

# Evaluating Web Archive Search Systems

Miguel Costa, Mário J. Silva

LaSIGE @ Faculty of Sciences, University of Lisbon  
Portuguese Foundation for National Scientific Computing  
IST/INESC-ID

*WISE 2012, Paphos, Cyprus*

# The Web contains all kind of Information

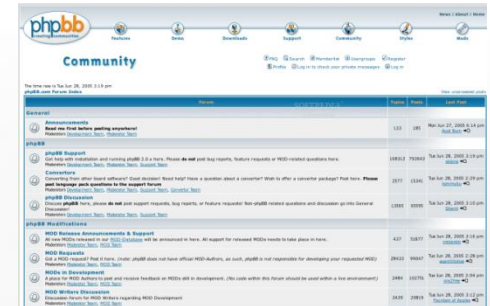
## E-books



## Web photo galleries



## Forums



## Blogs



## Online newspapers



## Social networks



# The Web is Ephemeral

- 50 days - 50% of documents are changed  
(Cho and Garcia-Molina. 2000)
- 1 year - 80% of documents become inaccessible  
(Ntoulas, Cho and Olson. 2004)
- 27 months - 13% of web references disappear  
(<http://webcitation.org/>. 2007)

# Will we face a Digital Dark Age?





## The page cannot be found

The page you are looking for might have been removed, had its name changed, or is temporarily unavailable.

---

Please try the following:

- If you typed the page address in the Address bar, make sure that it is spelled correctly.
- Open the [httpd.apache.org](http://httpd.apache.org) home page, and then look for links to the information you want.
- Click the  [Back](#) button to try another link.
- Click  [Search](#) to look for information on the Internet.

HTTP 404 - File not found  
Internet Explorer

**404**  
**ERROR**

- 5



# Portuguese Web Archive Search System



- Available since 2010: <http://archive.pt>
- 1 billion documents
  - searchable by full-text and URL
  - range between 1996 and 2011

# State-of-the-Art

- International Internet Preservation Consortium
  - 42 national libraries, archives and organizations



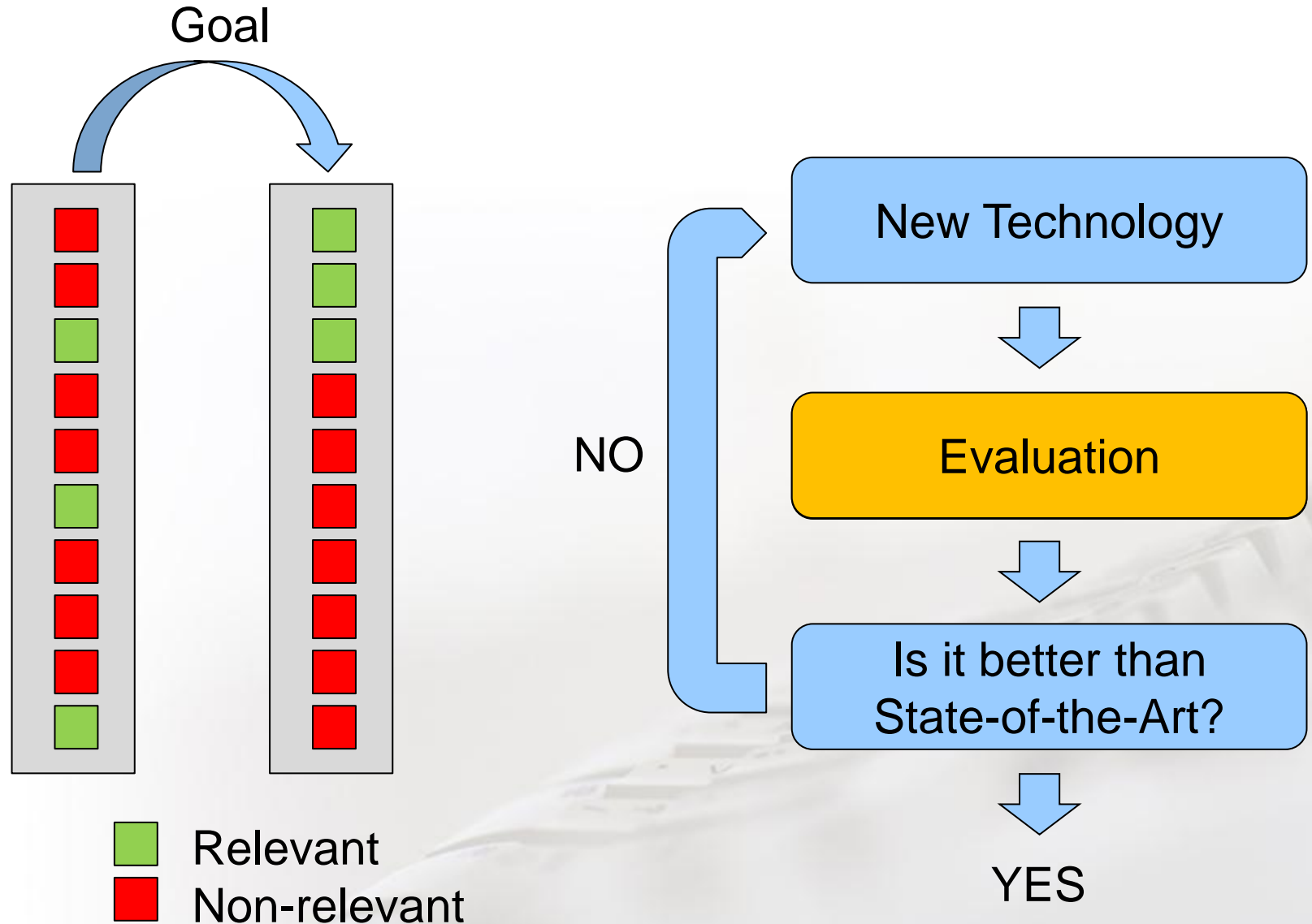
- **URL Search** – Internet Archive's Wayback Machine
  - difficult to remember or unknown



