

Citation Sentiment

Workshop on Open Citations (Sep 2018)

Work by David Ciudad, presented by Daniel Ecer

Data Scientist

Outline

About eLife

Project Overview

Athar Dataset

Inferring Sentiment from Retractions

Summary

Next

About eLife

@eLife @eLifeInnovation

“

Creativity, imagination, and
doing the **experiments**.
That's what **eLife** is all about.

”

-- Sir Mark Walport, UK Science Advisor
Founding member, eLife Board of Directors



eLife is a non-profit organisation inspired by research funders and led by scientists

Funding, exposure and mentorship

Supporting the open source community through the eLife Innovation Initiative

Get involved elifesci.org/innovate



#eLifeSprint. Credit: Orquidea Real Photobook - Julieta Sarmiento Photography @orquidea.real.pho



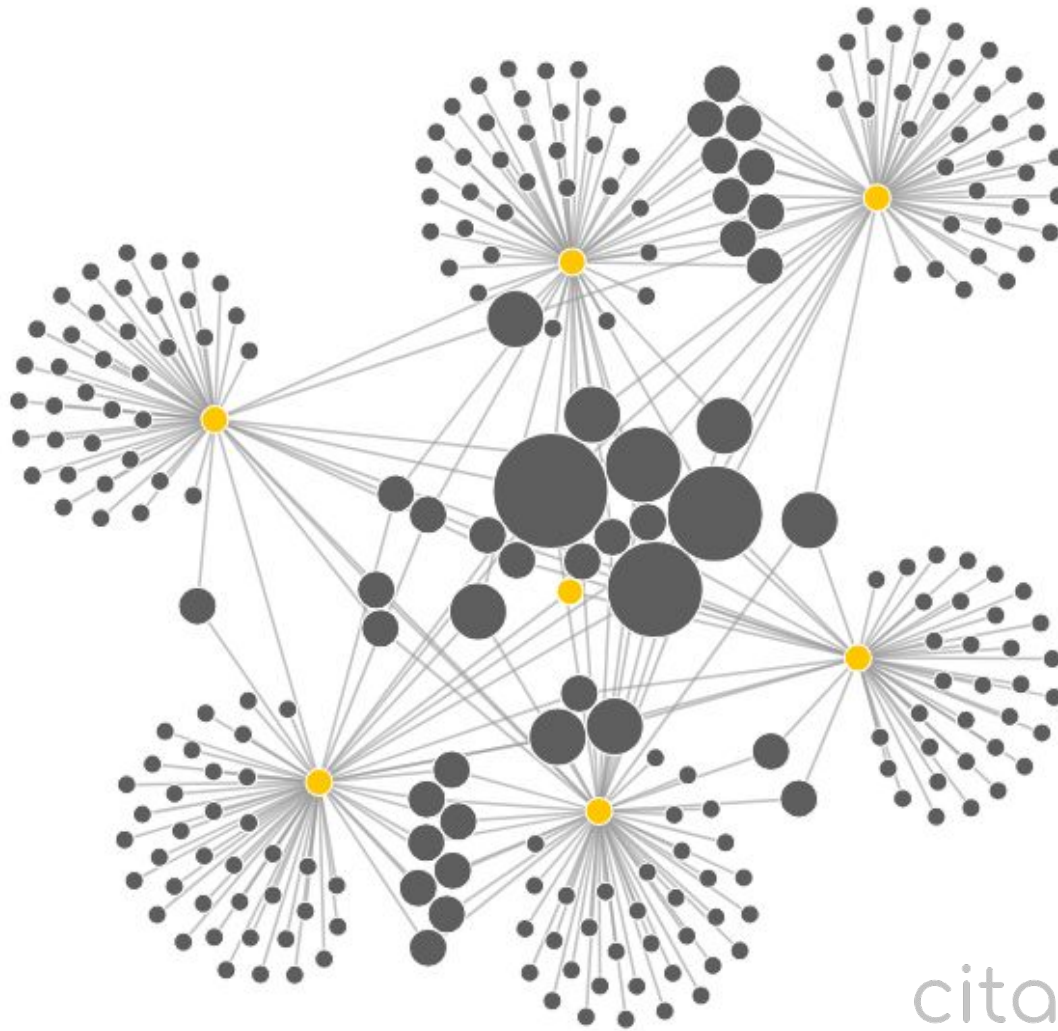
Project Overview

Work by David Ciudad

- ASI Fellowship project:
Investigating the context of citations
- Project continuation:
Investigate criticism / sentiment of citations
(preliminary results)



Example citation network visualisation



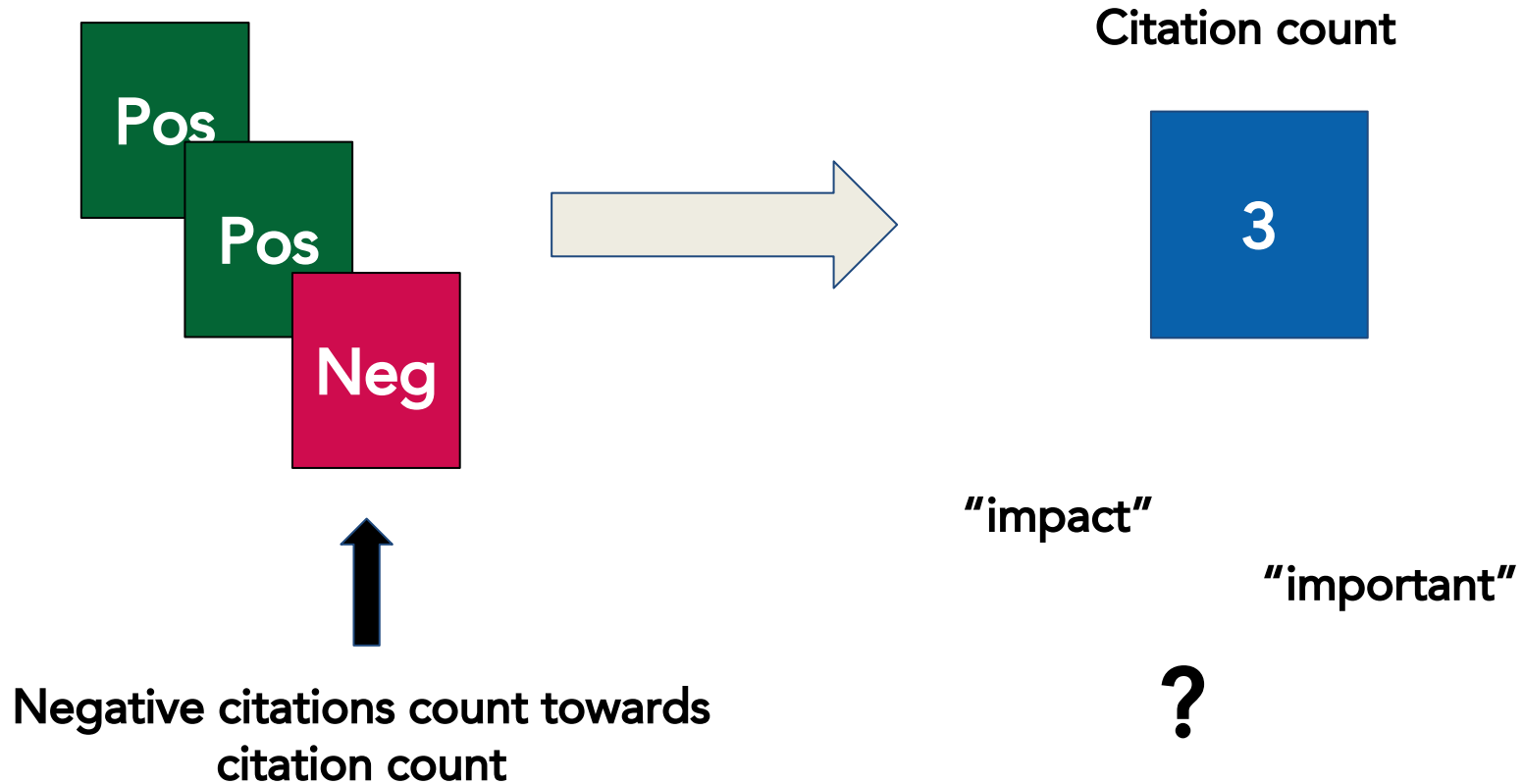
citation **gecko** 

Example citation

“Our results contrast with the high rate of XMRV detection reported by *Lombard et al.* among both CFS patients and controls, but are in agreement with recent data reported in two large studies in the UK and in the Netherlands...”

