Sub-Document Timestamping

A Study on the Content Creation Dynamics of Web Documents

Yue Zhao and Claudia Hauff

y.zhao-1@tudelft.nl

Web Information Systems group

Delft University of Technology

Agenda

- Motivations and Research Themes
- Pipeline for Sub-document Timestamping
- Exploratory Analysis
- Timestamp Inference
- Conclusions

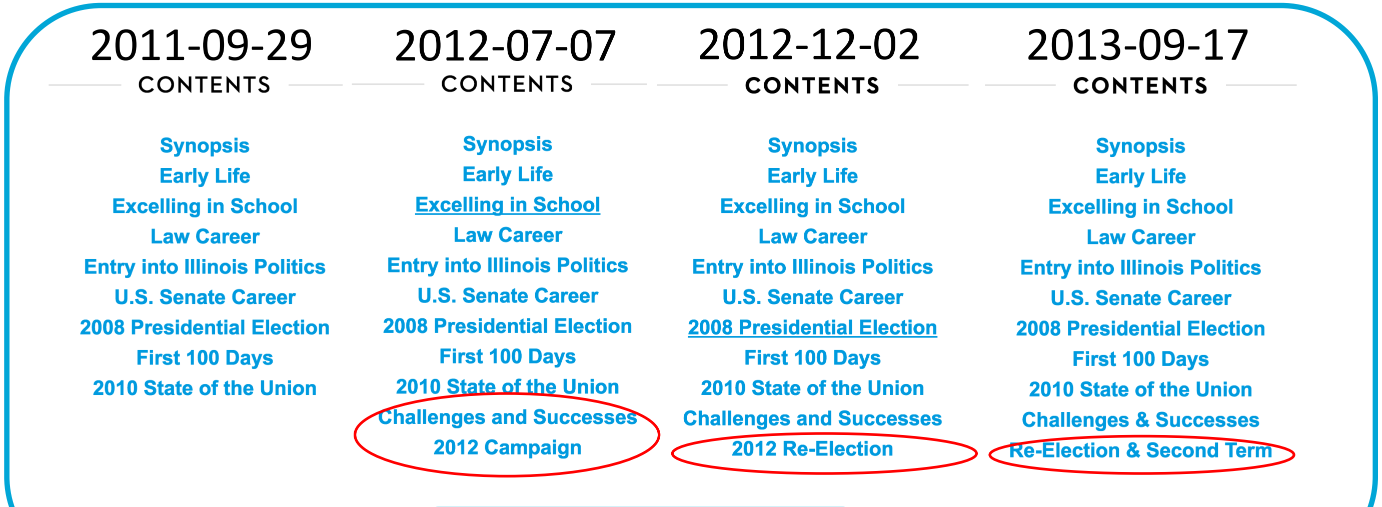
Motivations

- Document timestamping is an **important** step in temporal information retrieval.

- On the Web, documents are **dynamic**.

Motivations

- Document timestamping is an **important** step in temporal information retrieval.
- On the Web, documents are **dynamic**.



2011-09-29	2012-07-07	2012-12-02	2013-09-17
CONTENTS	CONTENTS	CONTENTS	CONTENTS
Synopsis	Synopsis	Synopsis	Synopsis
Early Life	Early Life	Early Life	Early Life
Excelling in School	Excelling in School	Excelling in School	Excelling in School
Law Career	Law Career	Law Career	Law Career
Entry into Illinois Politics	Entry into Illinois Politics	Entry into Illinois Politics	Entry into Illinois Politics
U.S. Senate Career	U.S. Senate Career	U.S. Senate Career	U.S. Senate Career
2008 Presidential Election	2008 Presidential Election	2008 Presidential Election	2008 Presidential Election
First 100 Days	First 100 Days	First 100 Days	First 100 Days
2010 State of the Union	2010 State of the Union	2010 State of the Union	2010 State of the Union
	Challenges and Successes	Challenges and Successes	Challenges & Successes
	2012 Campaign	2012 Re-Election	Re-Election & Second Term

Barack Obama's biography page

Research Themes

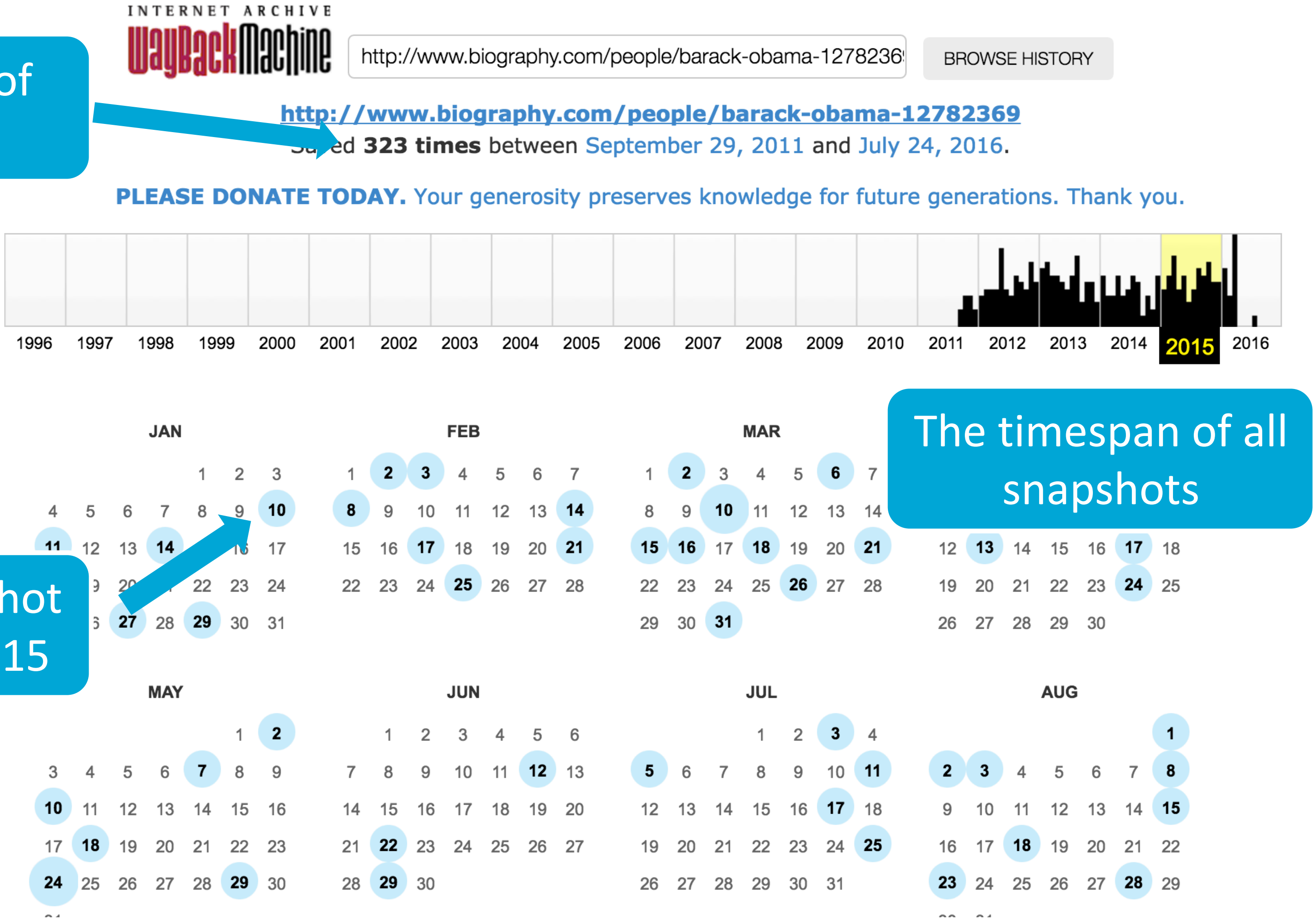
- RT 1: To what extent do Web documents consist of sub-documents created at different time?

- RT 2: To what extent can we infer the creation time of sub-documents on the Web?

Data Sets

Data Sets	*General*	*Quality*	*Seen*
Data Resources	ClueWeb12 Internet Archive	ClueWeb12 Internet Archive	ClueWeb12 Internet Archive
Selection Methods	Randomly sampled	Judged relevant to at least one specific topic	Marked as crawled from Twitter
# Documents	433,082	7,118	23,077
# Historical Versions	2,961,005	121,671	368,106
Characteristics		Each document has some meaningful content to TREC topics	Each document was of interest to some real users

Data Sets



http://www.biography.com/people/barack-obama-12782369

http://www.biography.com/people/barack-obama-12782369

Saved **323 times** between September 29, 2011 and July 24, 2016.

PLEASE DONATE TODAY. Your generosity preserves knowledge for future generations. Thank you.

