Web Application Testing: Using Tree Kernels to Detect Near-duplicate States in Automated Model Inference

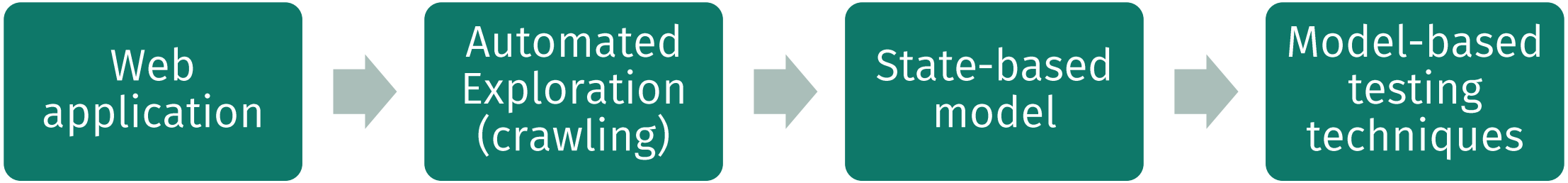
A. Corazza, S. Di Martino, A. Peron, and **Luigi Libero Lucio Starace**

Università degli Studi di Napoli Federico II, Naples, Italy

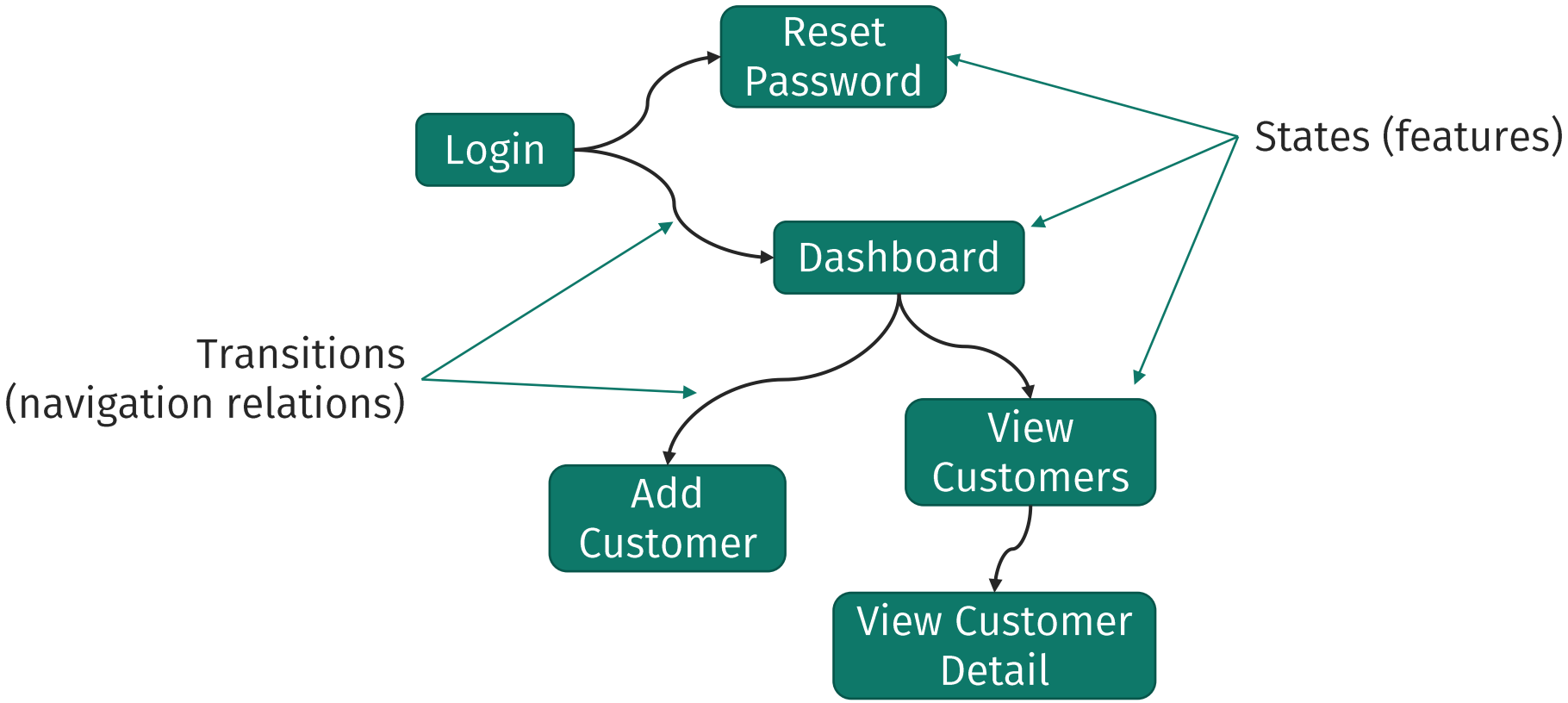
luigiliberolucio.starace@unina.it

ESEM ‘21 - October 12, 2021

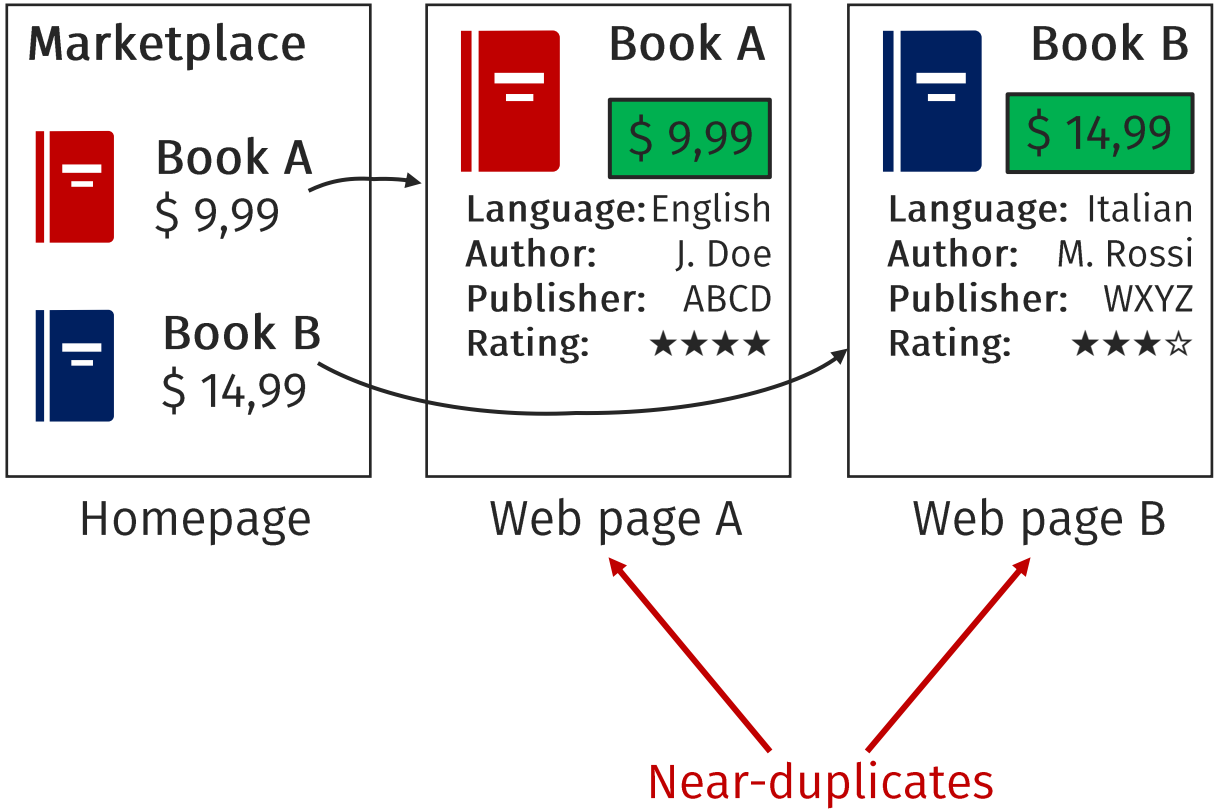
E2E testing of web applications



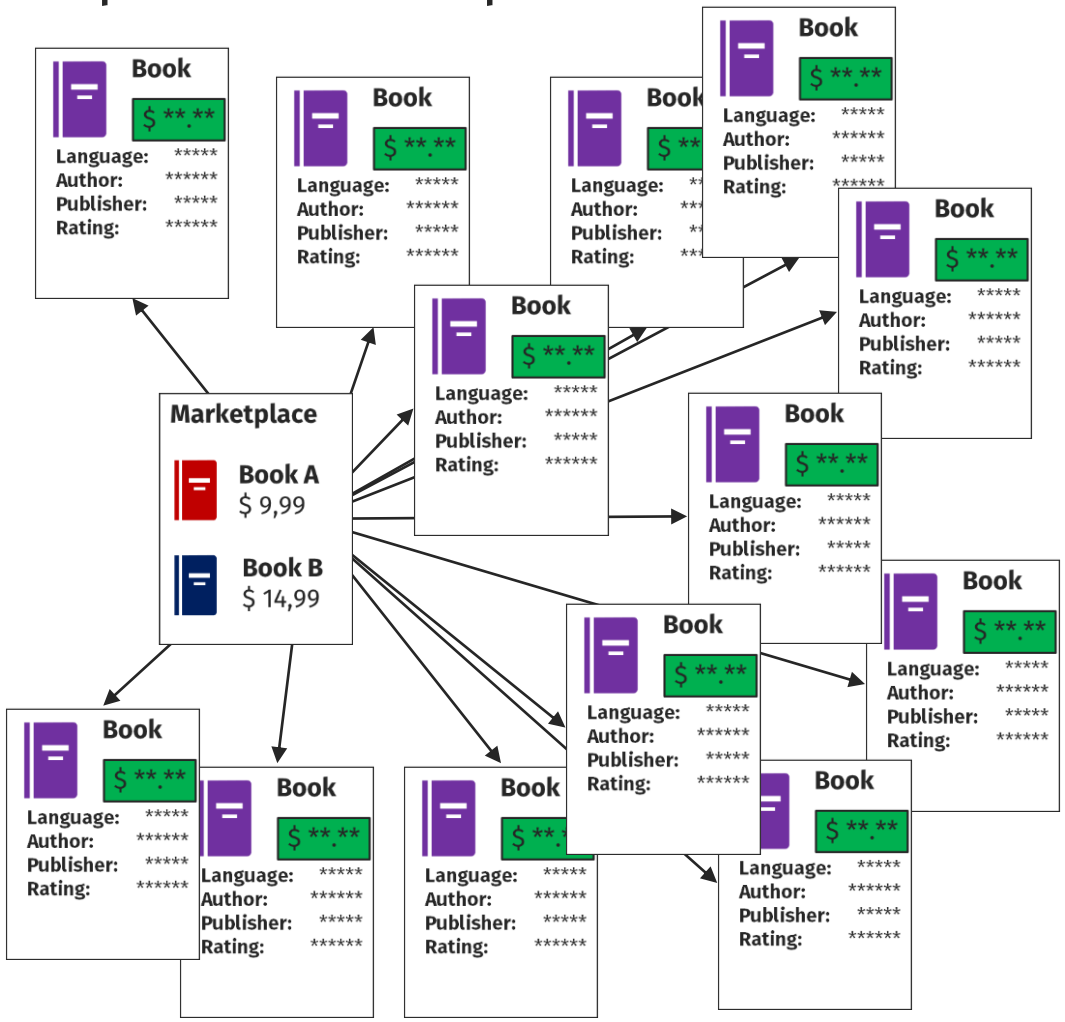
