amount of raw materials (minerals, sand, granite and iron). Its limited exploitation of these assets has, however, prevented Benin from becoming a major player in the regional market. To fully exploit its natural resources and better management of its natural resources Benin still needs to overcome several structural constraints, namely the lack of infrastructure, the lack of skills and the lack of access to finance, energy, roads, infrastructure and services associated with the exploitation of these resources.

For development purposes, Benin faces three main challenges: first, to implement its strategic plan for the revival of the agricultural sector (PGRSA), which is expected to further develop the country's economy and to diversify its economy; second, to move Benin from being a French colony to becoming a logistic and export hub, in particular thanks to an improved infrastructure and transport services system.

Figure 1 Real GDP growth 2012 (Year)

[African Economic Outlook](#)

Home | Country Notes | West Africa | Benin

Authors: Daniel Nohy, El Hadji Fall

Download this full country note in PDF

Driven by agriculture and trade, real GDP growth is estimated at 5.0% in 2013, down from 5.4% in 2012.

Reforms have continued in public finances and in the port sector, but a clear national strategy remains to be defined to support the private sector.

Global value chains (GVCs) are embryonic in Benin, but some activity sectors may be integrated into them provided the constraints weighing on the private sector are eased.

After having risen from 3.5% in 2011 to 5.4% in 2012, real gross domestic product (GDP) growth is estimated at 5.0% in 2013. This is mainly due to the recovery of agricultural activity, which has been fuelled in particular by: (i) an increase in agricultural production due to better access to irrigation systems, (ii) an increase in the number of agricultural workers, (iii) an increase in production and the distribution of inputs, and (iv) an increase in port traffic following port improvements. The growth rate of the real economy is projected to reach 4.9% in 2014 and 4.7% in 2015. The projected growth rate in 2014 is slightly lower than the 5.2% projected in 2013, because of Nigeria's reduction in supplies of fuel prices. Growth is projected at 4.9% in 2014 and 5.3% in 2015. The projected growth rate in 2015 is slightly higher than the 5.0% projected in 2014, because of the entry into production of a new cement plant and a number of agricultural processing units.

Macroeconomic stability should be strengthened by the ongoing reform, particularly in the agricultural sector. The government must continue to implement its fiscal policy in consultation with all the stakeholders, a clear strategy to achieve sustainable management of the cotton sector in order to increase its competitiveness and to diversify the economy away from agriculture towards the population. To support growth in Benin and reduce poverty incidence, there are also major issues related to the private sector, which must be addressed by the government. In this respect, in addition to ongoing efforts to make the business climate better and facilitate access to funding, the government must continue to implement its fiscal policy and its structural reform plan, implementing the recommendations of the round table on the development of the private sector held in December 2012.

The private sector is in fact in a good position to exploit the country's potential, notably agricultural, to the full, and its development is essential for Benin's integration into global value chains (GVCs). Although the GVCs are embryonic in Benin, they are developing rapidly, especially in the cotton sector, provided they are structured through appropriate policies. These include in particular: the cotton trade policy, the development of the cotton sector, the development of the port of Cotonou, the development of subsectoral markets, with the development of tourist areas based on public-private partnerships (PPPs), the development of infrastructure, and the development of the port of Cotonou. In this respect, the position in a port could (modernization of the port of Cotonou, construction of the Cotonou Free Zone, port of Parakou).

Table 1 Macroeconomic Indicators

	2012	2013(a)	2014(a)	2015(a)
Real GDP growth	5.4	5	4.9	5.3
Real GDP per capita growth	2.7	2.3	2.3	2.7
CPI inflation	6.6	2.6	2.3	2.9
Budget balance % GDP	-1.3	-1.2	-1.1	-1.2
Current account balance % GDP	-8.5	-8.2	-7.9	-7.8

Source: Data from domestic authorities; estimates (a) and projections (a) based on authors' calculations.