JAN

				1	2	3
4	5	6	7	8	9	10
11	12	13	14		16	17
	19	20	21	22	23	24
	26	27	28	29	30	31

FEB

1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28

MAR

1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30		

MAY

					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

JUN

	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30				

JUL

			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

AUG

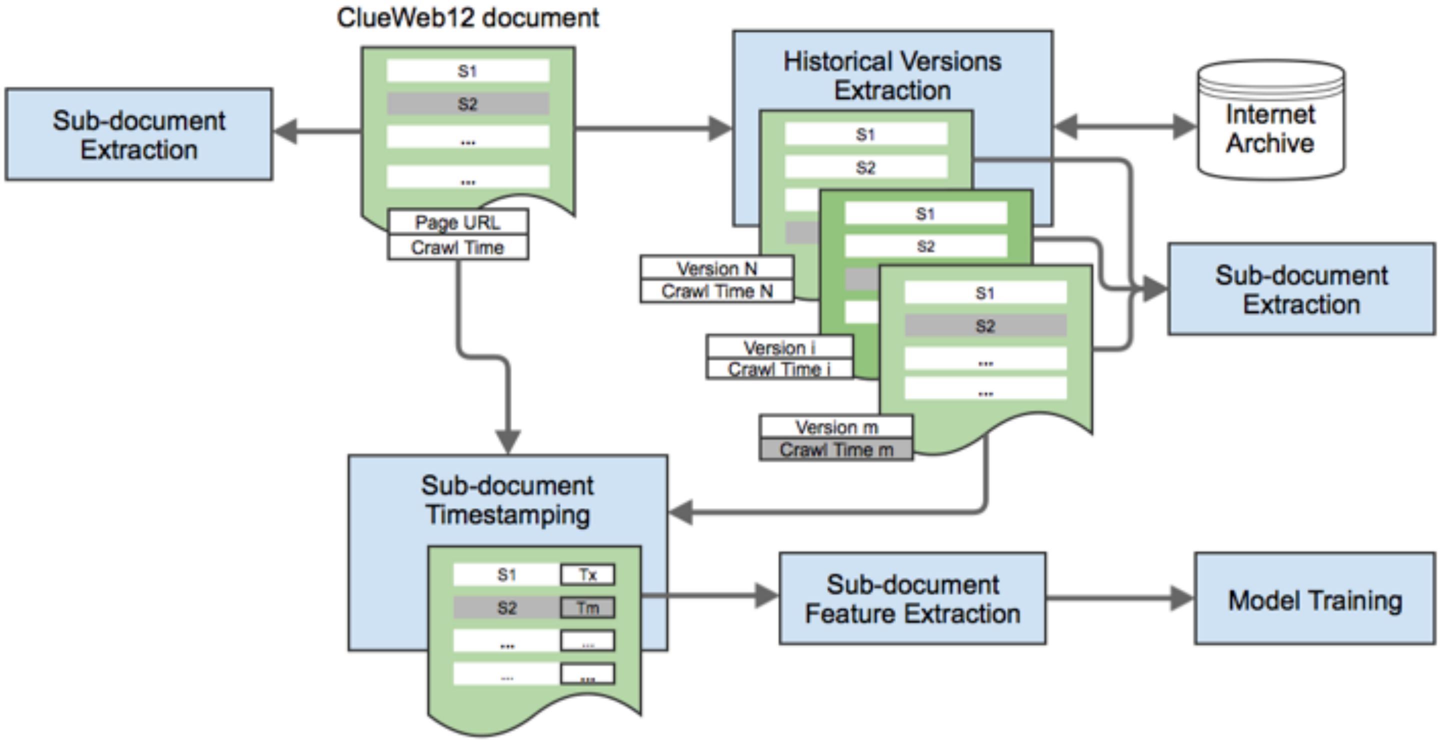
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29