- **Full-text Search** – Lucene extensions (NutchWAX & Solr)
  - does not scale for large collections
  - slow searches
  - poor quality results



# How to improve?





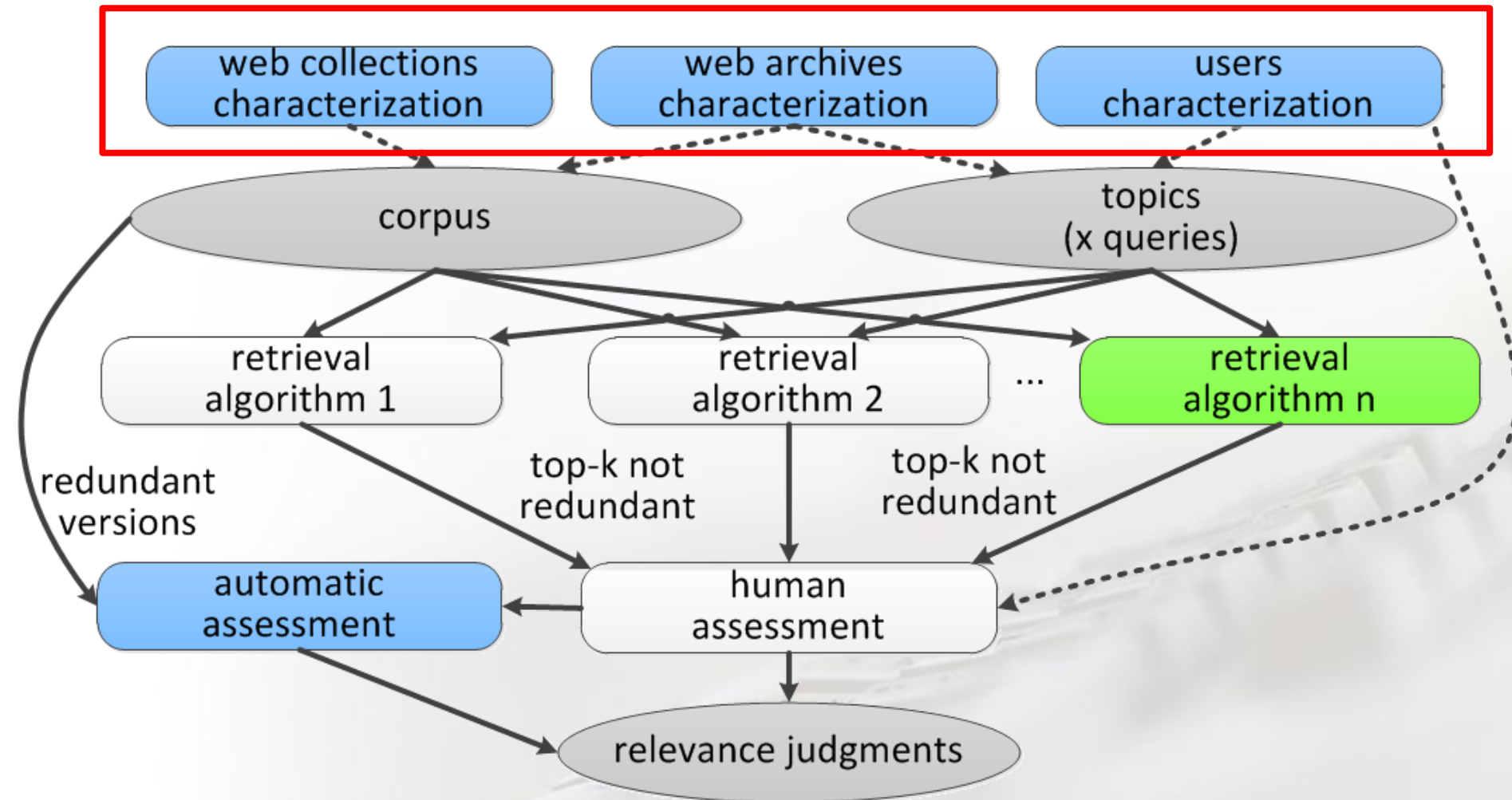


# How to evaluate?

# Evaluation – Cranfield Paradigm

- Test Collection:
  - Corpus
  - Topics
  - Relevance Judgments
  - Measures
- Evaluate system changes in a short time
- The basis of major evaluation initiatives in Information Retrieval (TREC, CLEF, NTCIR, INEX)