What a difference a year makes – completely rewritten economic summary of Benin with additional year's data

# Drift for collaborative knowledge archives

## • DBpedia

Average vsim score for arXiv Dbpedia references	83.1609944762
Median	83.3099367787
Standard Deviation	14.5843111047
Minimum	0
Count	114

Average vsim score for PLOS Dbpedia references	90.742775143
Median	91.5110663275
Standard Deviation	14.4120110005
Minimum	15.2128779096
Count	45

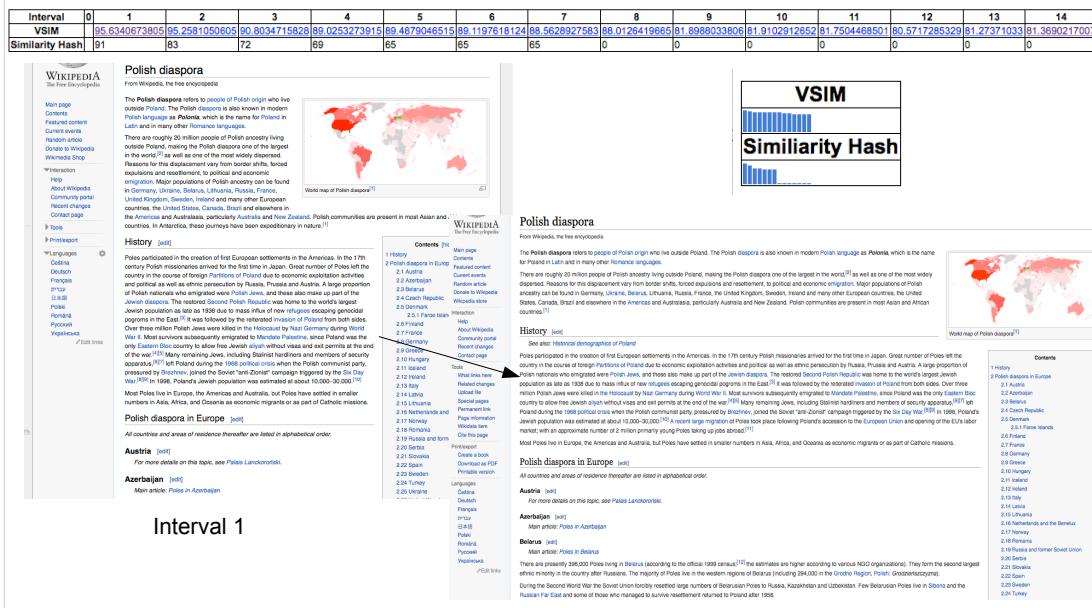
## • Wikipedia

Average vsim score for arXiv Wikipedia references	72.454779475
Median	81.1939942706
Standard Deviation	15.9157933471
Minimum	10.2760503055
Count	4103

Average vsim score for PLOS Wikipedia references	82.3733833883
Median	84.7083559866
Standard Deviation	12.005865573
Minimum	48.1399118472
Count	203

# Paper URI: plos/articles/2014-02-26/10.1371/journal.pone.0089094

URI [http://www.wikipedia.org/wiki/Polish\\_diaspora](http://www.wikipedia.org/wiki/Polish_diaspora)



## Observations

- Drift is sometimes the result of very general, or lazy, references, such as references to the root level of a project or organization website
- Although these references make sense in some ways, they are exceptionally vulnerable to content drift
- Some references drift less than the similarity measures indicate – the core of the page in the CDC example changed little over 13 months, it was the “What’s New” section that changed.

## Summary

- No matter content format or similarity measure used, measuring and characterizing content drift remains difficult and challenging
- In one sense, drift is a subjective event
- Some authors may consider significant changes in peripheral aspects of a cited Web at large reference as immaterial
- Perhaps some authors actually intend to cite the “*timeless*” version of a reference
- Yet activities such as analysing data and reproducing results might require precise alignment of software versions and other details which a snapshot would increase the likelihood of a reader finding

## Weka

- Converting forward experiment metadata and content metrics to ARFF attributes
  - Nominals, strings, numeric
- Unsupervised clustering
- Classifier for identifying drift based on attribute values
- Association algorithm and resulting rules
- Potential uses for...