Total number of snapshots

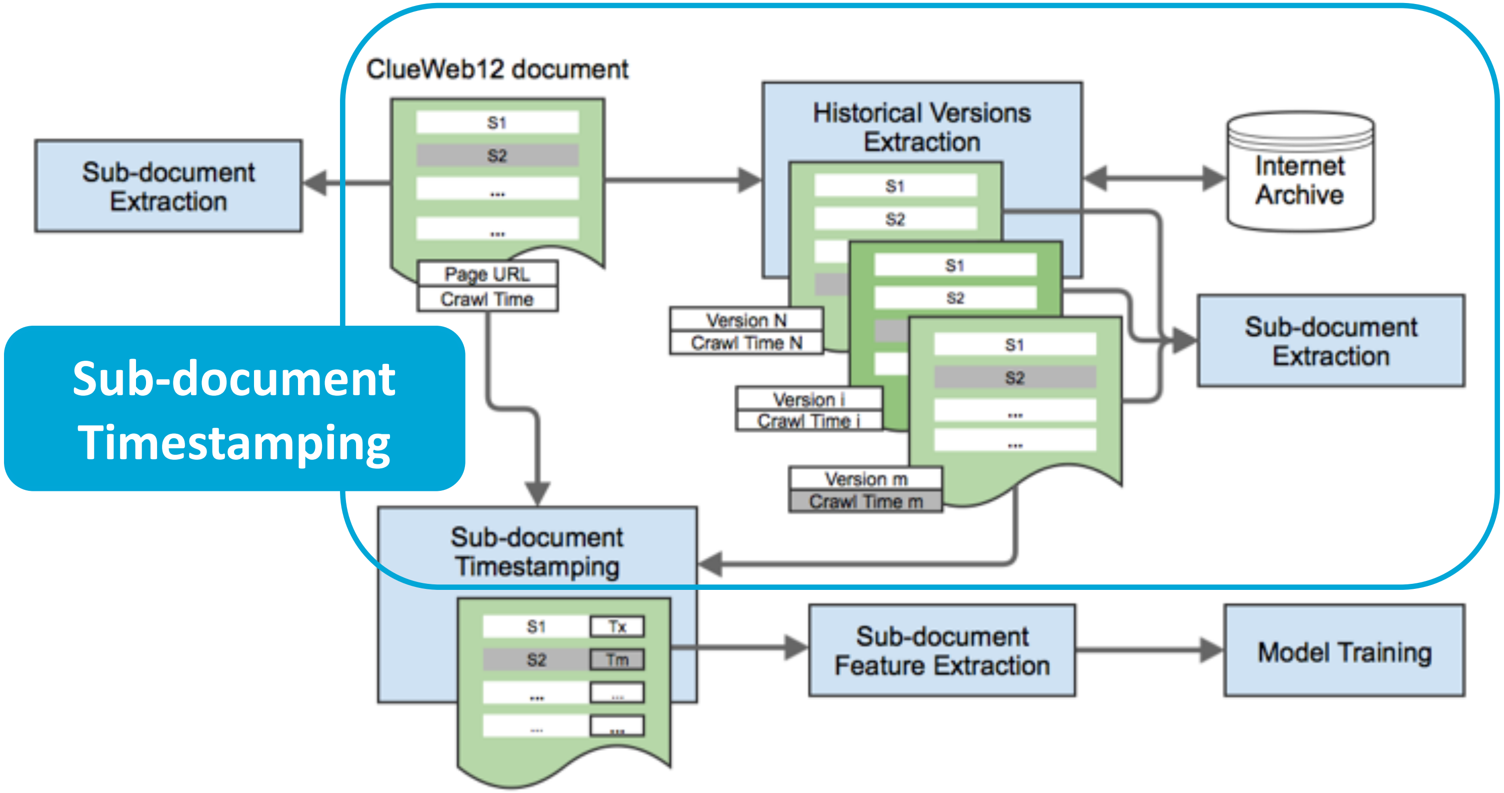
The timespan of all snapshots

At least 1 snapshot on 10th Jan, 2015

Pipeline

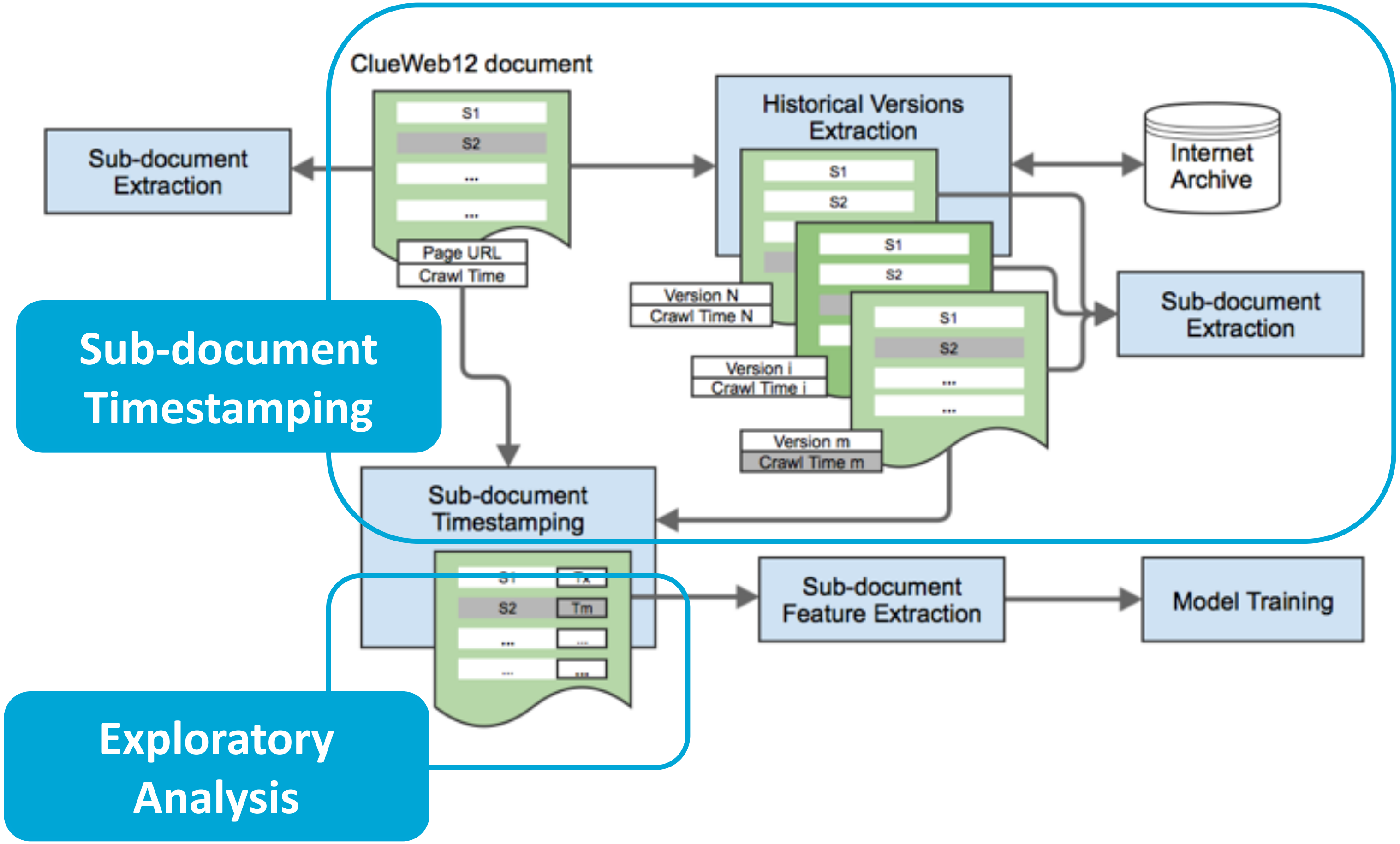

ClueWeb12 document
S1
S2
...
...
Page URL
Crawl Time
Sub-document Extraction
Historical Versions Extraction
Internet Archive
S1
S2
Version N
Crawl Time N
S1
S2
Version i
Crawl Time i
S1
S2
...
...
Version m
Crawl Time m
Sub-document Extraction
Sub-document Timestamping
S1
Tx
S2
Tm
...
...
...
...
Sub-document Feature Extraction
Model Training

Pipeline



Sub-document Timestamping

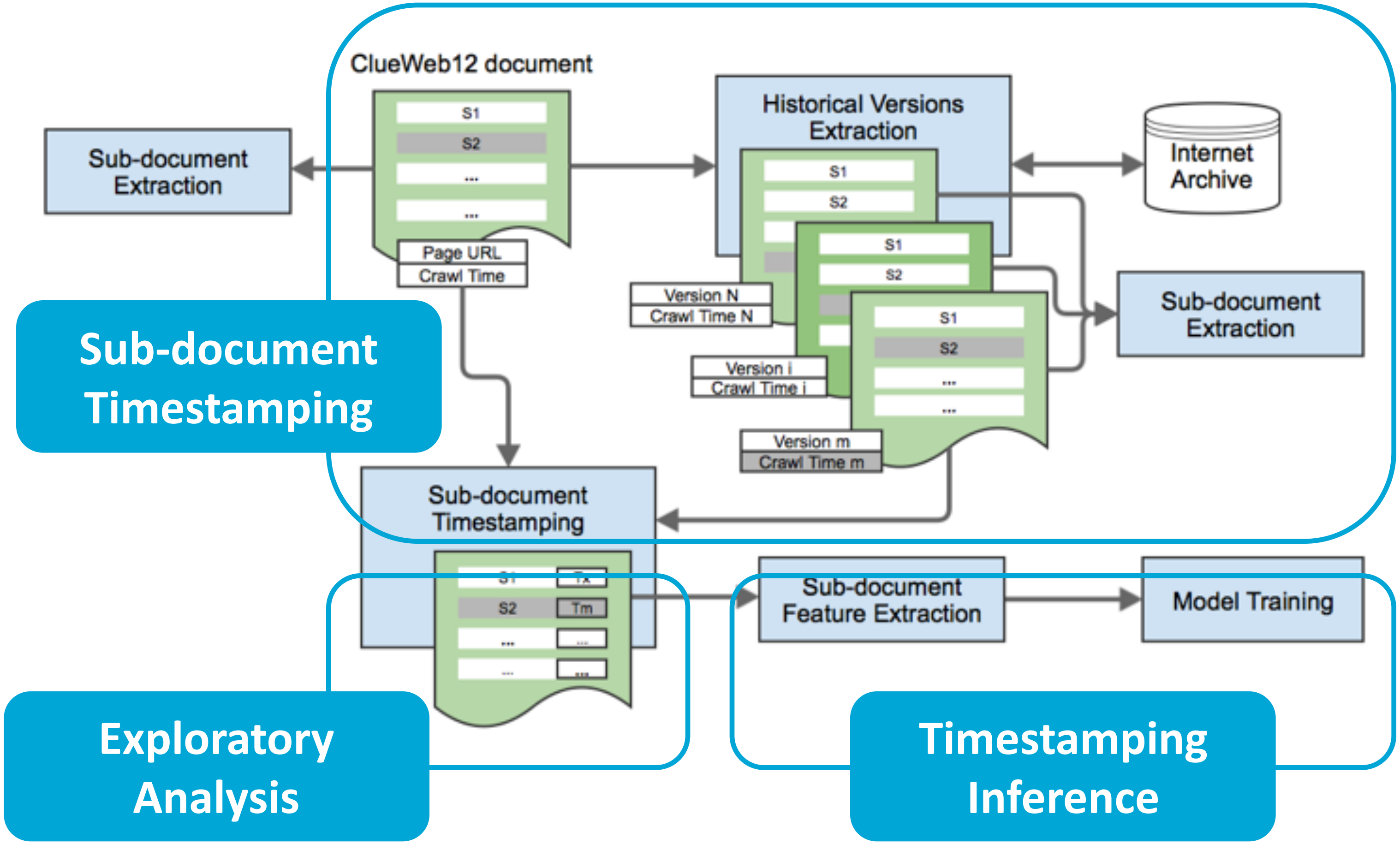
Pipeline



Sub-document Timestamping

Exploratory Analysis

Pipeline



Sub-document Timestamping

Exploratory Analysis

Timestamping Inference

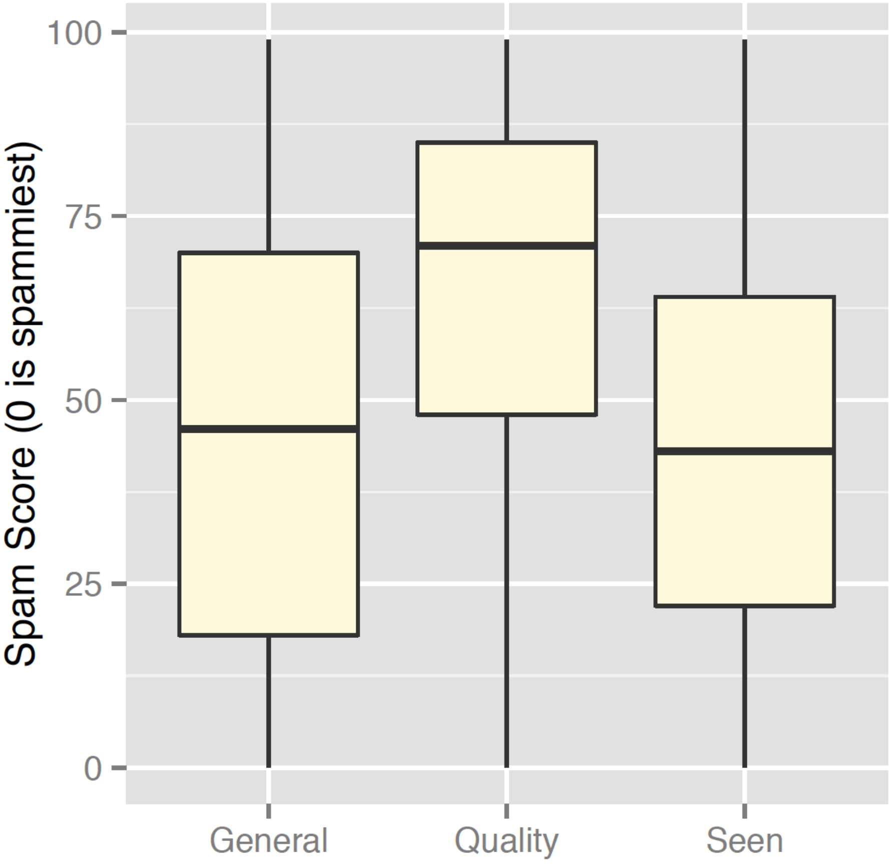
Exploratory Analysis

RQ 1: To what extent do the **document qualities** vary across the three sets?