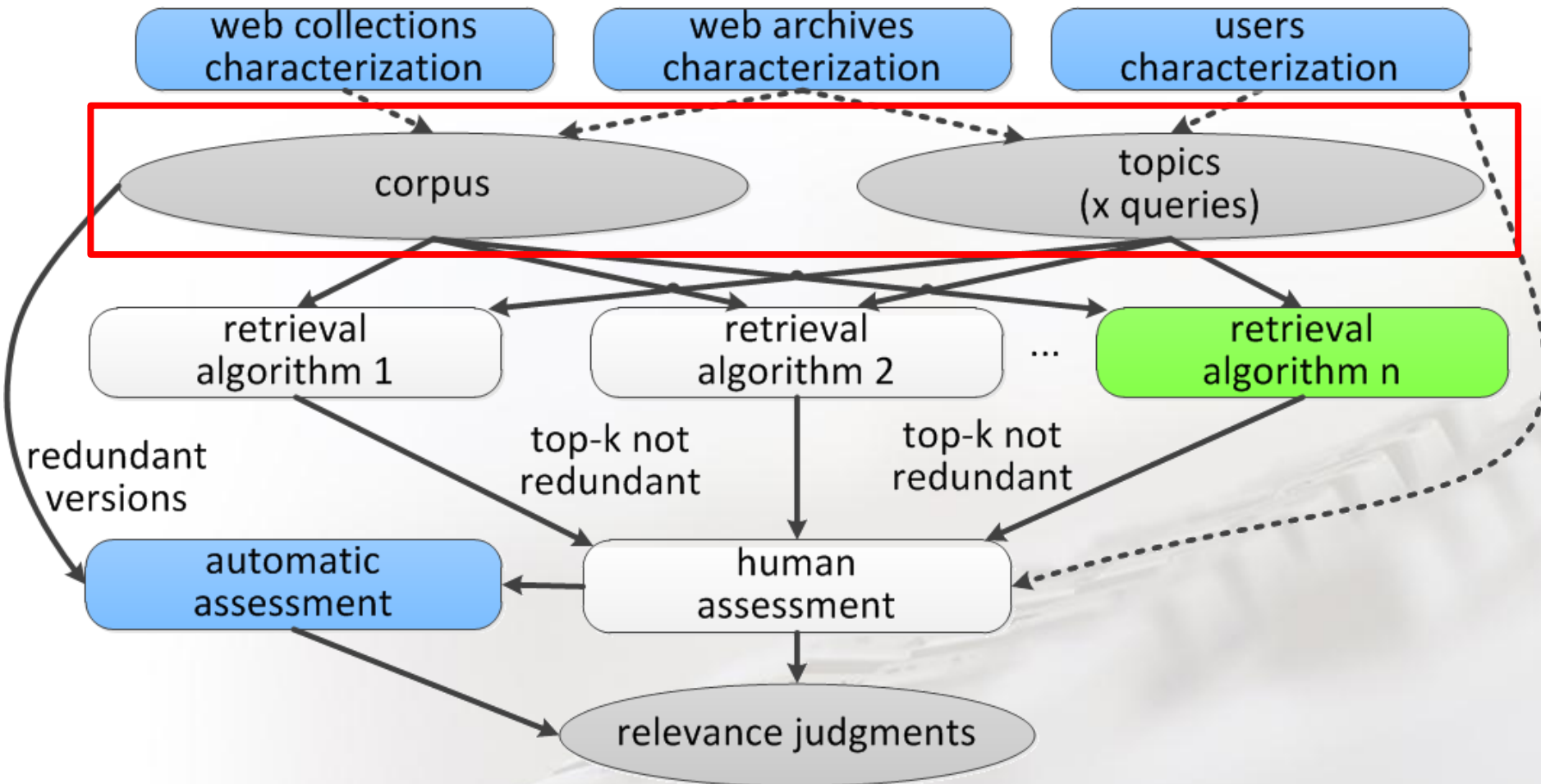
# Evaluation Methodology



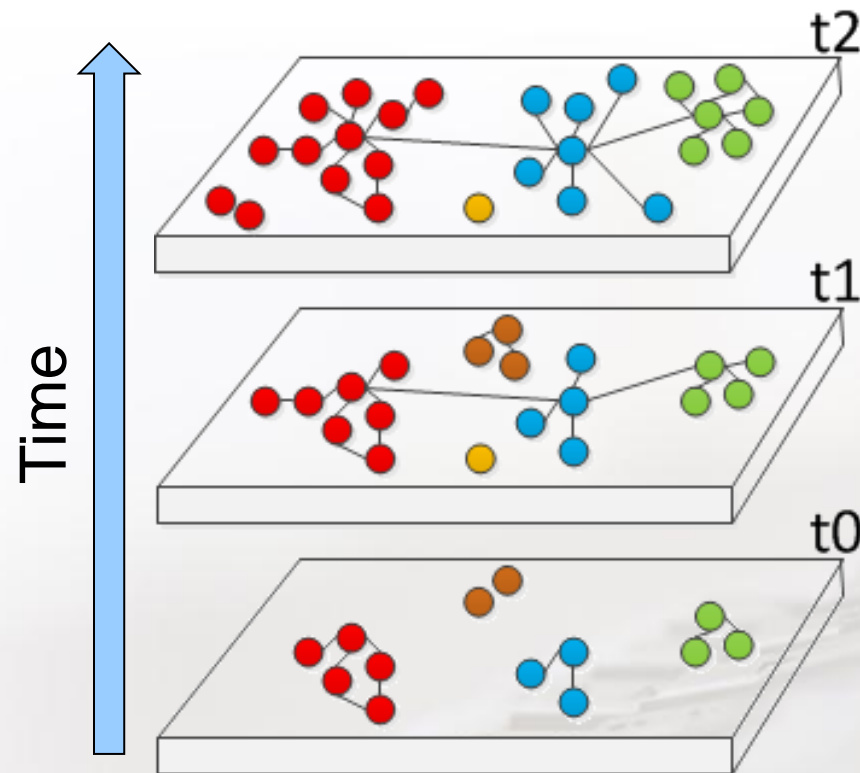
# Little Knowledge about Web Archives

- Wrong assumptions lead to wrong conclusions:
  - Corpus
    - **What** are the typical web collections?
  - Topics
    - **Why, what** and **how** do users search?
  - Relevance Judgments
    - **What** is relevant for users?
  - Measures
    - **What** and **how** many documents do users see?