(Switzer et al., 2010)

Why does the sentiment matter?



What do we mean by sentiment?

- Polarity:
 - Positive
 - Neutral
 - Negative (criticism)

Off-the-shelf sentiment analysis...

"Our results contrast ..."

< 0

+0.16

0.0

Expected

Textblob

NLTK
Vader

> 0

+0.07

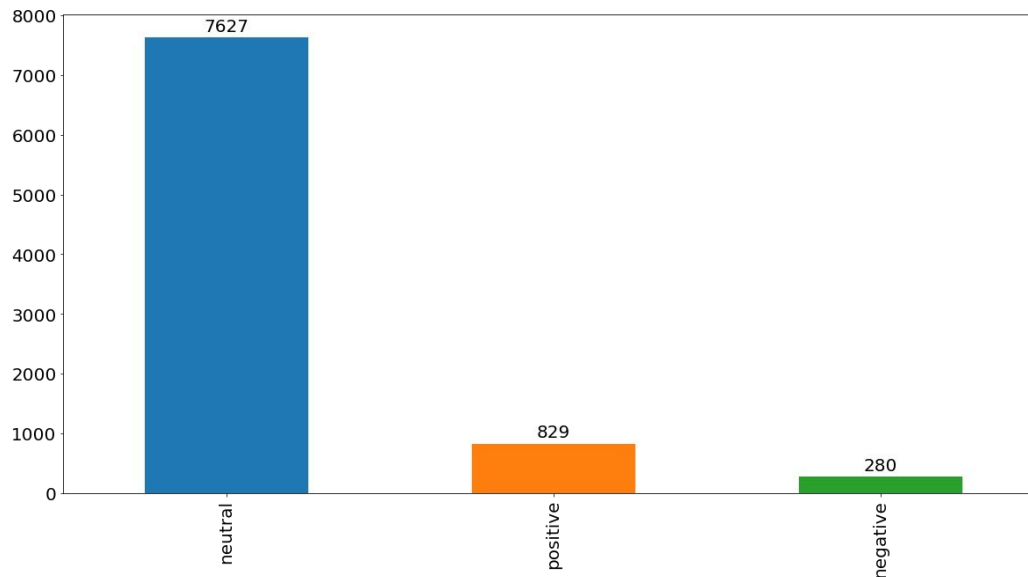
+0.18

"but are in agreement ..."

Athar Dataset

Athar Dataset

- “Sentiment Analysis of Citations using Sentence Structure-Based Features” (June 2011, Awais Athar)
- <http://cl.awaisathar.com/citation-sentiment-corpus/>
- 8736 manually annotated citations



Athar dataset cleaning

- There are 37 records with a citation text > 1000 characters (text shows missing sentence separator)
- Some citation texts are missing word separations or include single character tokens
- Filter tokens, keep only:
 - known English words
 - that are at least three characters
 - except the word "no"
(we don't remove most nltk stop words as they might have an influence on sentiment, e.g. "won't")

Preliminary Results - predict negative vs neutral

- Naive Bayes (bag of words):
 - ~0.77 (0.84 for more biased subset)
 - Words are stemmed

Negative	Neutral
outperform	bilingu
while	similar
no	updat
although	constitu
limit	posit
small	follow
show	describ

Table: Most indicative word (stems)

Athar Dataset Conclusion

- Dataset too small (only 280 negative citations)
- “Messy” data (could be cleaned up via source ACL Anthology Network corpus?)

Inferring Sentiment from Retractions

Retraction Watch Database

Retraction or Other Notices Title/Subject(s)/Journal — Publisher/Affiliation(s)/Retraction Watch Post URL(s) 600 Items Displayed Out of 18956 Item(s) Found	Reason(s)	Author(s)	Original Paper Date/PubMedID/DOI	Retraction or Other Notices Date/PubMedID/DOI
A robust technique based on VLM and Frangi filter for retinal vessel extraction and denoising (HSC) Medicine - Ophthalmology; (HSC) Radiology/Imaging; <i>PLoS One</i> — PLoS Department of Telecommunication Engineering, The Islamia University Bahawalpur, Pakistan Department of Electronic Engineering, International Islamic University, Islamabad, Pakistan Al-Khwarizmi Institute of Computer Science, UET Lahore, Pakistan	+Duplication of Article	Khan Bahadar Khan Amir A Khaliq Abdul Jalil Muhammad Shahid	02/18/2018 29432464 10.1371/journal.pone.0192203	08/29/2018 00000000 10.1371/journal.pone.0203418
Losing control: Mostly incongruent lists postpone, but do not eliminate, the Stroop effect (SOC) Psychology; <i>Attention, Perception & Psychophysics</i> — Springer Department of Psychology, Saint Mary's University, Halifax, Canada Department of Psychology and Neuroscience, Dalhousie University, Halifax, Canada	+Error in Analyses +Error in Results and/or Conclusions	Jason Ivanoff Nicole E Webb Harjot Chahal Virginia P Palango Raymond M Klein Steven R Carroll	03/08/2018 29520714 10.3758/s13414-018-1496-9	08/28/2018 00000000 10.3758/s13414-018-1591-y
A new approach to the mass production of titanium carbide, nitride and carbonitride whiskers by spouted bed chemical vapor deposition (B/T) Business - Manufacturing; (PHY) Chemistry; <i>Materials Letters</i> — Elsevier Department of Materials Science and Engineering, Tsinghua University, Beijing, 100084, China	+Date of Retraction/Other Unknown +Notice - Limited or No Information +Withdrawal	Jinsheng Pan Ruixiang Cao Yongwen Yuan	03/15/2006 00000000 10.1016/j.matlet.2005.02.092	08/27/2018 00000000 10.1016/j.matlet.2005.02.092
Intrathoracic transmural esophageal perforation (Boerhaave syndrome): Challenges in management of the delayed presentation (HSC) Medicine - Gastroenterology; (HSC) Medicine - Otorhinolaryngology; <i>Journal of Trauma and Acute Care Surgery</i> — Wolters Kluwer Denver Health Medical Center, University of Colorado School of Medicine, Department of General Surgery, Aurora, Colorado St. Joseph Hospital, Denver, Colorado	+Error in Text +Retract and Replace	Irada Ibrahim-Zada Phil Ernest Ernest E Moore	11/01/2017 29073118 10.1097/TA.0000000000001662	08/26/2018 30142106 10.1097/TA.0000000000002052
Clinical Importance of Somatostatin Receptor 2 (SSTR2) and Somatostatin Receptor 5 (SSTR5) Expression in Thyrotropin-Producing Pituitary Adenoma (TSHoma)	+Plagiarism of Article	Benxia Yu Zhongsheng Zhang Hao Song Xuebin Shi	04/23/2017 28434012 10.12659/MSM.903377	08/24/2018 30142144 10.12659/MSM.912715