State-based web application models



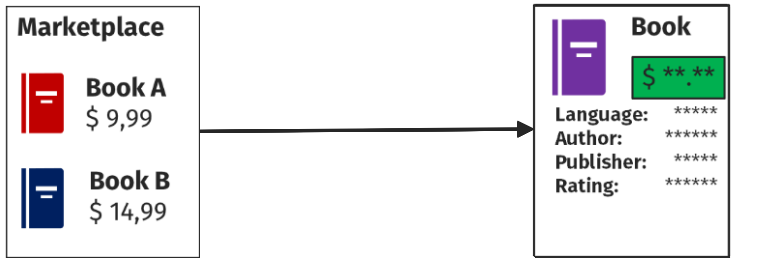
The near-duplicate problem



The near-duplicate problem



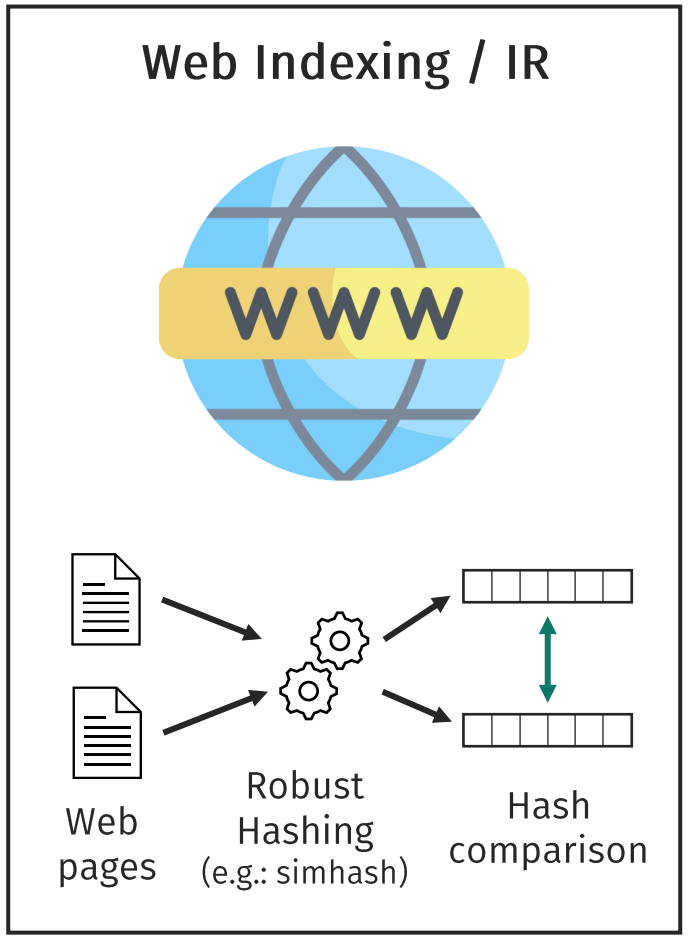
The challenge: detecting near-dupes

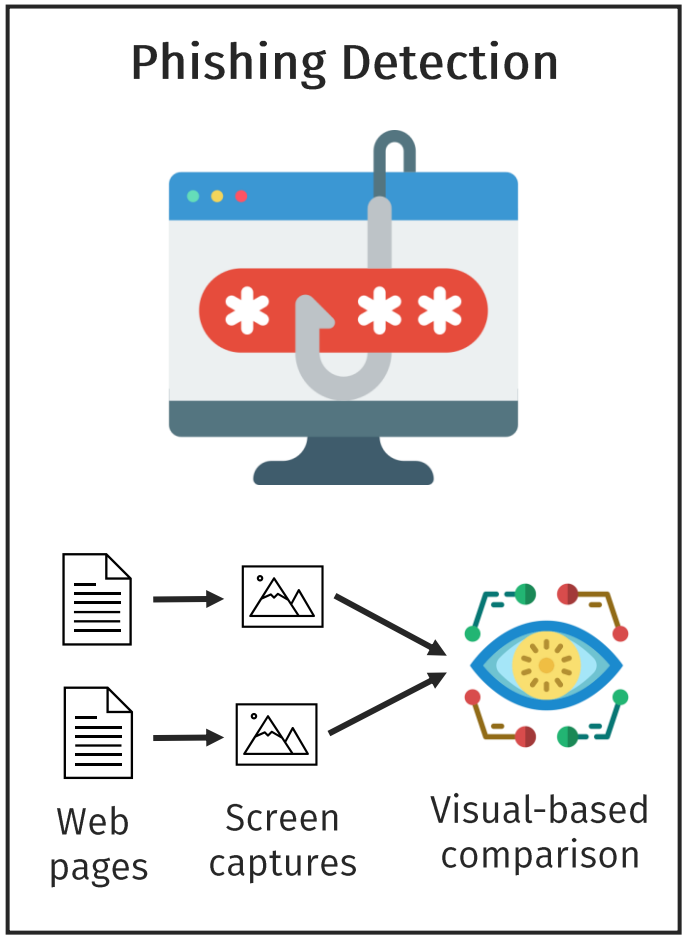


Catalog