Pre-computed Web spam scores for ClueWeb12

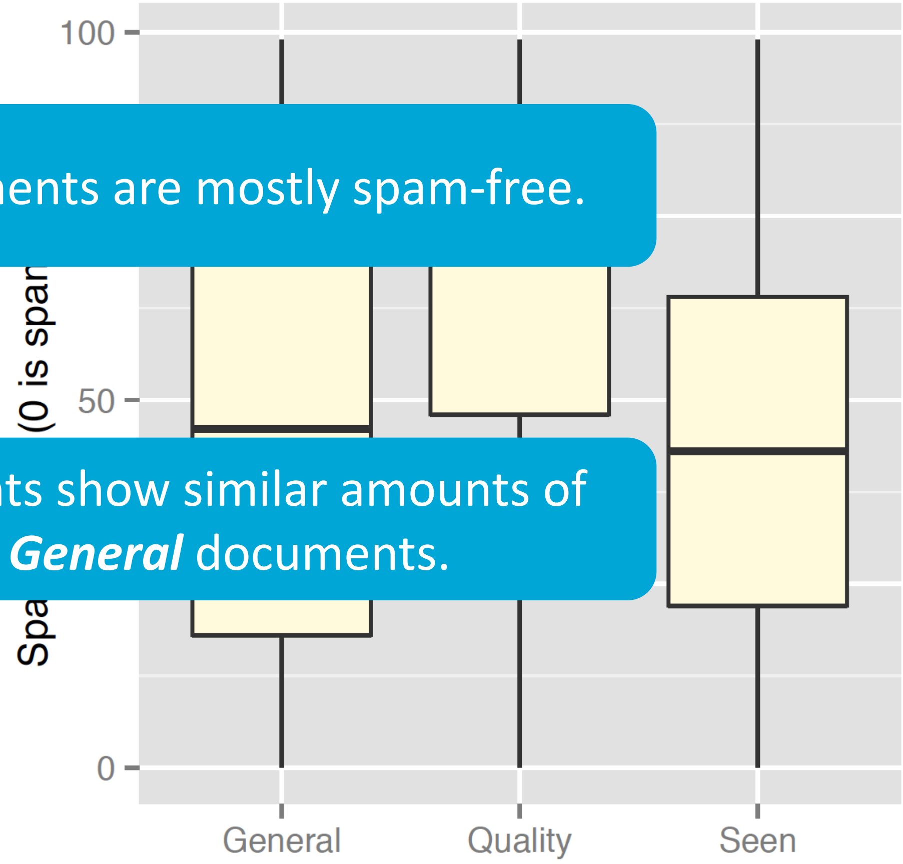
0 is most likely to be spam and 100 is least likely to be spam.

Below 70 are considered to have at least some spam in them



Exploratory Analysis

RQ 1: To what extent do the **document qualities** vary across the three sets?

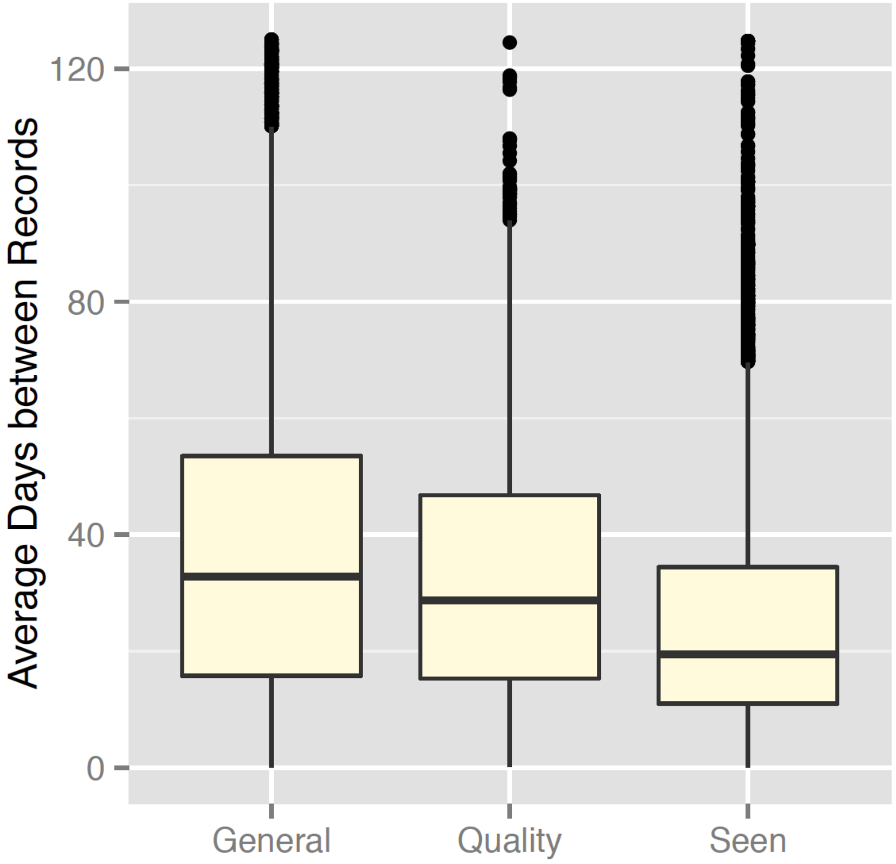


Quality documents are mostly spam-free.

Seen documents show similar amounts of spam as ***General*** documents.

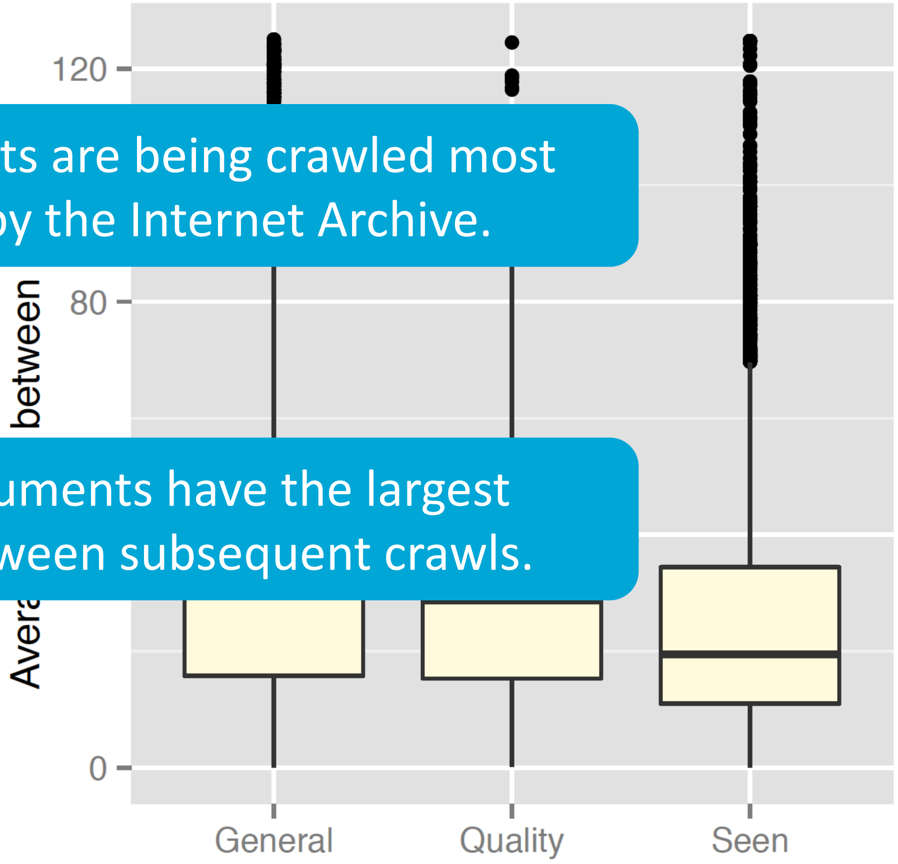
Exploratory Analysis

RQ 2: Do the **crawl frequencies** of documents differ in the Internet Archive?



Exploratory Analysis

RQ 2: Do the **crawl frequencies** of documents differ in the Internet Archive?

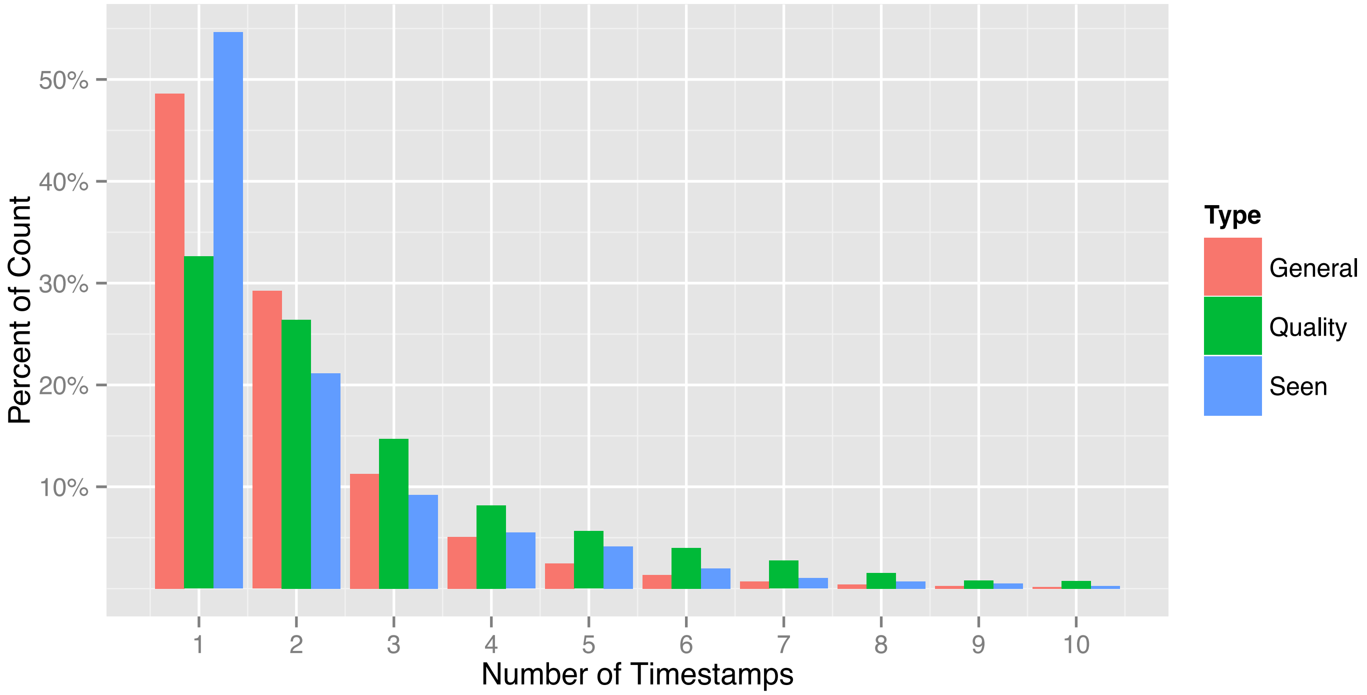


Seen documents are being crawled most frequently by the Internet Archive.