# Evaluation Methodology



# Multi-version Corpus





# Navigational Topics restricted by Date Range

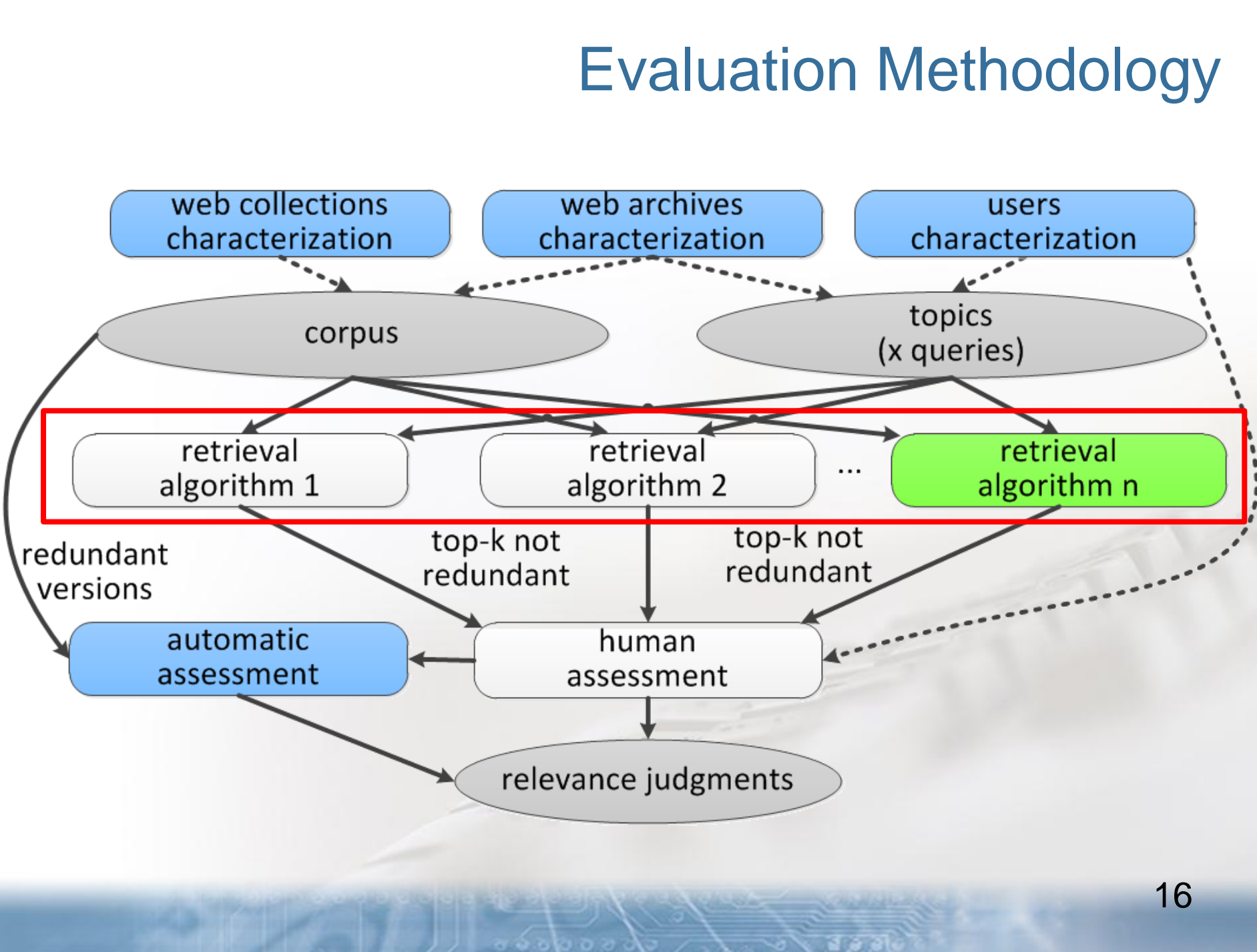
- **Navigational** – 53% to 81%
  - seeing a web page in the **past** or how it evolved
- **Informational** – 14% to 38%
  - collecting information about a topic written in the **past**
- **Example:**
  - I want to see the web page of *WISE@2010*?

# Evaluation Methodology

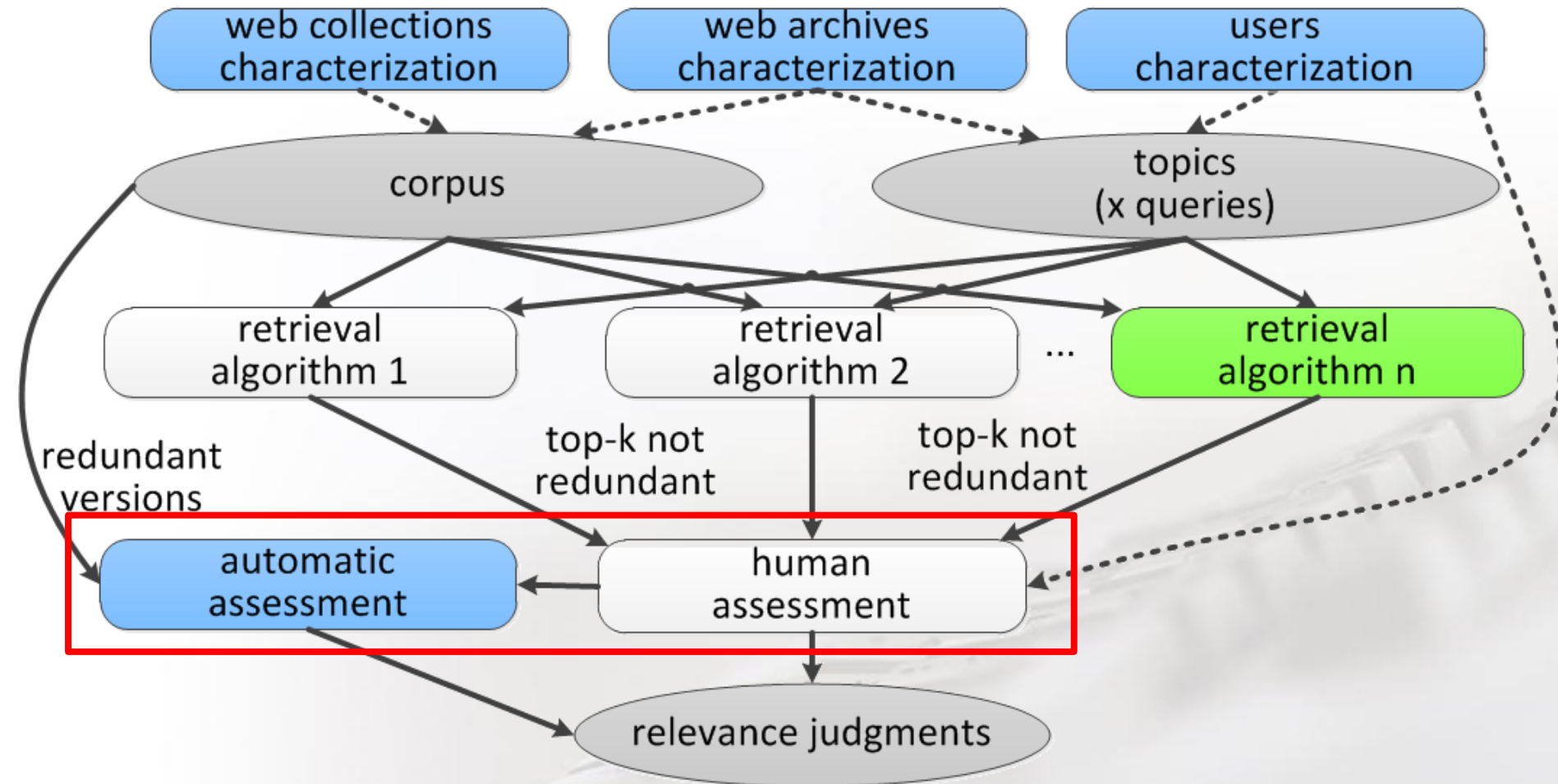
```
graph TD; WC[web collections characterization] -.-> C([corpus]); WA[web archives characterization] -.-> C; U[user characterization] -.-> T([topics x queries]); C --> R1[retrieval algorithm 1]; C --> R2[retrieval algorithm 2]; C --> Rn[retrieval algorithm n]; T --> R1; T --> R2; T --> Rn; R1 -- "top-k not redundant" --> HA[human assessment]; R2 -- "top-k not redundant" --> HA; Rn -- "top-k not redundant" --> HA; HA --> RJ([relevance judgments]); C -- "redundant versions" --> AA[automatic assessment]; AA --> RJ;
```