<http://retractiondatabase.org/>

Retraction Note example


Retraction Note to: The *BRCA2* variant c.68–7 T > A is associated with breast cancer

Pål Møller [1,2,3](#) ✉ and Eivind Hovig [2,4,5](#)

Hereditary Cancer in Clinical Practice 2018 **16**:10

<https://doi.org/10.1186/s13053-018-0093-1> | © The Author(s). 2018

Received: 16 April 2018 | Accepted: 16 April 2018 | Published: 2 May 2018

 The [original article](#) was published in *Hereditary Cancer in Clinical Practice* 2017 15:20

Retraction

This article [[1](#)] has been retracted at the request of the authors. Upon re-review of the data, the authors identified coding errors in this study. Due to an error in the SQL query, the conclusions drawn in the article are incorrect. A re-examination of the data shows that there is no association between familial breast cancer and the *BRCA2* variant c.68–7 T > A. Another recent study suggests that the variant is not pathogenic [[2](#)]. All authors agree to this retraction.

Retraction Note references

References

1. Møller P, et al. The BRCA2 variant c.68-7 T>a is associated with breast cancer. Hered Cancer Clin Pract. 2017;15:20.
[View Article](#) [PubMed](#) [PubMed Central](#) [Google Scholar](#)
2. Colombo M, et al. The BRCA2 c.68-7T > a variant is not pathogenic: a model for clinical calibration of spliceogenicity. Hum Mutat. 2018;39:729-41.
[View Article](#) [PubMed](#) [Google Scholar](#)

Retracted article

Criticising article

**Can we harvest criticising citations
from retraction notes?**

Retraction note - not linking to criticising papers

Retraction

The authors retract this article [1] following an investigation by Leiden University Medical Centre into the research activities of the last author. The investigation identified a discrepancy between the data reported in the article and the original collected data. The investigation committee concluded that this undermined the scientific basis of the publication and advised that the publication should be retracted.

The online version of this article contains the full text of the retracted article as electronic supplementary material (Additional file 1).

Reference

1. Suurmond J, Dorjée AL, Boon MR, Knol EF, Huizinga TWJ, Toes REM, et al. Mast cells are the main interleukin 17-positive cells in anticitrullinated protein antibody-positive and -negative rheumatoid arthritis and osteoarthritis synovium. *Arthritis Res Ther*. 2011;13:R150.

[PubMed Central](#) 

[View Article](#) 

[PubMed](#) 

[Google Scholar](#) 

Retraction note - not including any references

Retraction Note to: Cryptomycota: the missing link

[Krishna Bolla](#)[✉] and [Elizabeth Jane Ashforth](#)

[Author information](#) ► [Copyright and License information](#) ► [Disclaimer](#)

This retracts the article "[RETRACTED ARTICLE: Cryptomycota: the missing link](#)" in volume 3 on page 161.

Retraction Note to: *Protein Cell* 2012, 3(3): 161–162 DOI
[10.1007/s13238-012-2013-x](https://doi.org/10.1007/s13238-012-2013-x)

Go to: ☐

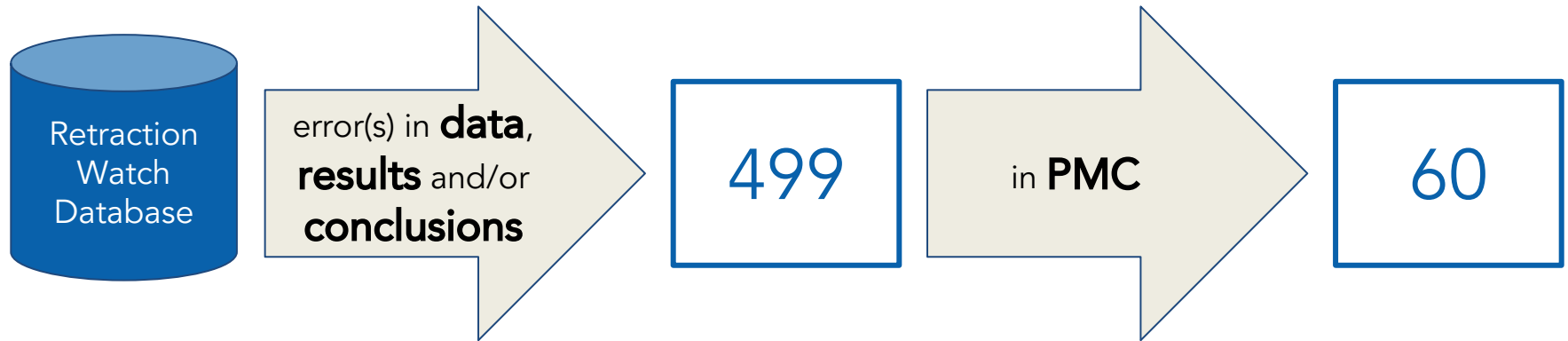
This article has been retracted. Although the article “James, T.Y., and Berbee, M.L. (2011). No jacket required new fungal lineage defies dress code. *Bioessays*. doi:10.1002/bies.201100110” was cited, Bolla K unintentionally copied smaller parts of the text. The authors apologize to the authors of the *BioEssays* paper and the readers of *Protein and Cell*.