General documents have the largest timespan between subsequent crawls.

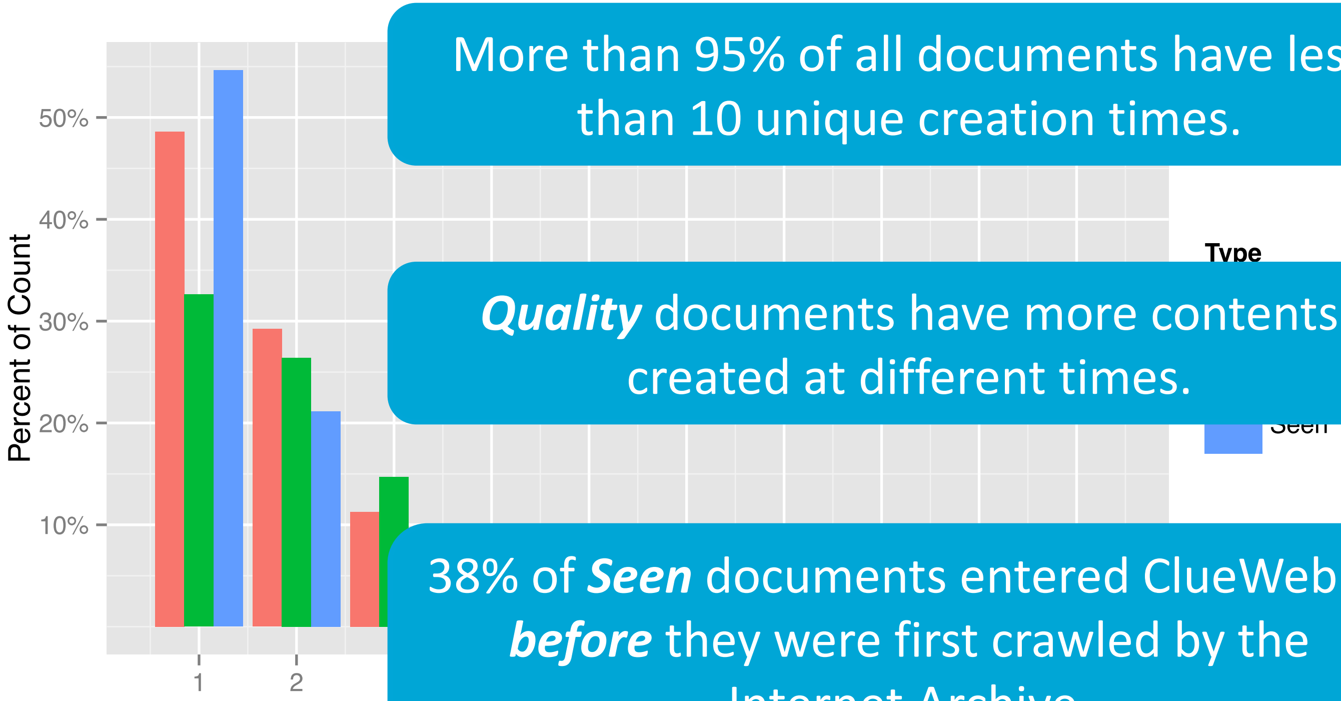
Exploratory Analysis

RQ 3: What proportion of Web documents are created at multiple points in time?



Exploratory Analysis

RQ 3: What proportion of Web documents are created at multiple points in time?



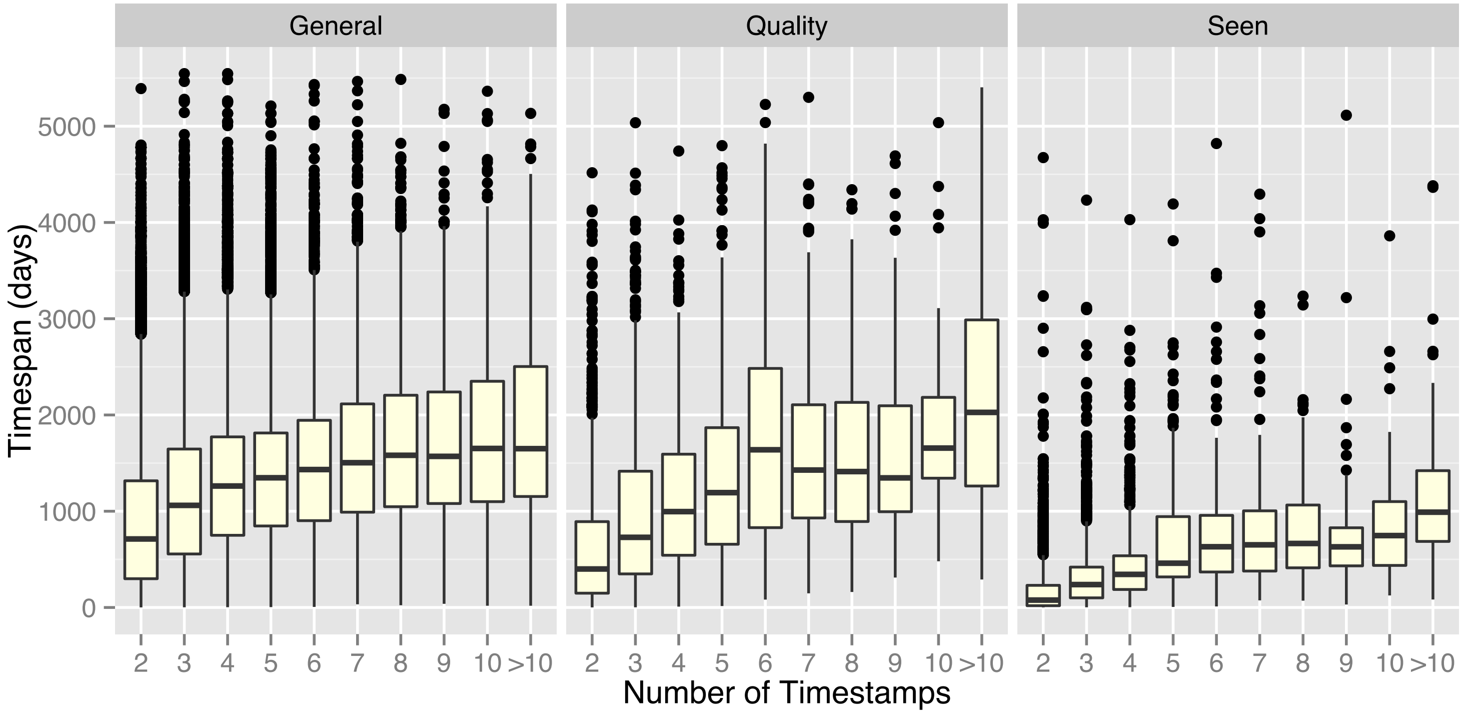
More than 95% of all documents have less than 10 unique creation times.

Quality documents have more contents created at different times.

38% of ***Seen*** documents entered ClueWeb12 ***before*** they were first crawled by the Internet Archive.