The diagram illustrates the evaluation methodology process:

- Data Sources:** web collections characterization, web archives characterization, and users characterization.
- Corpus and Topics:** The first two sources feed into the corpus, while all three feed into topics (x queries).
- Retrieval Algorithms:** Both the corpus and topics are used by multiple retrieval algorithms (1, 2, ..., n). These algorithms are highlighted with a red border.
- Assessment:** The top-k not redundant results from each algorithm are sent to human assessment. Redundant versions from the corpus are sent to automatic assessment.
- Final Output:** Human and automatic assessments lead to relevance judgments.

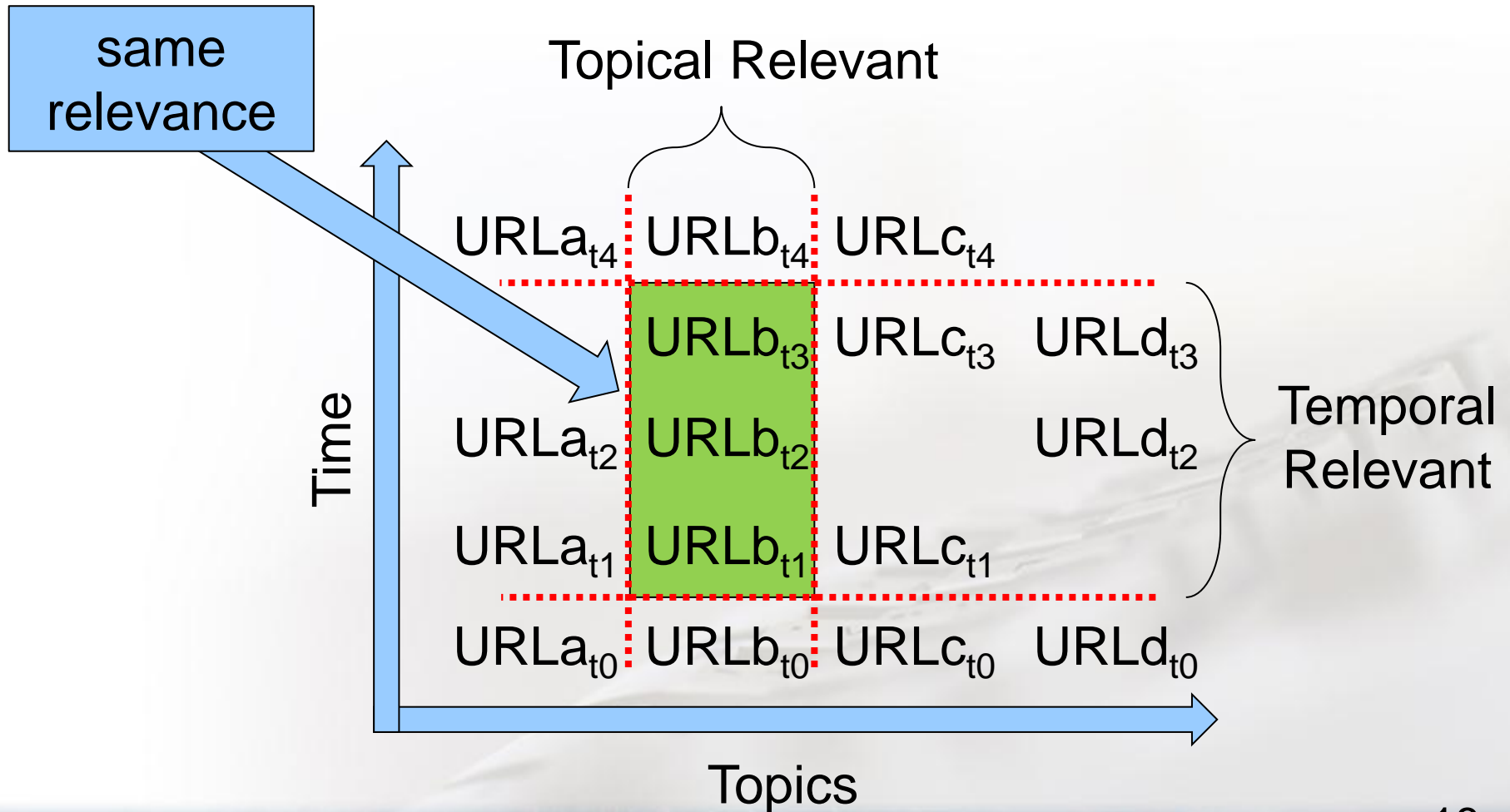