Footnotes

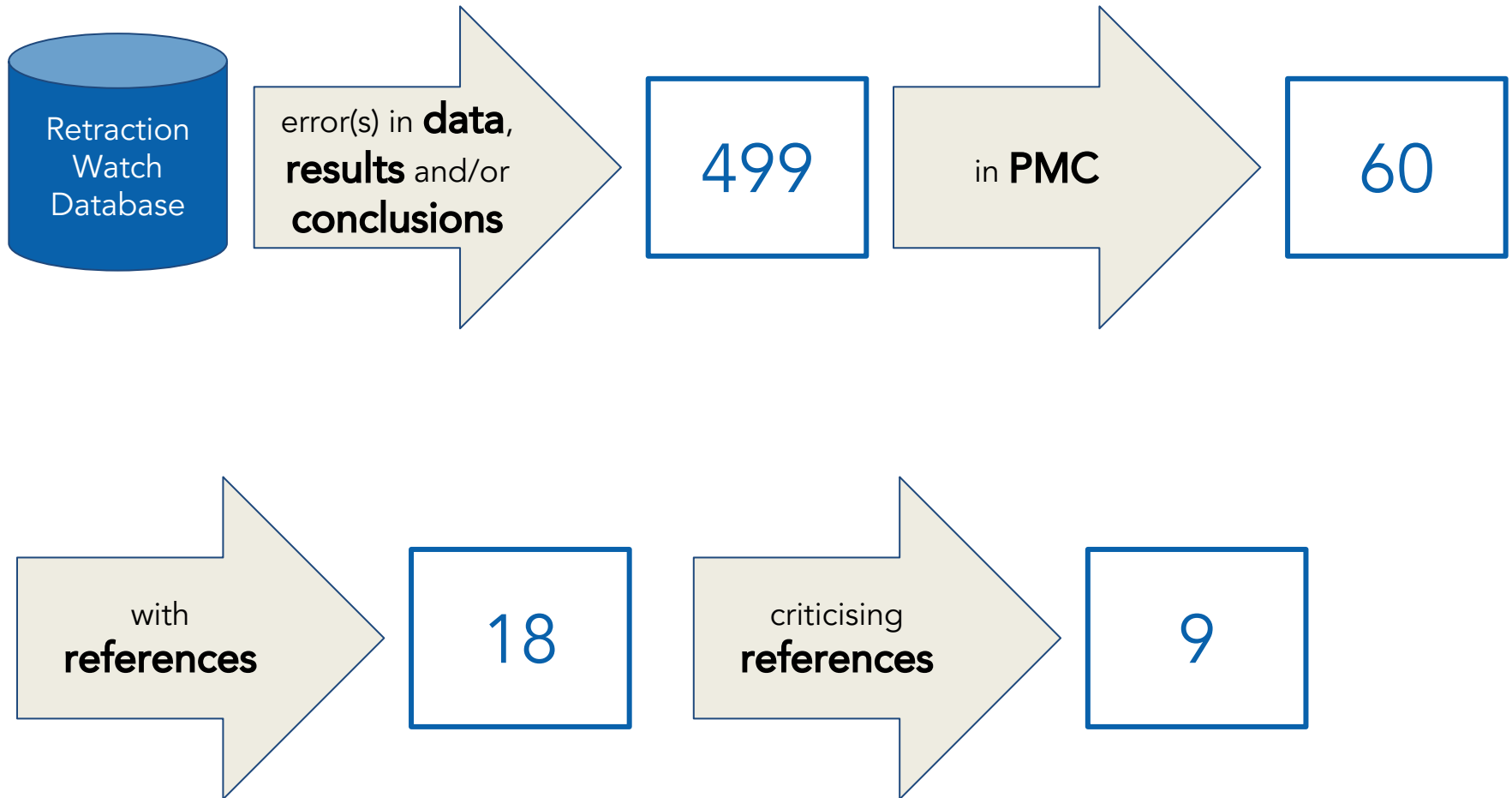
Go to: ☐

The online version of the original article can be found under doi:[10.1007/s13238-012-2013-x](https://doi.org/10.1007/s13238-012-2013-x).

Retraction Notes in PMC



Retraction Notes in PMC with Criticism



Retraction Notes (reality)

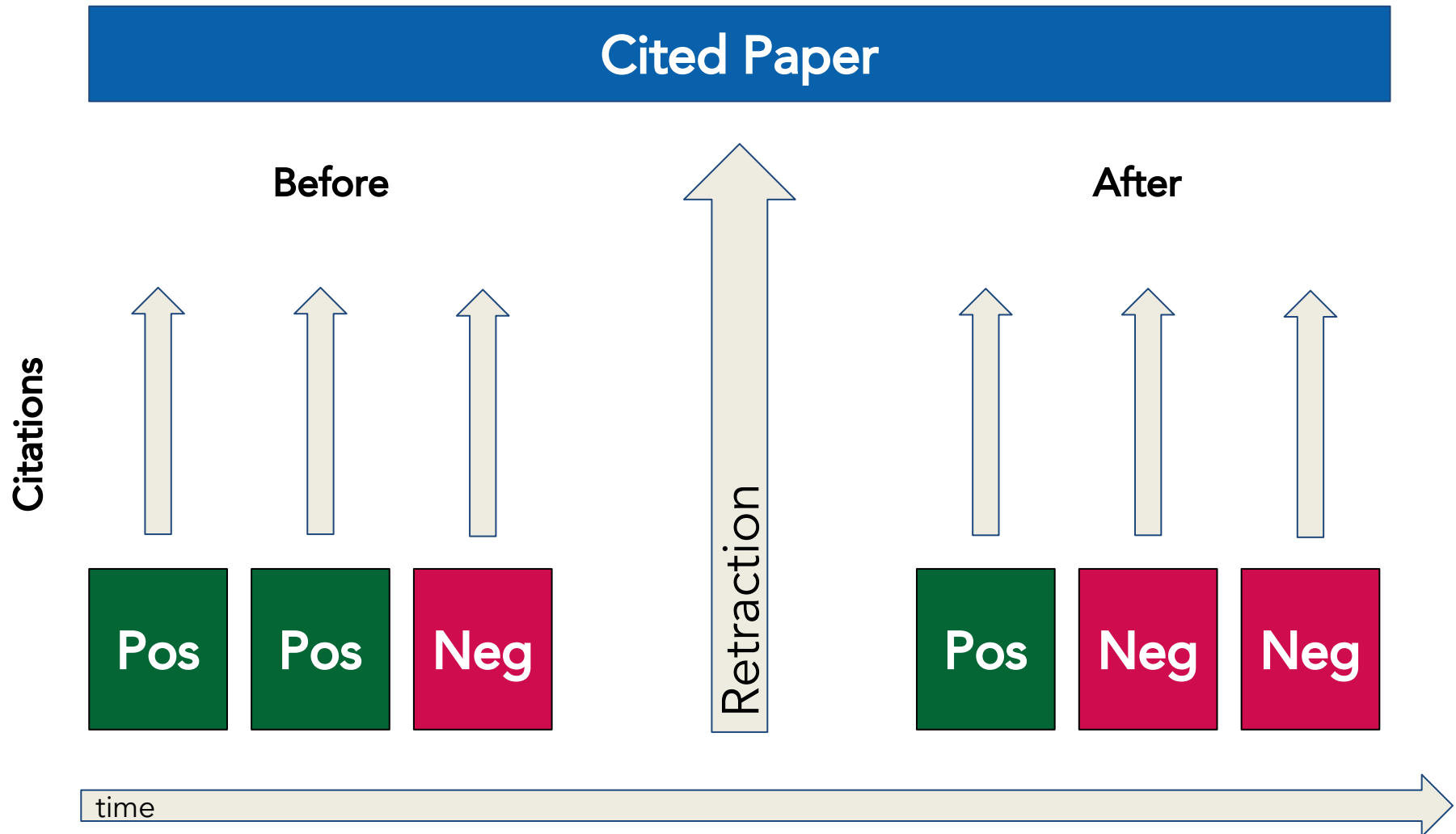
- Most retraction notes do not link to criticising papers
- Retraction Notes not always tagged accurately in PMC
- Retraction reason may not be machine readable

Can we still infer sentiment from retractions?