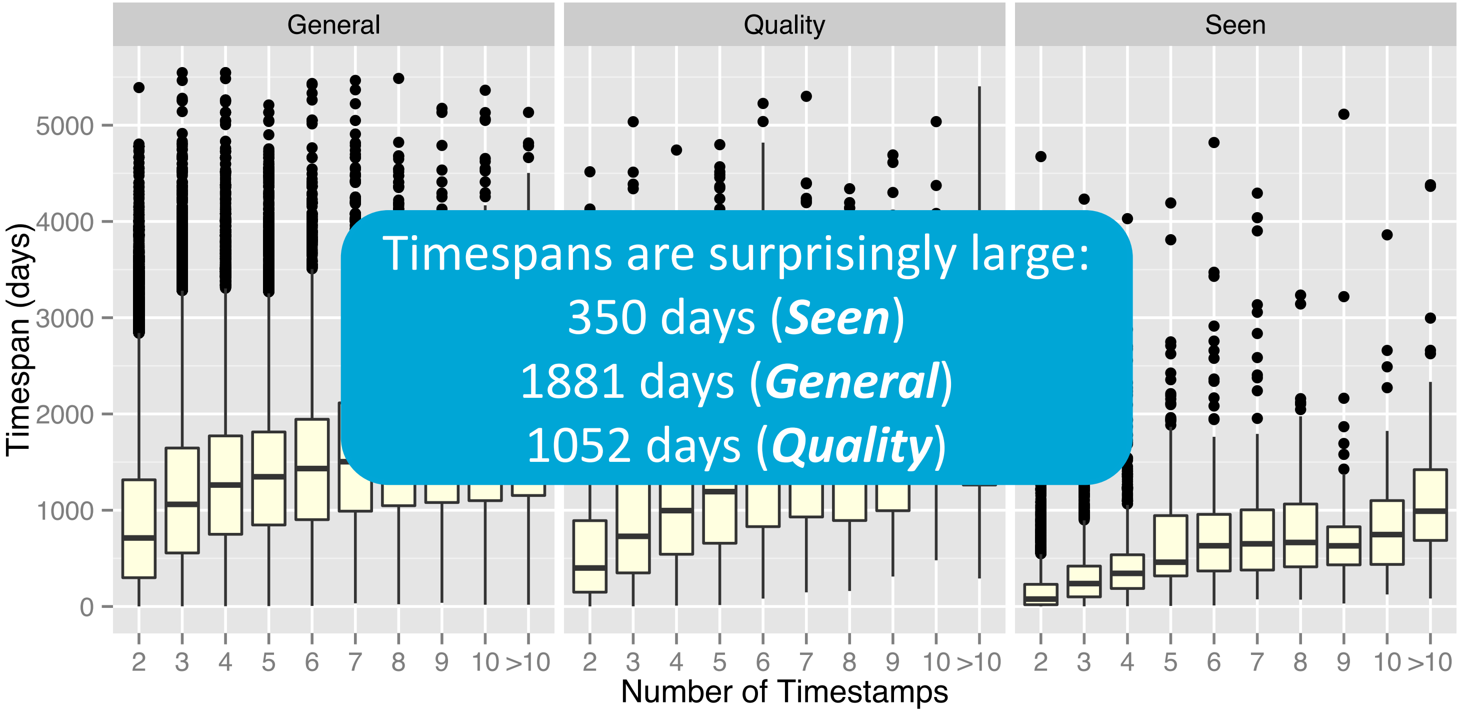
Exploratory Analysis

RQ 4: How much time passes between content updates?



Exploratory Analysis

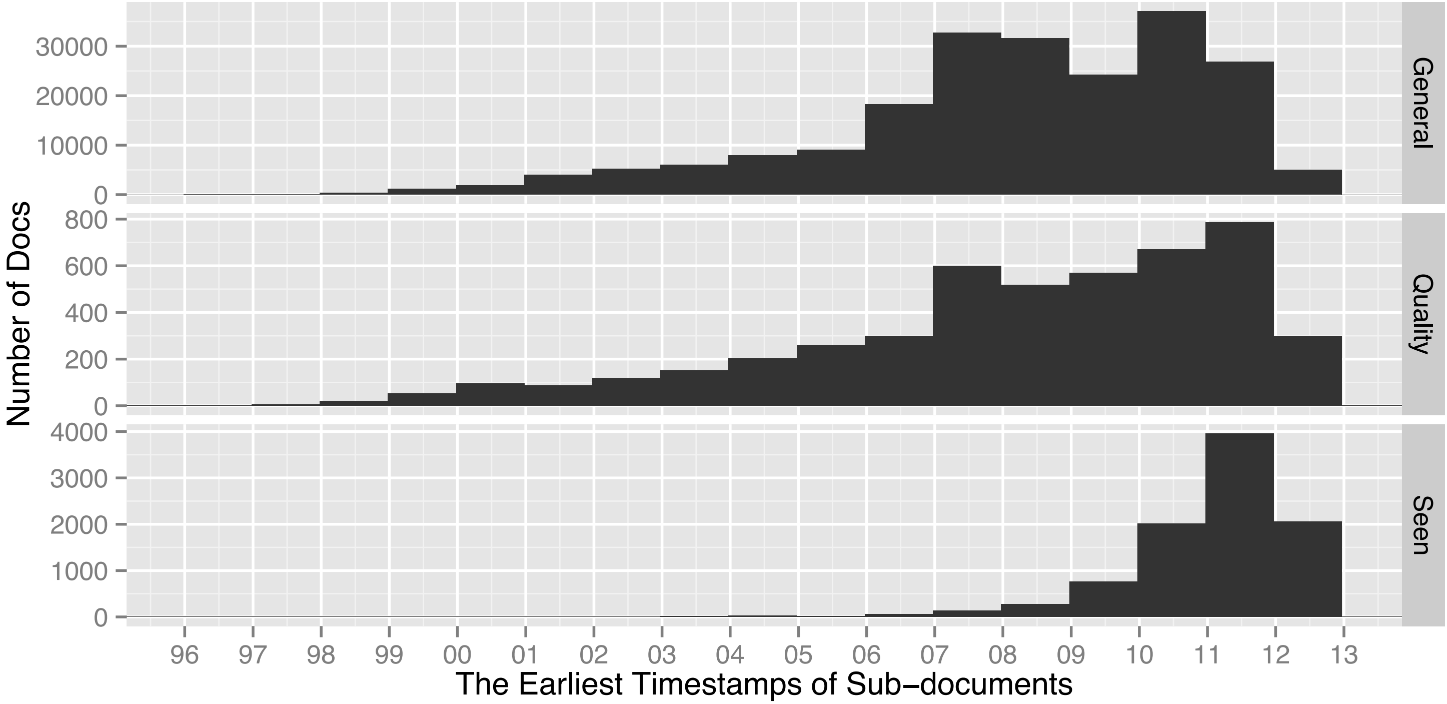
RQ 4: How much time passes between content updates?



Timespans are surprisingly large:
350 days (***Seen***)
1881 days (***General***)
1052 days (***Quality***)

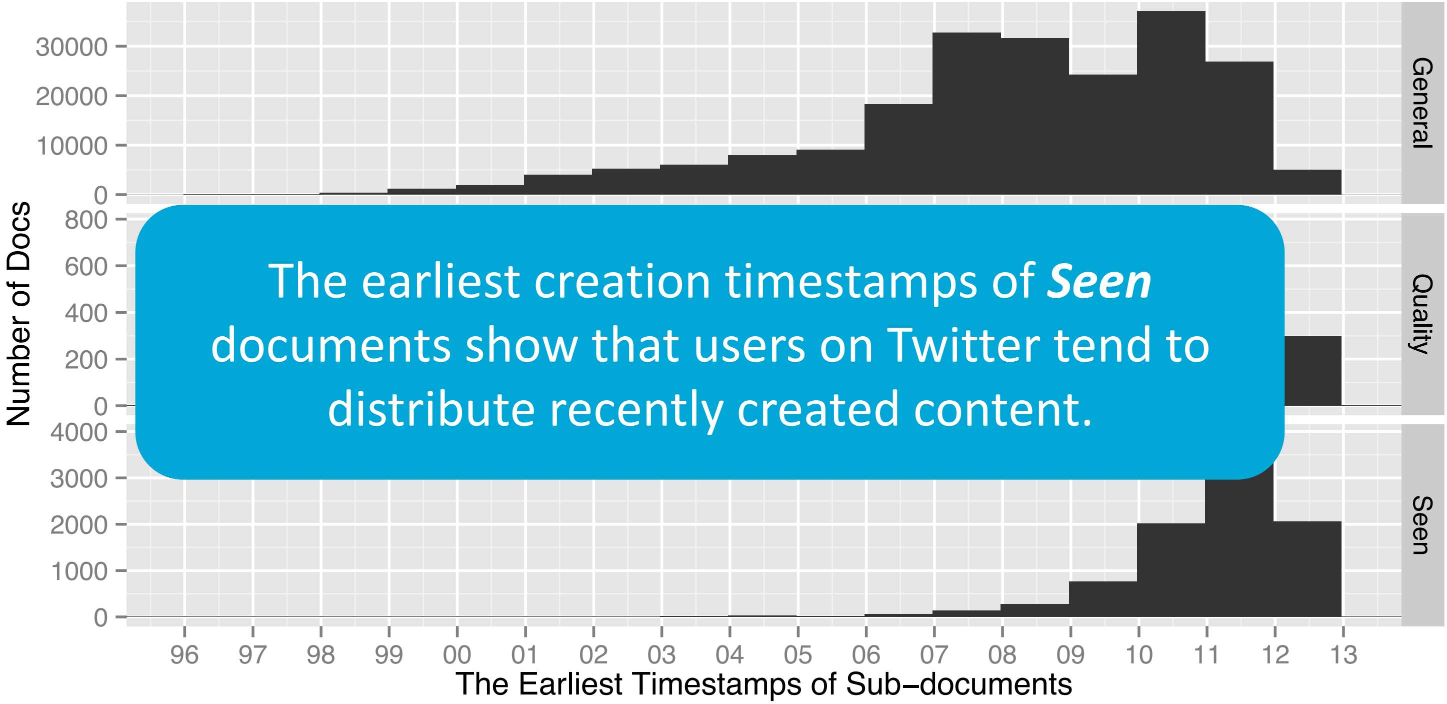
Exploratory Analysis

RQ 4: How much time passes between content updates?