# ML algorithms work with “concepts”

- Just converting your existing data and importing it may not yield good results
- What are the characteristics of the thing or phenomena that you are describing?
- Are they adequately represented within the data?
- Categorical data and numeric data is useable by more algorithms, strings are generally not useful unless you are specifically analysing text
- Concepts can be used to classify, cluster, find associations or to predict values

## Prepping data for Weka

- Metadata about cited references

```
@attribute interval {0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16}
@attribute url string
@attribute citedby string
@attribute last_val_laststatus {200, 203, 301, 302, 400, 401, 403, 404, 405, 406, 408, 409, 410, 416, 429, 490, 500, 501,
502, 503, 504}
@attribute current_val_laststatus {200, 203, 301, 302, 400, 401, 403, 404, 405, 406, 408, 409, 410, 416, 429, 490, 500,
501, 502, 503, 504}
@attribute status_chain string
@attribute vsim {0,1-9,10-19,20-29,30-39,40-49,50-59,60-69,70-79,80-89,90-99,100}
@attribute sim {0,1-9,10-19,20-29,30-39,40-49,50-59,60-69,70-79,80-89,90-99,100}
@attribute htype {*, application/json, application/msword, application/octet-stream, ... }
@attribute size numeric
@attribute tld { nominal value list }
```

- Derived attributes about the content

```
@attribute vsim numeric
@attribute sim numeric
```

## Building training data for classifier my rules for content drift

- If it's the zero'th interval and the vector similarity is greater than 80%, then flag as no drift
- If vector similarity is less than 80%, content drift has occurred, drift set to yes
- If size has changed more than 10% then drift has occurred, drift set to yes
- Otherwise flag as no drift

## Preprocessing the data

- Many algorithms don't work with strings, used unsupervised text filter to convert status chains (StringToNominal in Weka)
- All Url strings were omitted
- Tests consisting of applying various classes of algorithms against the data were conducted on instance data for the PLOS collection.
- Each instance represents a concept that has various retrieve-time metadata, a previous retrieval HTTP status and some derived values based on the actual retrieved content

# Cluster: Simple Kmeans

Clustered Instances

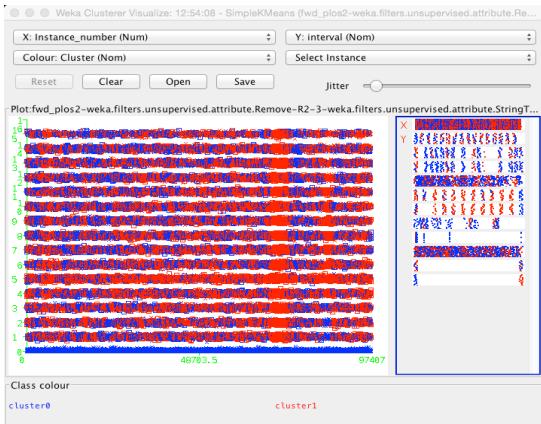
0	41599	( 43%)
1	55809	( 57%)

Class attribute: drift  
Classes to Clusters:

0	1	<-- assigned to cluster
12817	29006	yes
28782	26803	no

Cluster 0 <-- no  
Cluster 1 <-- yes

Incorrectly clustered instances : 39620.0 40.6743 %



# Classifiers: BayesNet

==== Summary ===

Correctly Classified Instances	95087	97.6172 %
Incorrectly Classified Instances	2321	2.3828 %
Kappa statistic	0.9511	
Mean absolute error	0.0337	
Root mean squared error	0.1336	
Relative absolute error	6.8687 %	
Root relative squared error	26.9827 %	
Total Number of Instances	97408	