Book detail page

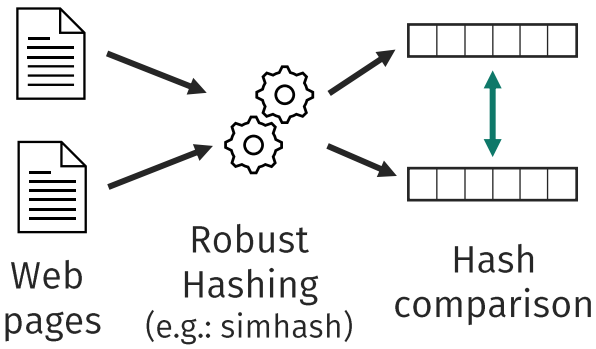
Related works and proposed solution



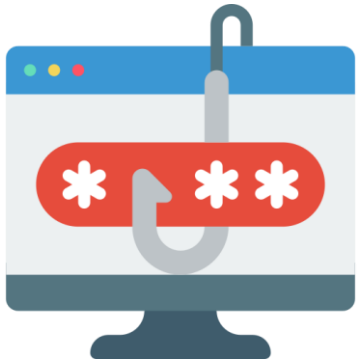


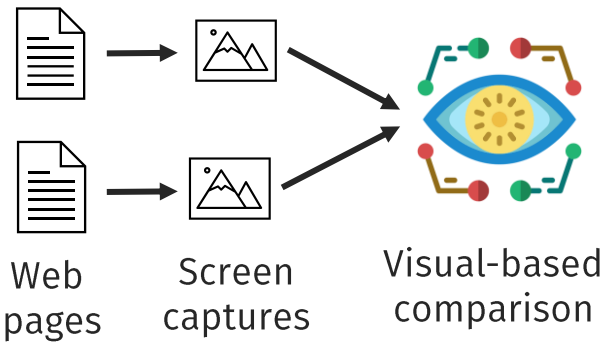
Related works and proposed solution

Web Indexing / IR

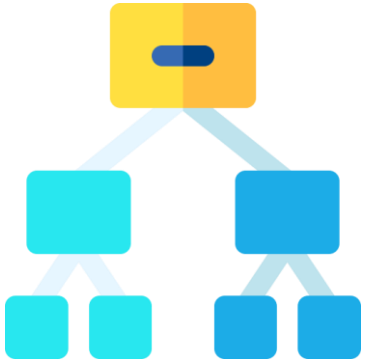