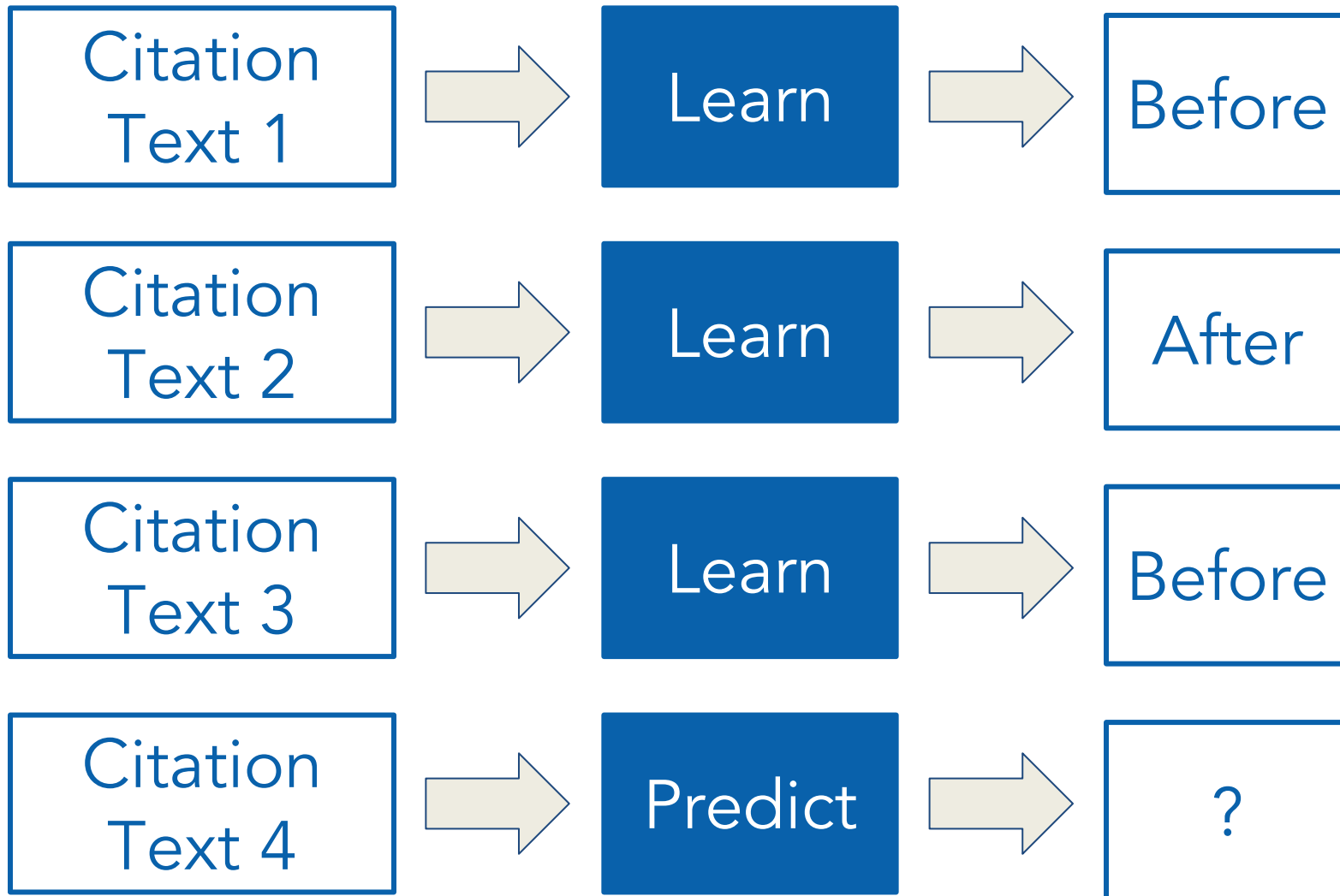
Hypothesis: Retraction ~ Negative sentiment

- **2009**: Paper **B1** published
- **2010**: Neutral citation: "These deficiencies in NK activity may increase viral load in CFS, incidentally a recent study observed increases in xenotropic murine leukemia virus-related virus (XMRV) in peripheral blood samples of CFS patients [**B1**]"
- **2011**: Paper **B1** retracted
- **2016**: Criticising citation: "The recent link between ME/CFS and xenotropic murine leukemia virus-related virus (XMRV) [**B1**] has largely been disproven and attributed to laboratory contaminants [**B2**]"

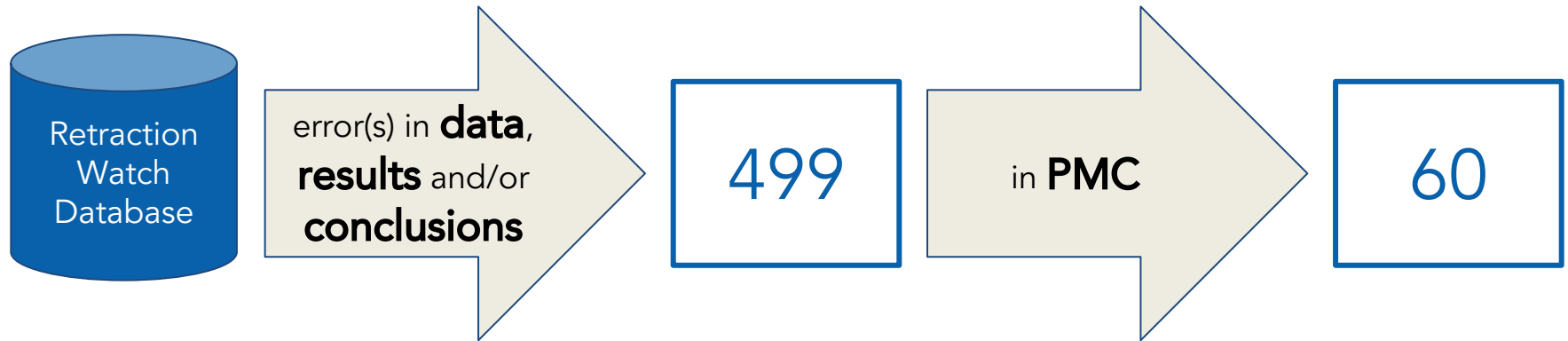
Are citations after retraction more likely negative?