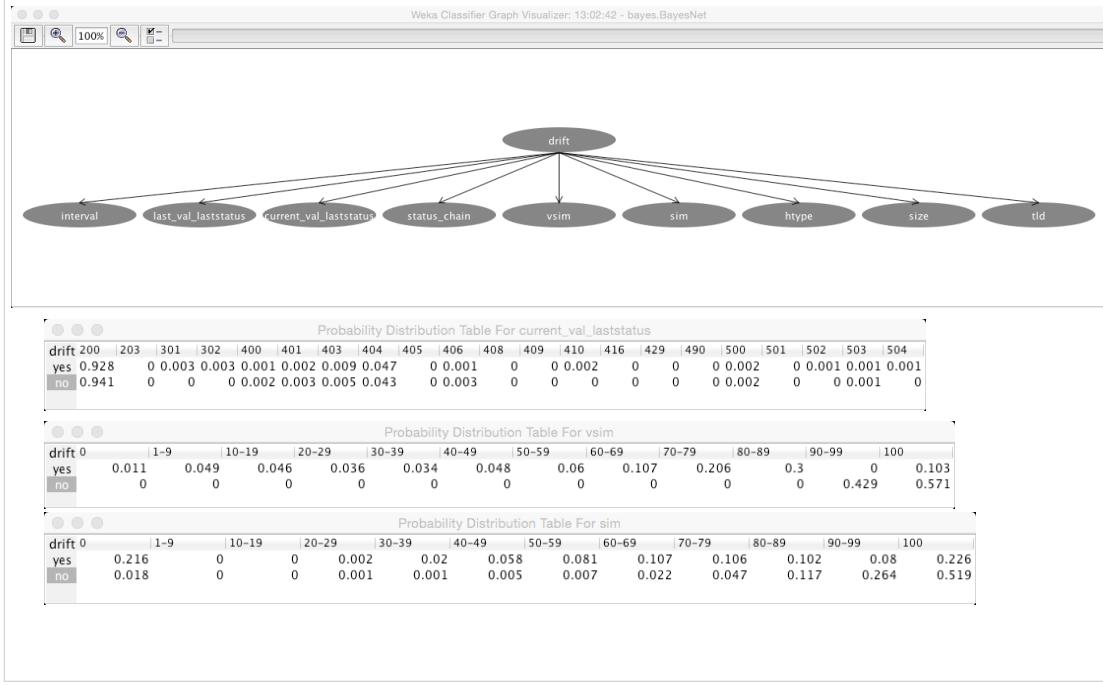
==== Detailed Accuracy By Class ===

Area	Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC
	yes	0.947	0.002	0.998	0.947	0.972	0.997
	no	0.998	0.053	0.961	0.998	0.98	0.997
	Weighted Avg.	0.976	0.031	0.977	0.976	0.976	0.997

==== Confusion Matrix ===

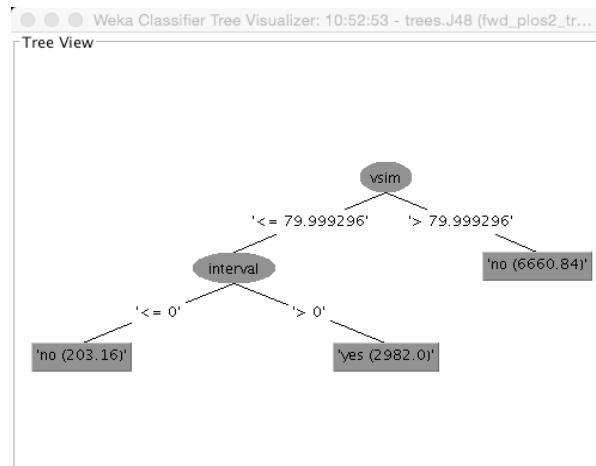
a	b	<-- classified as
39601	2222	a = yes
99	55486	b = no

# BayesNet visualized



# Decision trees

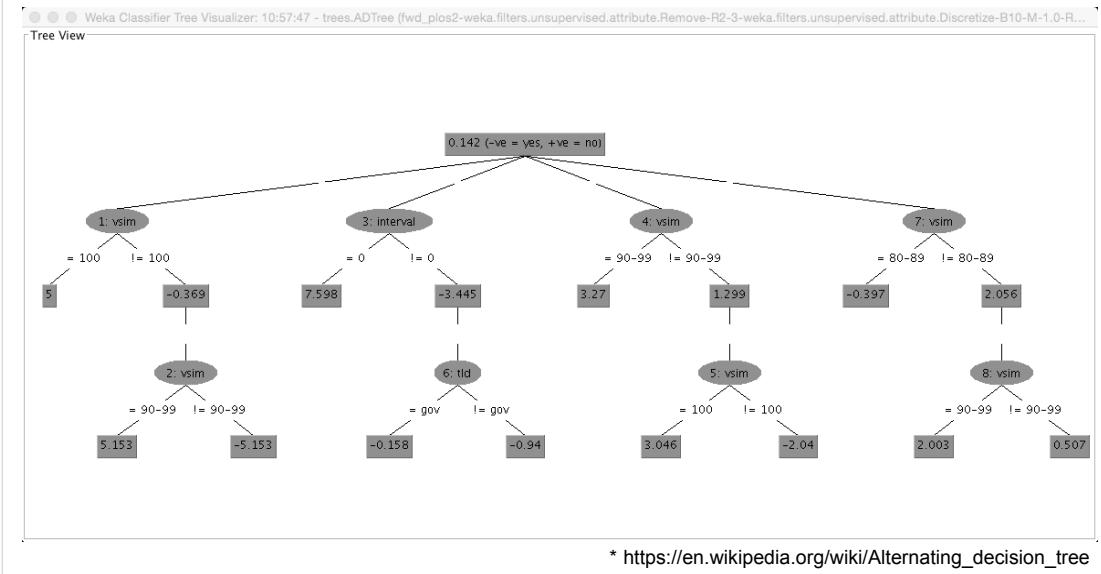
- J48\* Decision Tree



\* <http://www.d.umn.edu/~padhy005/Chapter5.html>

# Decision trees

- ADTree\* (alternating decision)