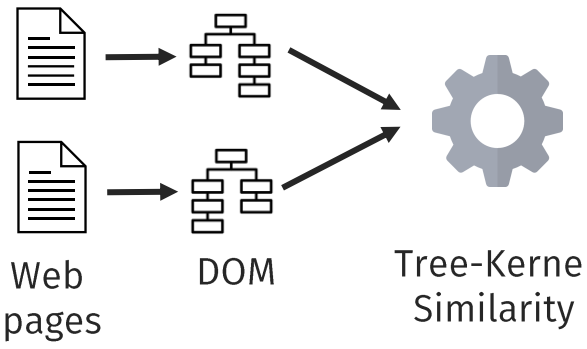
Phishing Detection



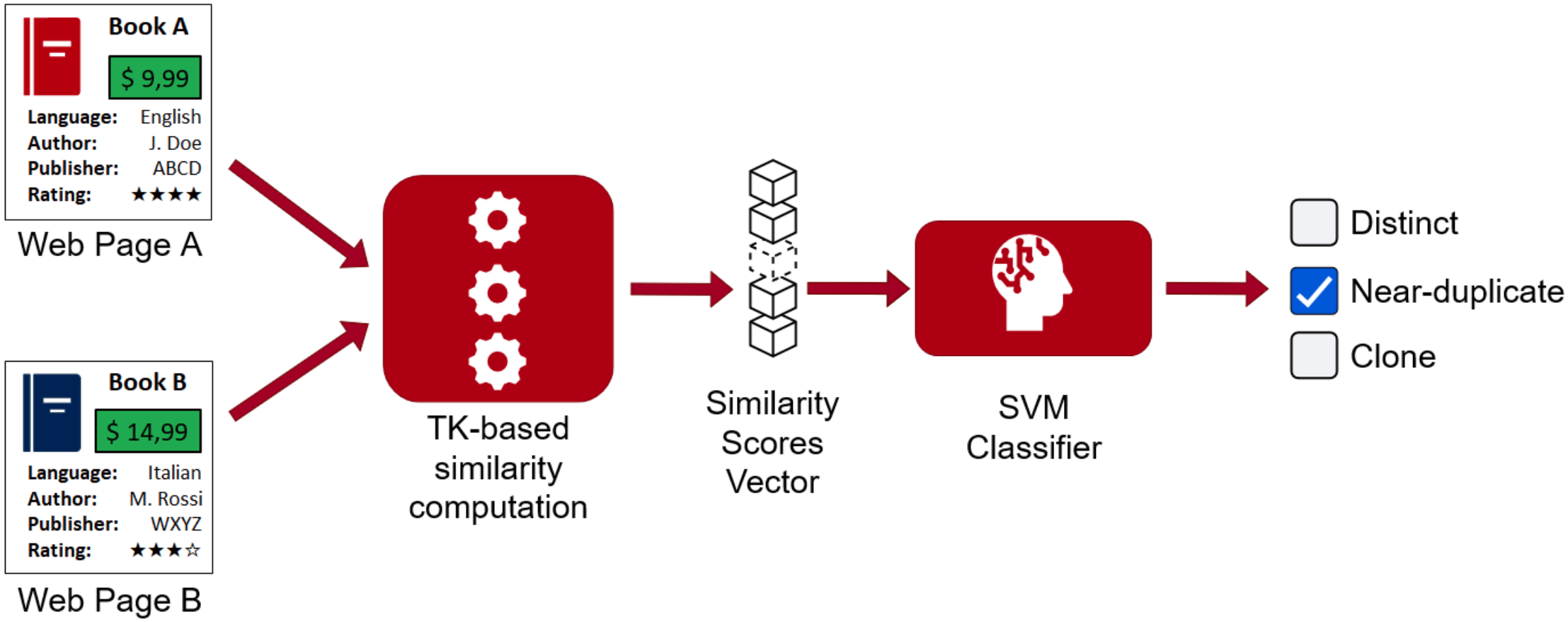


Our Approach





Tree Kernel-based near-dupes detection



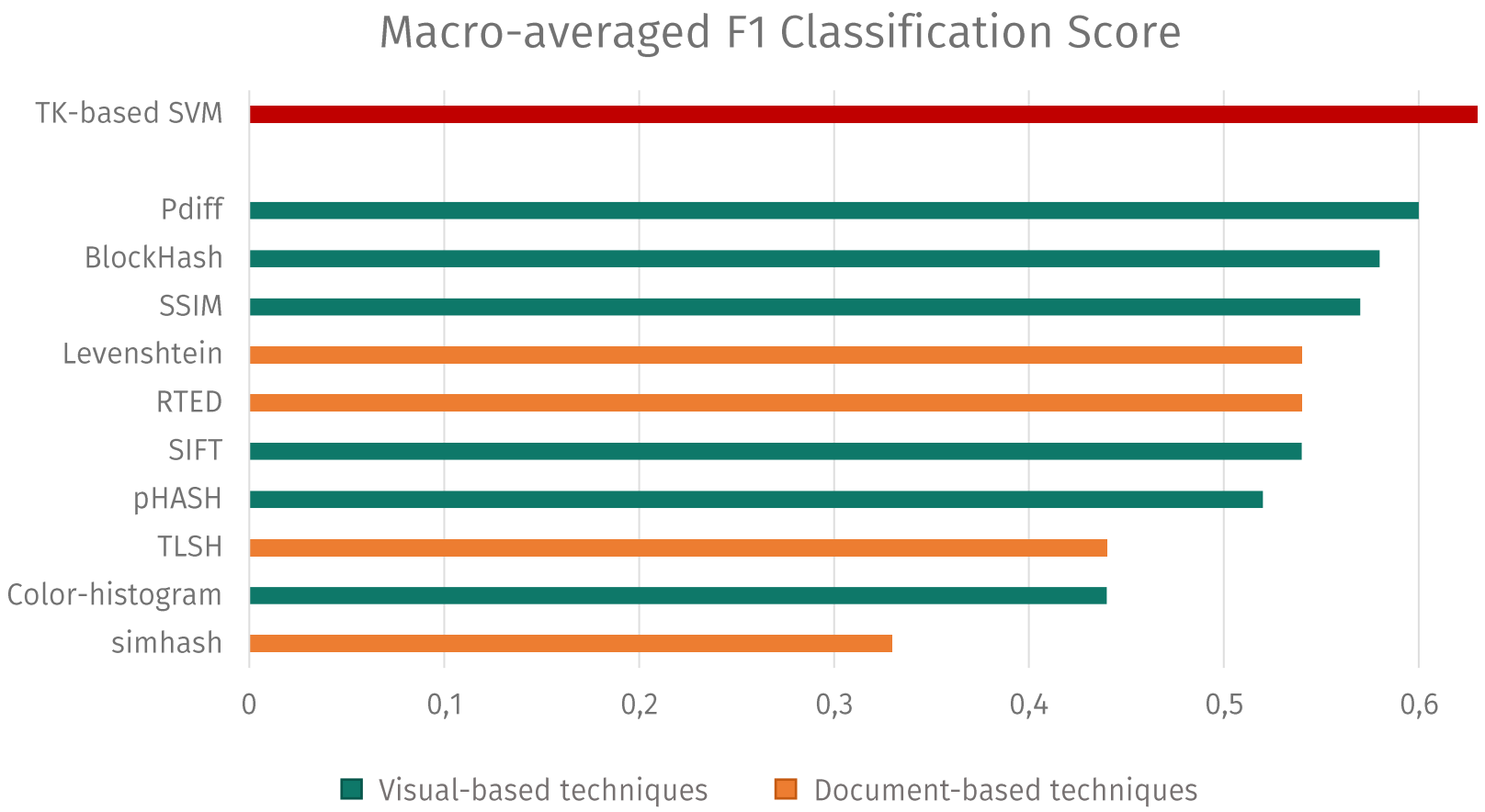
Experimental Framework