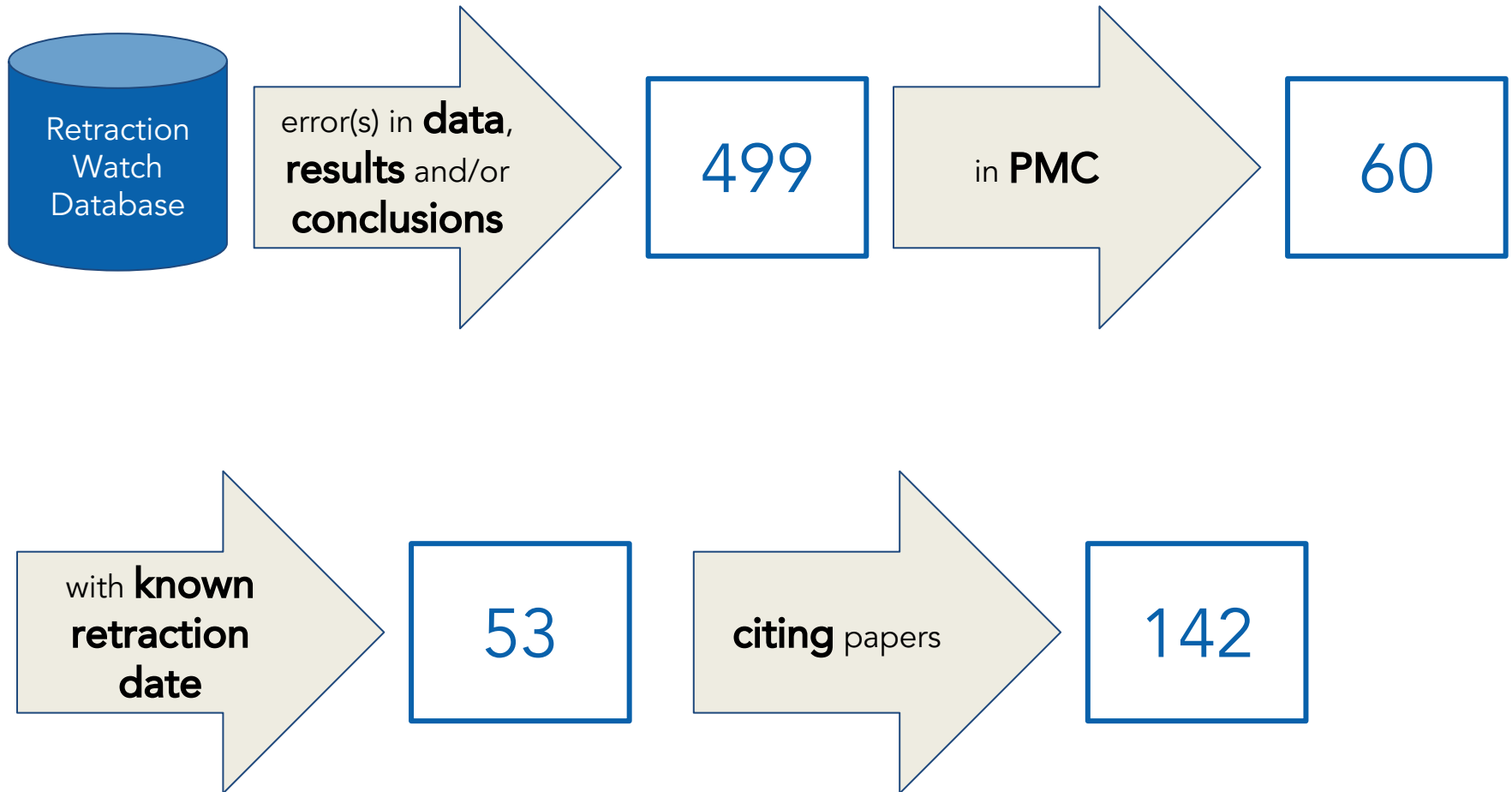
Predict: before / after retraction



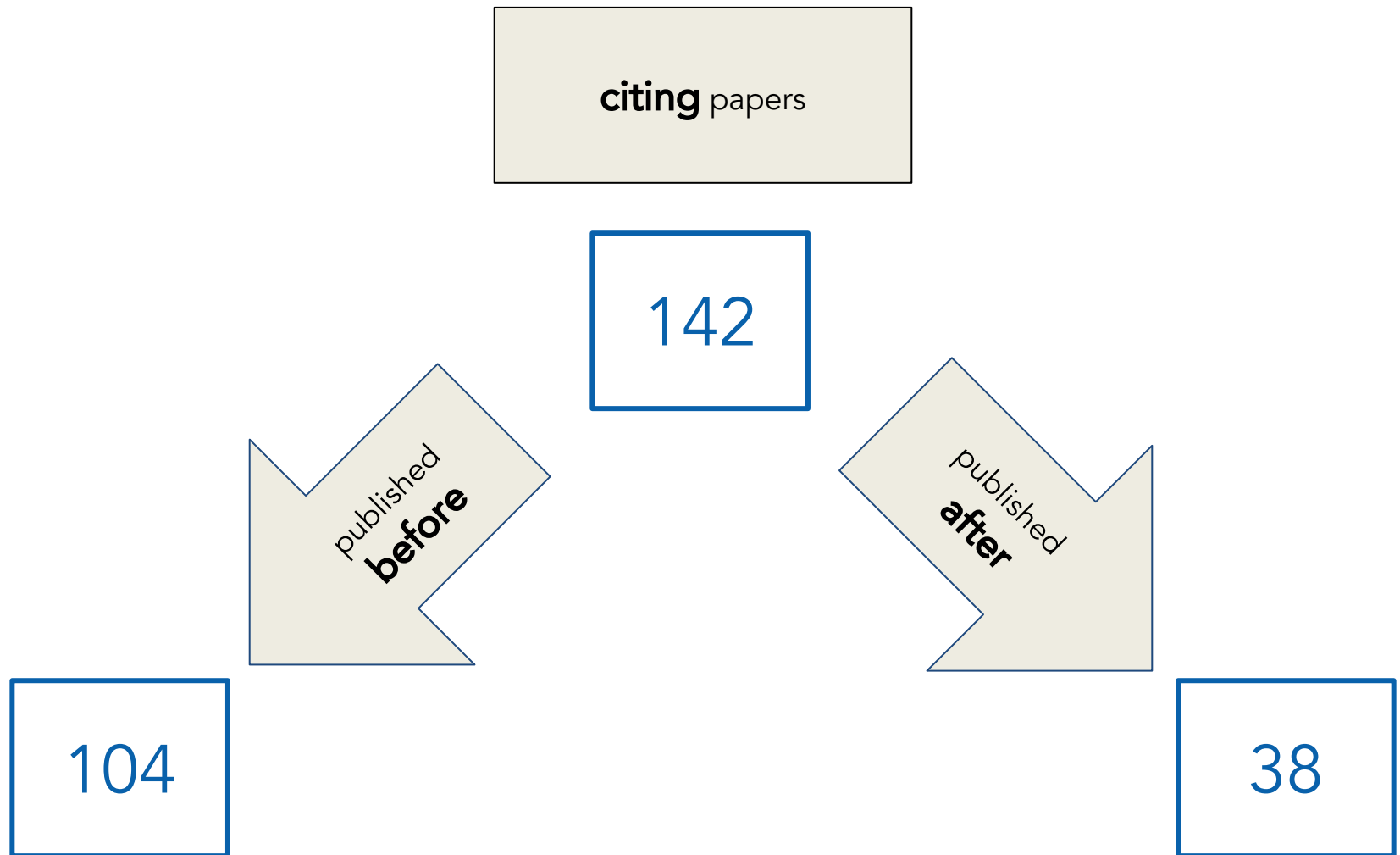
Retraction Notes in PMC (reminder)