# Evaluation Methodology



# Human Assessment

Is **URL x** collected at **time t** relevant for **topic q**@[t1,t3]?



# Automatic Assessment

- Relevance propagated between versions of the same URL
- 135 times more assessments
- 4K hours per assessor saved

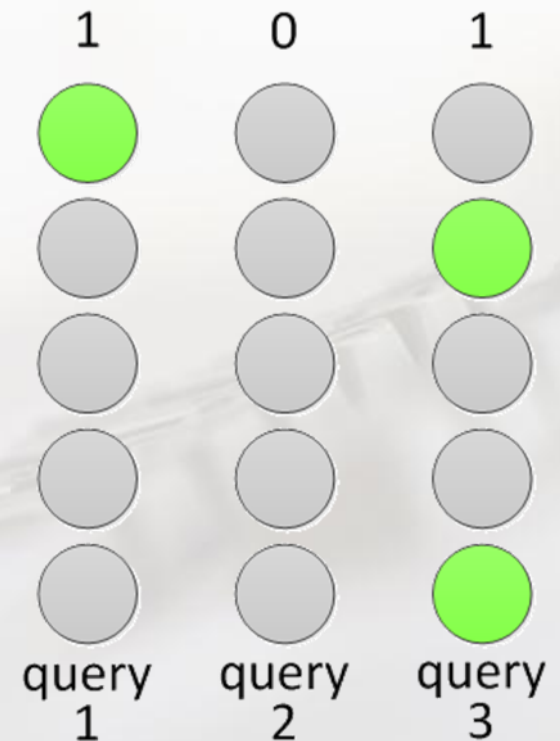
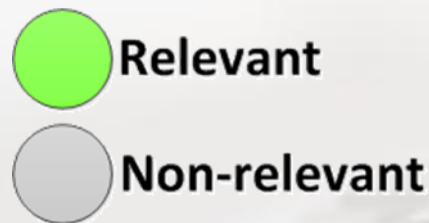
manual assessments	1 979
automatic assessments	267 822

# Results



# Evaluation Metrics

- Success@k
  - 1 if a relevant version has been found on the top-k
  - 0 otherwise
- Example: Success@5 = 2/3



# State-of-the-art (SoA) Effectiveness





# How to improve?

- Using **temporal information** intrinsic to web archives improves their search effectiveness.
  - What temporal information can I use?
    - Number of versions
    - Version's age
    - Temporal expressions in text
    - ...
  - And how?