We're replicating the procedure presented at ICSE20 [1]

- Near-duplicate detection framed as a **classification task**
- Leverage the same massive dataset (~100k web page pairs)
- Measure performance using **F1 classification score**

[1] Yandrapally et al. "*Near-duplicate detection in web app model inference.*" *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*. 2020.

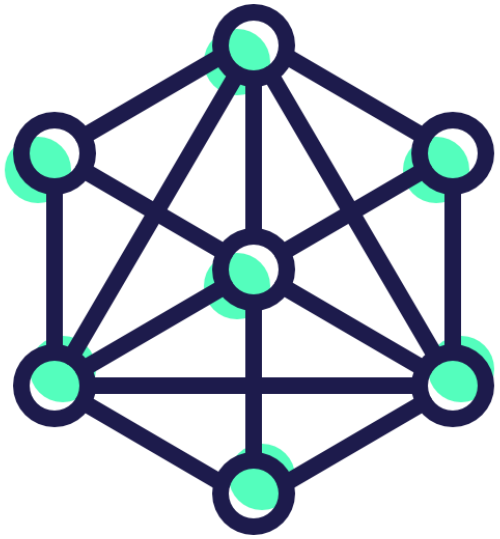
Emerging Results



3% improvement w.r.t. Pdiff (best, **visual-based** technique, but ~**5 times slower**)