Retraction Notes in PMC with known retraction date



Citing papers before/after date of retraction



Citation Text - Citing Sentence

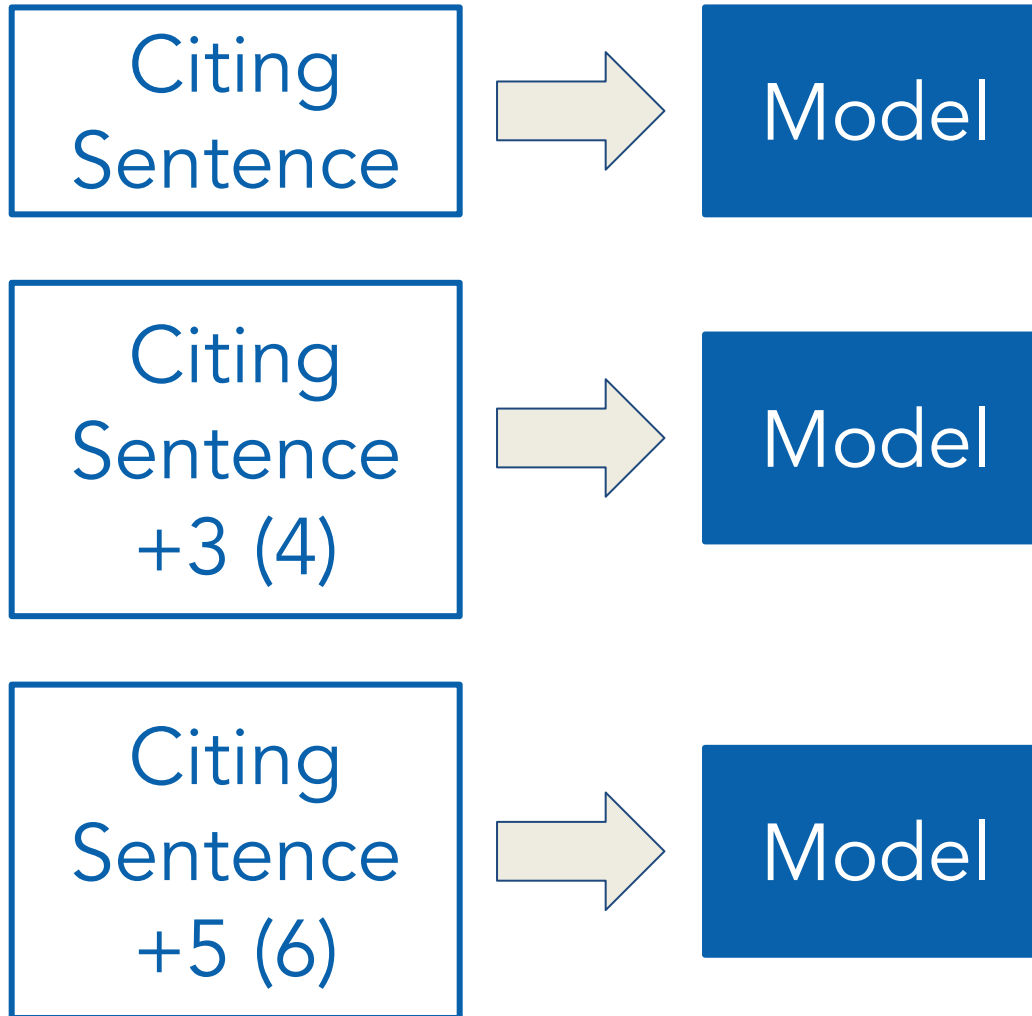
citing sentence

“Conversely, a recent article asserted that the variant was associated with breast cancer, based on a relatively limited case control association study in the Norwegian population (Møller & Hovig, 2017).”

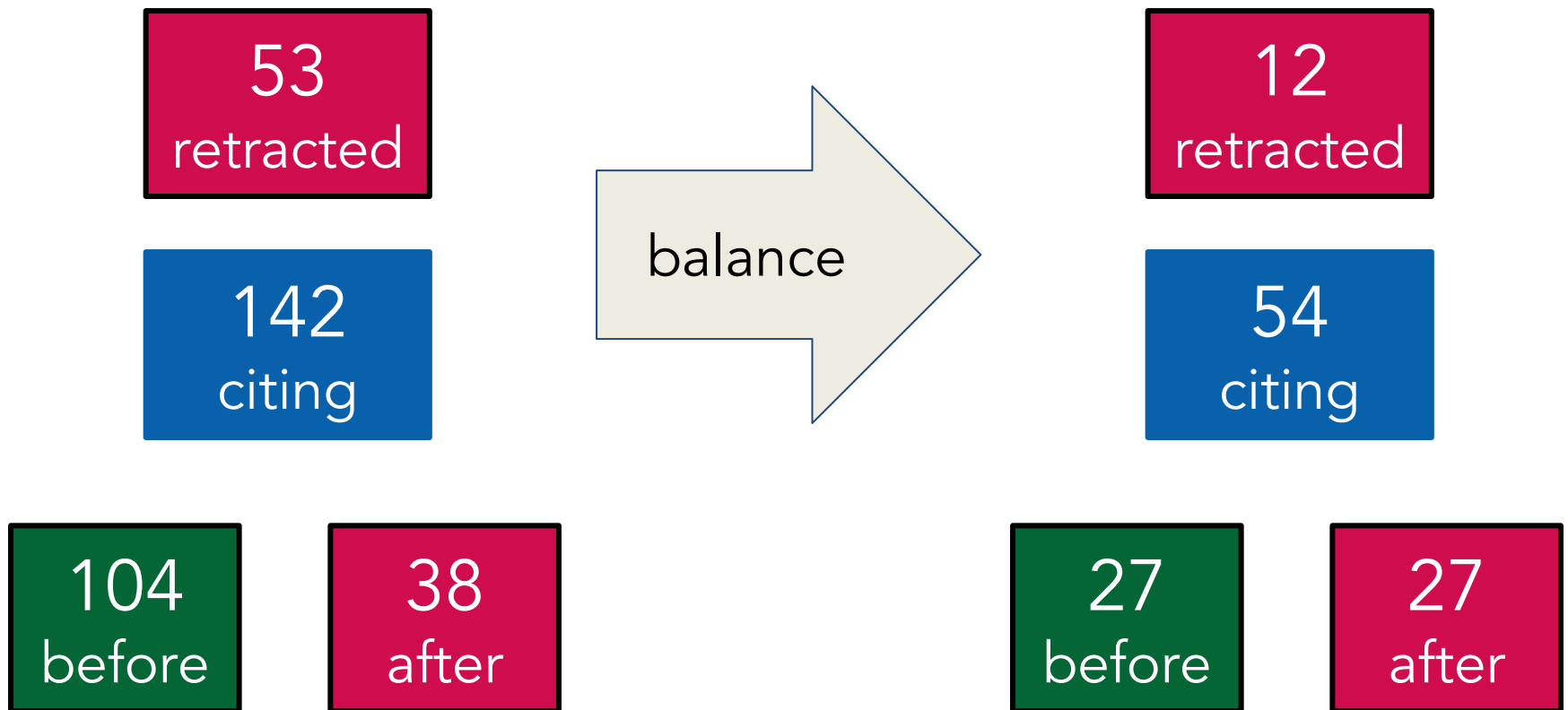
Citation Text - Citing Sentence + 3 (4 sentences)

- citing sentence
- +1
- +2
- +3
- "Conversely, a recent article asserted that the variant was associated with breast cancer, based on a relatively limited case control association study in the Norwegian population (Møller & Hovig, 2017).
- As a consequence, to date the classification of c.68-7T > A reported in databases aggregating information on genomic variations has remained inconclusive.
- In particular, ClinVar reports conflicting interpretations classifying the variant as benign (seven entries), likely benign (nine entries) and of uncertain significance (four entries).
- Moreover, the BIC database presently annotates the variant as of unknown clinical importance, pending classification, while the BRCA Share™ classifies it as likely benign."

Different models with citing sentences + context



Balance dataset (control for subject area etc.)



Detailed results*

	Undersampled	Undersampled + reduced bias
Citing Sentence	0.846	0.750
Citing Sentence +3 (4)	0.538	0.833
Citing Sentence +5 (6)	0.615	0.583

On closer inspection, used words like "cancer"
(indication of biased dataset)