Exploratory Analysis

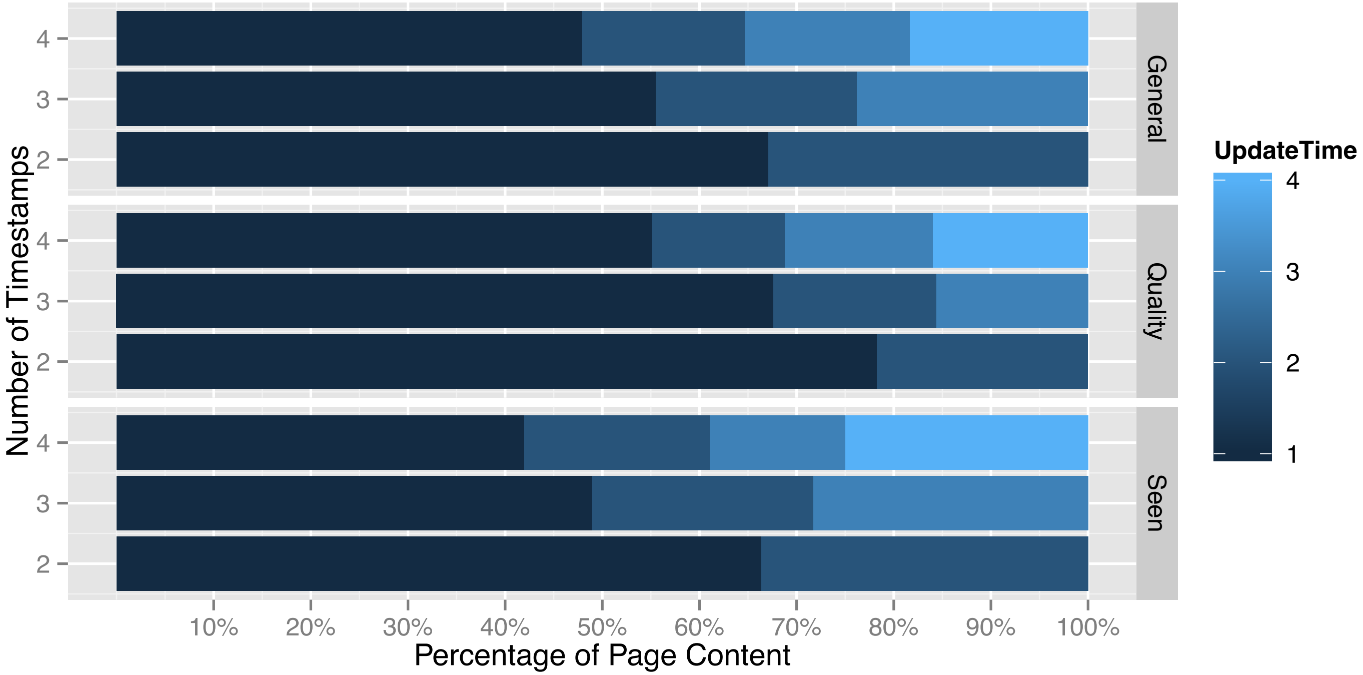
RQ 4: How much time passes between content updates?



The earliest creation timestamps of ***Seen*** documents show that users on Twitter tend to distribute recently created content.

Exploratory Analysis

RQ 5: What proportion of content is created over time?

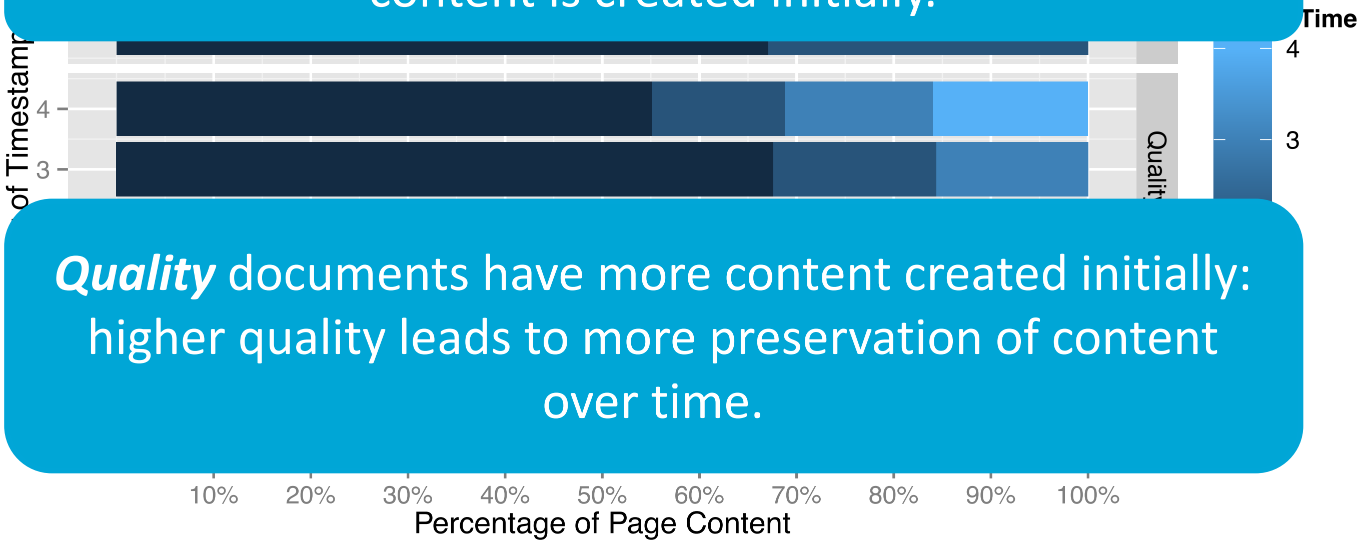


TUDelft

Exploratory Analysis

RQ 5: What proportion of content is created over time?

The more creation timestamps a document has, the less content is created initially.

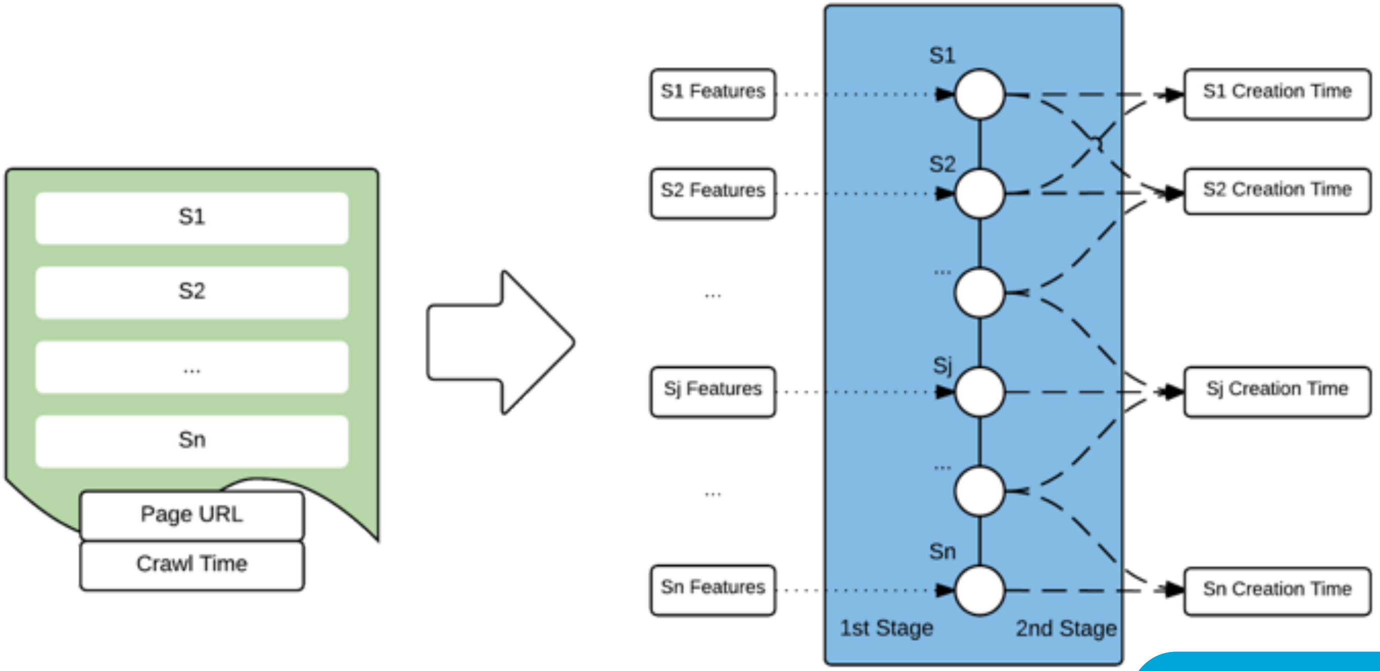


Quality documents have more content created initially: higher quality leads to more preservation of content over time.