## Alternate concept model

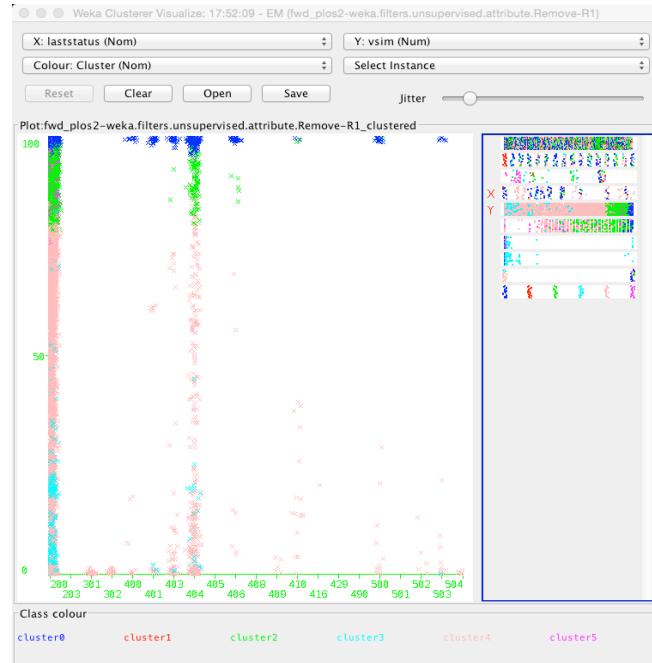
- Metadata about cited references

```
@attribute url string
@attribute interval {0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15}
@attribute htype {*, application/json, application/msword, application/octet-stream, application/pdf, application/postscript,
application/rar, application/rdf+xml, application/rss+xml, application/unknown, application/vnd.google-earth.kmz,
application/vnd.ms-excel, application/vnd.ms-powerpoint, application/x-dvi, application/x-gzip, application/x-rar,
application/x-rar-compressed, application/x-tar, application/x-troff-man, application/xhtml+xml, application/xml,
application/zip, Content-Type:text/html, image/jpeg, image/png, image/svg+xml, text/csv, text/html, text/plain, text/turtle,
text/vnd.wap.wml, text/x-csrc, text/x-server-parsed-html, text/xml, video/mp4, video/quicktime, video/x-msvideo }
@attribute laststatus {200, 203, 301, 302, 400, 401, 403, 404, 405, 406, 408, 409, 410, 416, 429, 490, 500, 501, 502, 503,
504}
@attribute size numeric
@attribute drifted {yes, no}
```

- Derived attributes about the content

```
@attribute sizechange numeric % current content length as compared to original
@attribute vsim numeric
@attribute sim numeric
```

# Simple EM (expectation maximisation)



# Classifiers: Logistic Regression

==== Summary ===

Correctly Classified Instances	19259	99.4218 %
Incorrectly Classified Instances	112	0.5782 %
Kappa statistic	0.9862	
Mean absolute error	0.0116	
Root mean squared error	0.0653	
Relative absolute error	2.78 %	
Root relative squared error	14.2883 %	
Total Number of Instances	19371	

==== Detailed Accuracy By Class ===

Area	Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC
	yes	0.991	0.004	0.989	0.991	0.99	1
	no	0.996	0.009	0.996	0.996	0.996	1
	Weighted Avg.	0.994	0.008	0.994	0.994	0.994	1

==== Confusion Matrix ===

a	b	<-- classified as
5699	51	a = yes
61	13560	b = no

## Some classification rules, by algorithm

- Ripple Down Rule Learner(Ridor) rules

```
drifted = yes  
Except (vsim > 80.00467) => drifted = no  
Except (interval = 0) => drifted = no
```