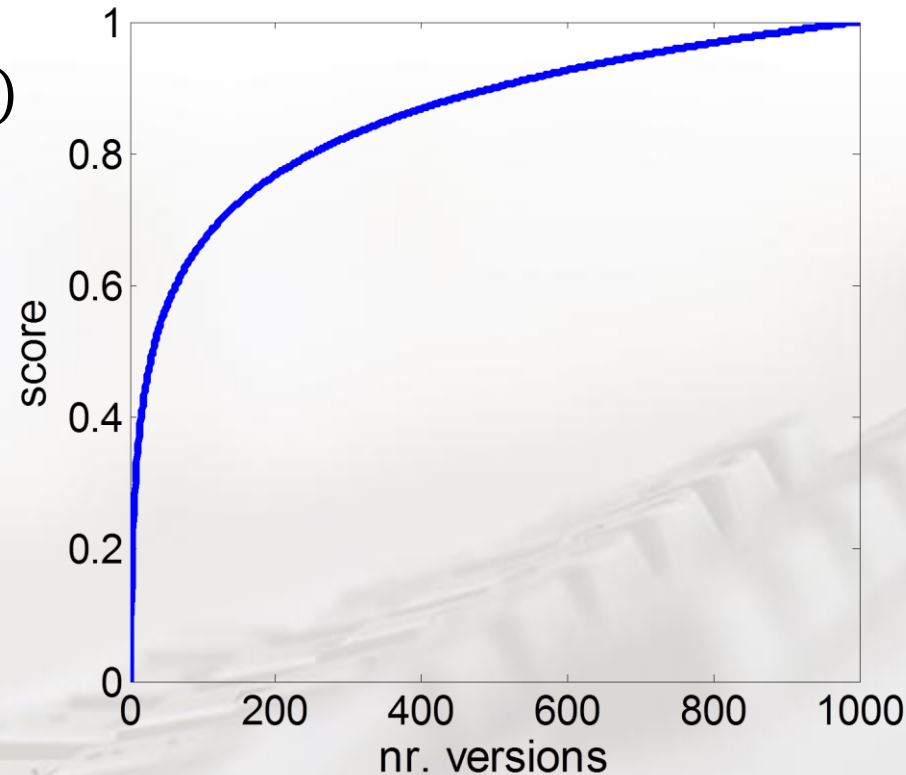
# Modelling Temporal Information

$$1. f_{Versions}(v_t^d) = \frac{\log_{10}(x)}{\log_{10}(y)} = \log_y(x)$$

*Parameters:*

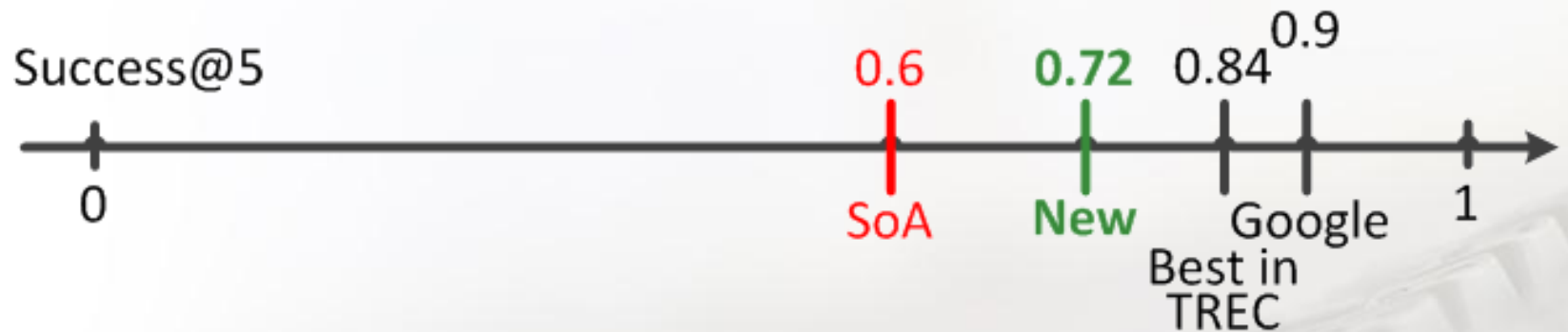
*$x$  = nr. versions of document  $d$*

*$y$  = max nr. versions of a document*



$$2. f_{Model}(v_t^d) = \lambda_1 * f_{SoA}() + \lambda_2 * f_{Versions}()$$

# New Ranking Model: $f_{\text{SoA}} + f_{\text{Versions}}$





# Conclusions

# Conclusions

- A methodology to evaluate search effectiveness.
- SoA was measured for the 1<sup>st</sup> time.
- SoA provides poor results.
- Temporal information improves search.

**Thank you.**



<http://archive.pt>

# Research Resources

- Test collection available for research.
  - <https://code.google.com/p/pwa-technologies/wiki/TestCollection>
- All code available under the LGPL license.
  - <https://code.google.com/p/pwa-technologies/>