More useful

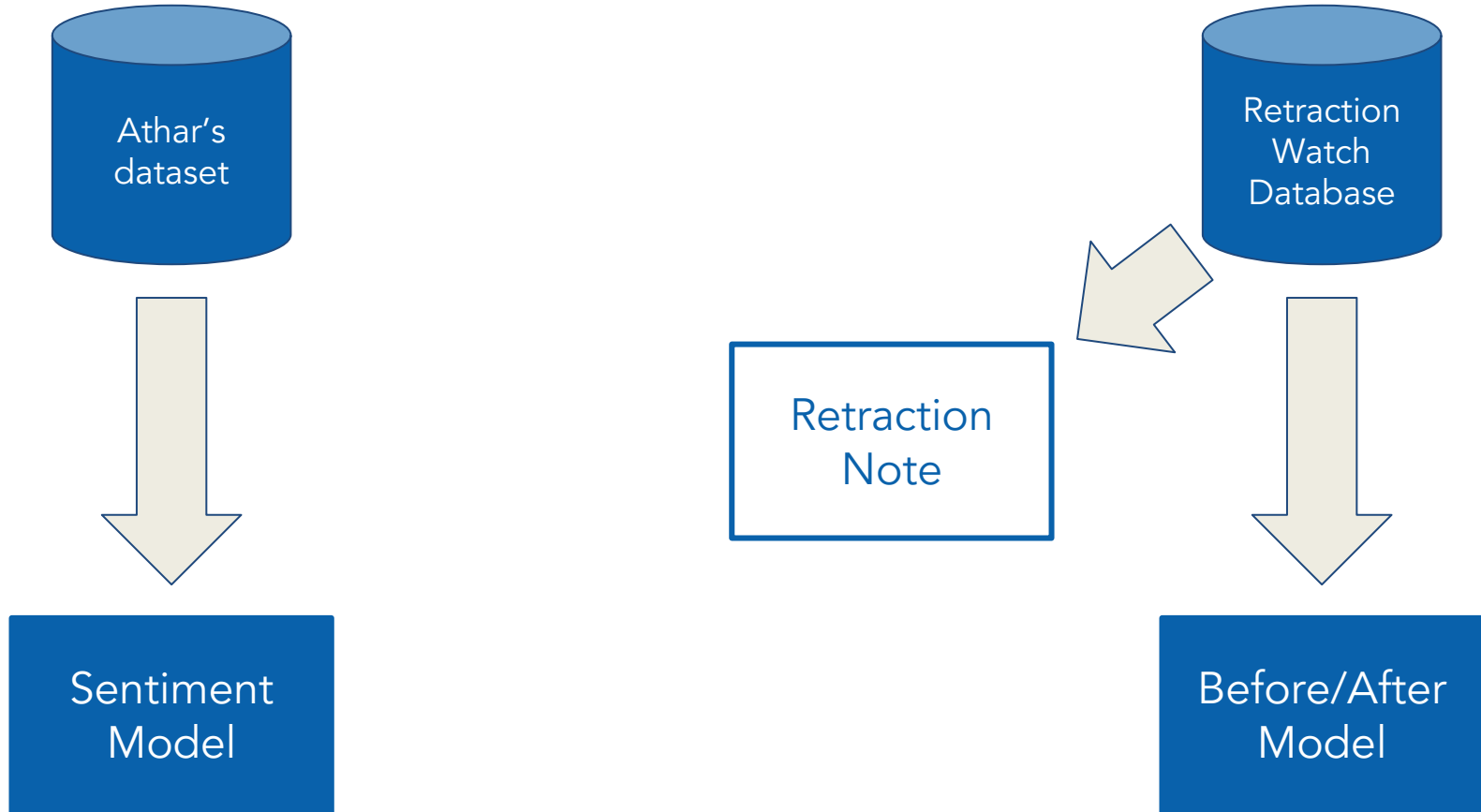
Top 7 stemmed words for 3-sentence model*

After	Before
remain	model
present	subject
assess	function
fail	lack
conflict	discoveri
induct	research
individu	no

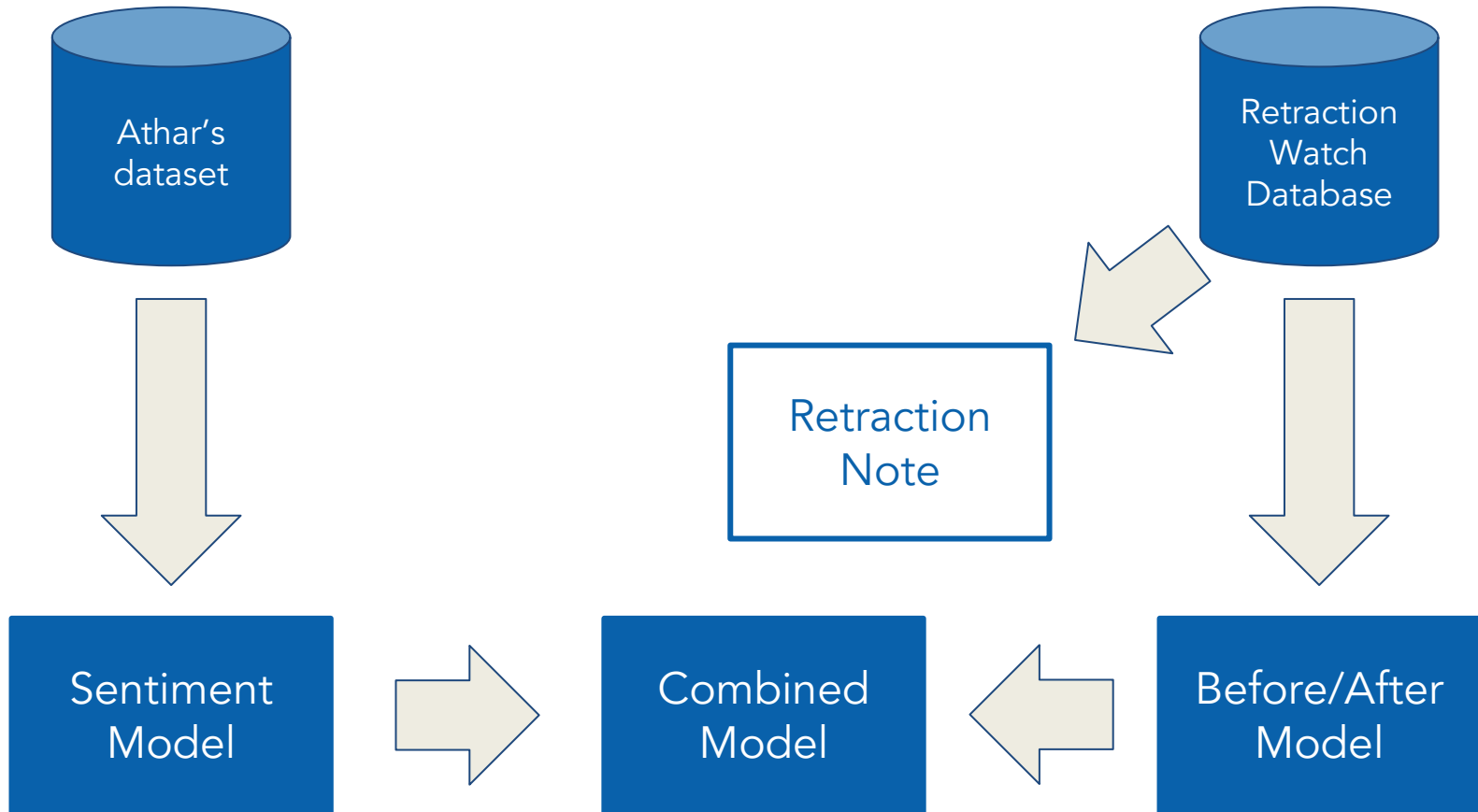
Table: Most indicative word (stems)

Summary

Summary



Summary (combined model)



Next

Next

- Getting more training/test data:
 - Pay for manual annotation?
 - Other ideas to automatically infer sentiment?
- Use cases for reliable sentiment detection of a citation:
 - Better metric? (Colin will talk more about metrics)
 - Identify contentious work?
 - <your suggestion here>
- More suggestions, feedback, collaboration please

Links

- https://github.com/dciudadr/eLife_retractions
- <https://github.com/elifesciences/citation-sentiment-analysis>
- [Blog: Investigating the context of citations](#)



Fin