- Jrip

```
(vsim <= 79.997759) => drifted=yes  
=> drifted=no
```

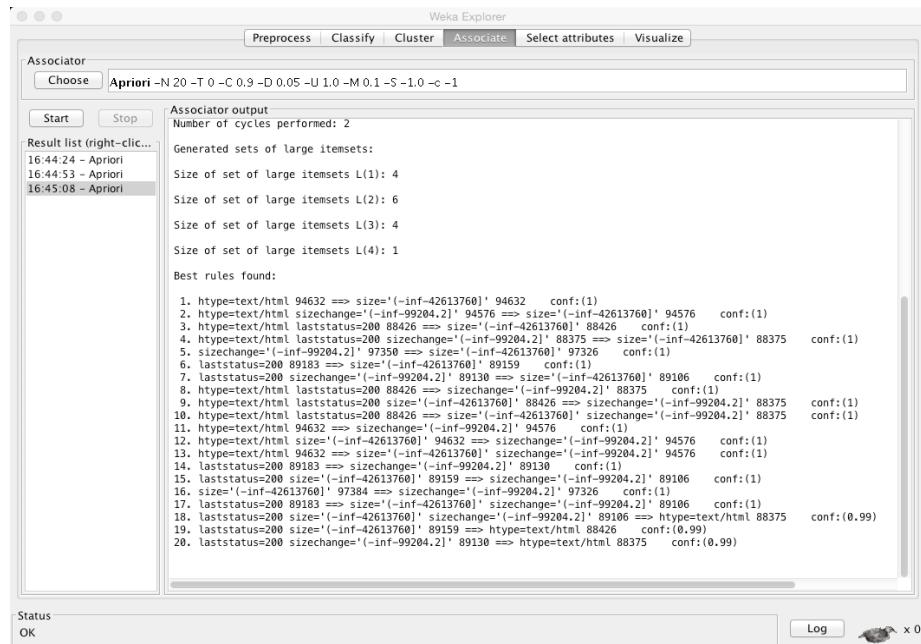
- OneR

```
vsim:  
< 80.00218866579999    -> yes  
=> 80.00218866579999   -> no  
?      -> no
```

## Applying association learner

- Algorithm will not work with numeric data
- Discretize step required for these values
  - Vsim
  - Sim
  - Size
  - Sizechange
- Tuning the Discretize algorithm
  - Somewhat arbitrarily chose 100 bins for nominals derived from numeric values – poor results. 1000 bins yields more reasonable nominal values.
  - “Use equal frequency” setting also yields better nominal values

# Association in the data: PLOS



## Apriori association learner

- High confidence level rules tend to be obvious and/or not useful associations within the data
- Setting “number of rules to find” higher is one way to see more variety and more interesting associations
- Of course, confidence level is lower for some of these rules so you have to keep this in mind when reviewing them

# Examples

- Spurious associations

- 1. sim='(99.5-inf)' 23768 ==> htype=text/html 23768 conf:(1)
- 4. laststatus=200 sim='(99.5-inf)' 21178 ==> htype=text/html 21178 conf:(1)
- 30. drifted=no 68142 ==> htype=text/html 67380 conf:(0.99)
- 38. vsim='(100-inf)' sizechange='(-0.5-0.5]' drifted=no 17066 ==> htype=text/html 16746 conf:(0.98)

- Valid associations

- 49. sim='(99.5-inf)' sizechange='(-0.5-0.5]' 21514 ==> drifted=no 20402 conf:(0.95)
- 53. drifted=no 68142 ==> laststatus=200 64395 conf:(0.95)
- 79. vsim='(100-inf)' 18520 ==> sizechange='(-0.5-0.5]' drifted=no 17066 conf:(0.92)
- 99. sim='(99.5-inf)' 23768 ==> sizechange='(-0.5-0.5]' 21514 conf:(0.91)

## What's a use cases for ML?

- Let's say our goal is to warn a user about the potential for content drift having occurred...

- Were a classifier to arrive at a combination of size, status and content type/content length as another path to accurate drift classification, then yes but preliminary tests didn't result in anything quite so simple.
- A couple of rules-based classifiers did arrive at some very simple rules which all emphasized vector similarity
- Generally, size and similarity measures are the best indicators of content drift. Size is available via HEAD HTTP, but similarity requires calculation prior to the classification step.