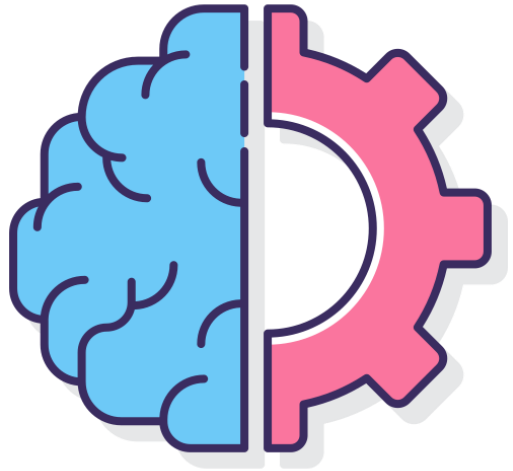
10% to **30% improvement** w.r.t. similar document-based techniques

Future research



Apply to Model Inference!

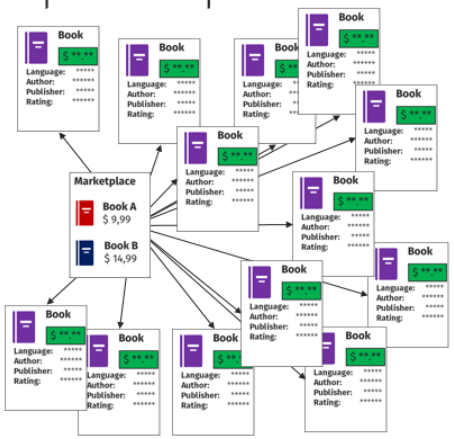
Custom Tree-Kernels



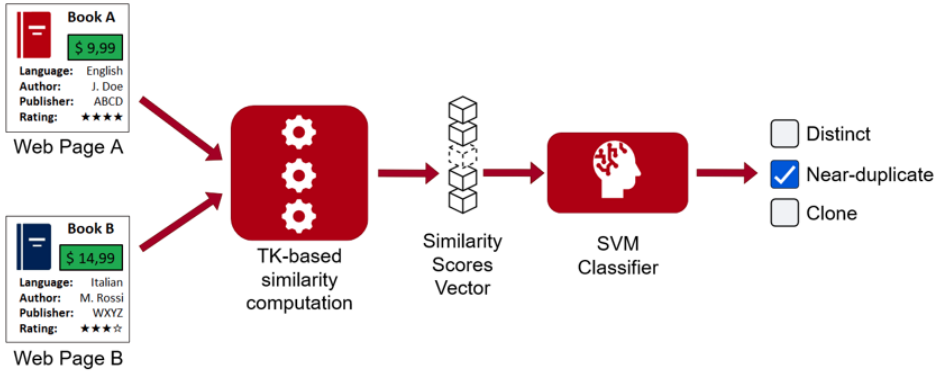
Deep Learning / Embeddings

Luigi Libero Lucio Starace
luigiliberolucio.starace@unina.it

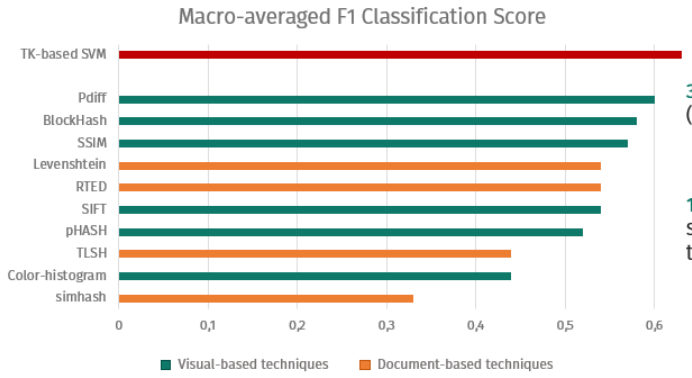
The near-duplicate problem



Tree Kernel-based near-dupes detection



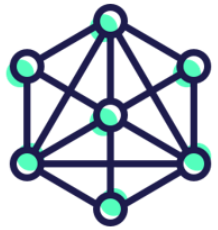
Emerging Results



3% improvement w.r.t. Pdiff (best **visual-based** technique)

10% to **30% improvement** w.r.t. similar document-based techniques

Future research



Apply to Model Inference!

Custom Tree-Kernels

Deep Learning / Embeddings