Temporal Inference

Two-Stage Model

- A new classification method



Web documents

Random Forest (RF)

2-stage model

Conditional Random Fields (CRFs)

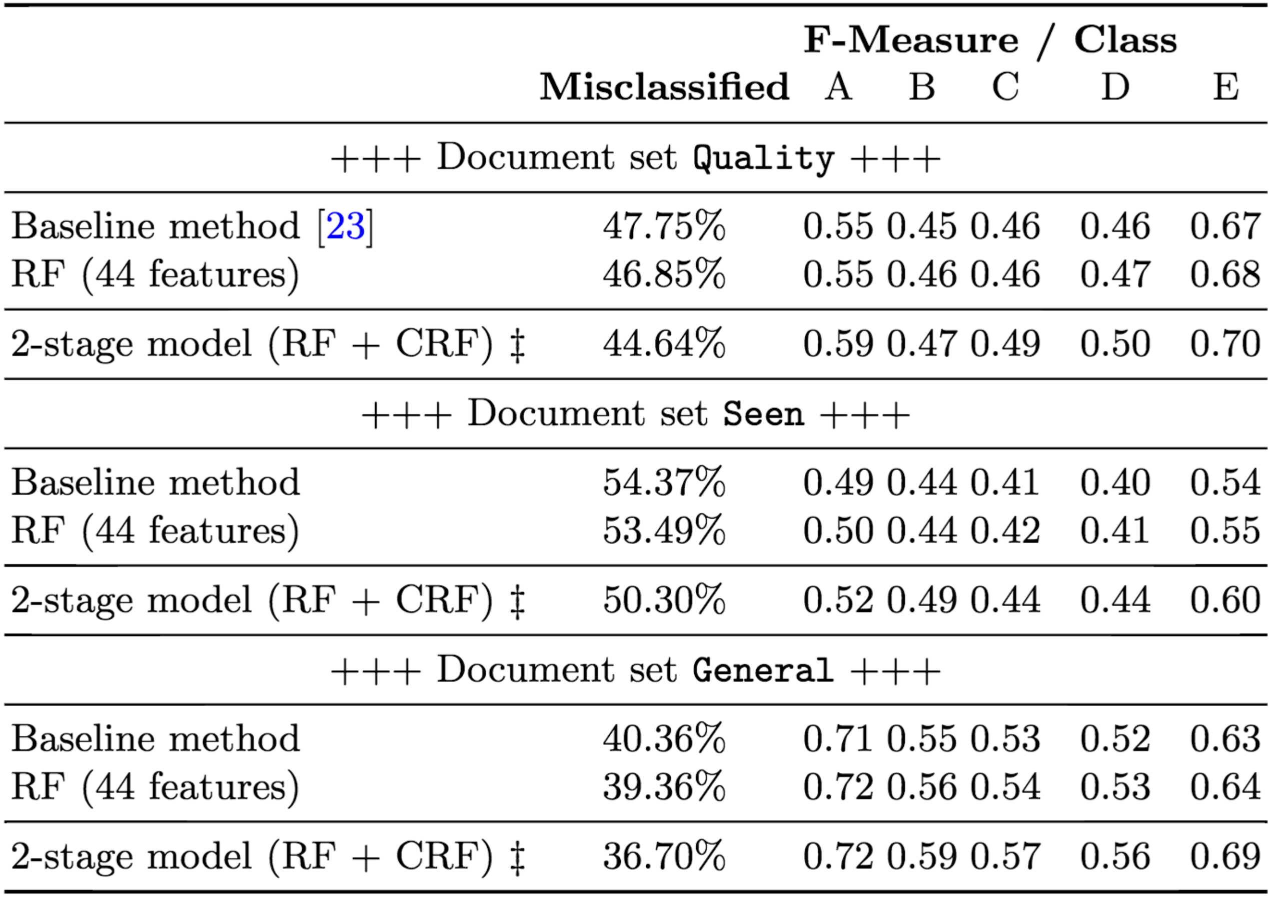
Temporal Inference

Classification Methods

- Baseline: **RF** with **21 features** about sub-documents statistics and temporal expressions.
- **RF** with extended features (**44** in total) which also consider explicit temporal expressions and verb tenses.
- Two-stage model: **CRFs** which use **RF results** as features and consider tags of **4 recent neighbors**.

Temporal Inference

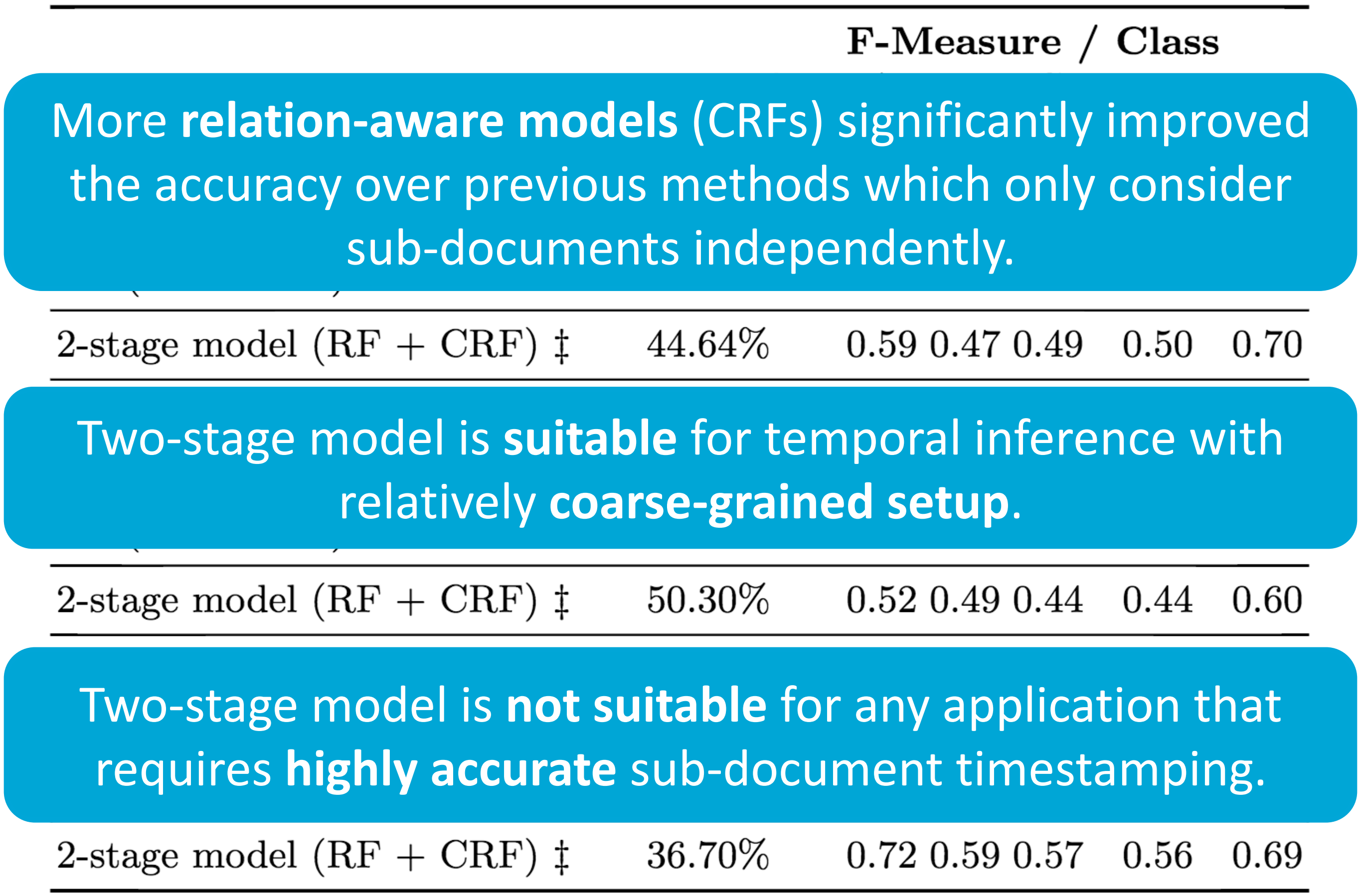
Classification Results



		F-Measure / Class				
	Misclassified	A	B	C	D	E
+++ Document set **Quality** +++						
Baseline method [23]	47.75%	0.55	0.45	0.46	0.46	0.67
RF (44 features)	46.85%	0.55	0.46	0.46	0.47	0.68
2-stage model (RF + CRF) ‡	44.64%	0.59	0.47	0.49	0.50	0.70
+++ Document set **Seen** +++						
Baseline method	54.37%	0.49	0.44	0.41	0.40	0.54
RF (44 features)	53.49%	0.50	0.44	0.42	0.41	0.55
2-stage model (RF + CRF) ‡	50.30%	0.52	0.49	0.44	0.44	0.60
+++ Document set **General** +++						
Baseline method	40.36%	0.71	0.55	0.53	0.52	0.63
RF (44 features)	39.36%	0.72	0.56	0.54	0.53	0.64
2-stage model (RF + CRF) ‡	36.70%	0.72	0.59	0.57	0.56	0.69

Temporal Inference

Classification Results



		F-Measure / Class				
2-stage model (RF + CRF) ‡	44.64%	0.59	0.47	0.49	0.50	0.70
2-stage model (RF + CRF) ‡	50.30%	0.52	0.49	0.44	0.44	0.60
2-stage model (RF + CRF) ‡	36.70%	0.72	0.59	0.57	0.56	0.69

More **relation-aware models** (CRFs) significantly improved the accuracy over previous methods which only consider sub-documents independently.

Two-stage model is **suitable** for temporal inference with relatively **coarse-grained setup**.

Two-stage model is **not suitable** for any application that requires **highly accurate** sub-document timestamping.

Conclusions

- A large proportion of Web documents do have sub-documents with different timestamps.
    - In general, about half of documents have 2+ timestamps.
    - Most Web documents have < 10 timestamps.
    - Timespan of sub-documents are really large.
    - The more creation timestamps, the less initial content.

- Our two-stage model are suitable for temporal inference with coarse-grained setup.
    - **63.3%** accuracy on coarse-grained classification.

- Future work will focus on the improvement of the sub-document timestamping pipeline in order to be able to reliably timestamp all of the Web (or more realistically all of ClueWeb12).

Questions?

Thanks

Yue Zhao y.zhao-1@tudelft.nl