

# Computing social constructions: visual analytics for text-intensive research on immigration

Serperi Sevgür<sup>1</sup>, Mariano Maisonnave<sup>2,3</sup>, Eugena Kwon<sup>1</sup>, Evangelia Tastsoglou<sup>1</sup>, Evangelos Milios<sup>2</sup>, Ana Maguitman<sup>3</sup> and Axel Soto<sup>3</sup>

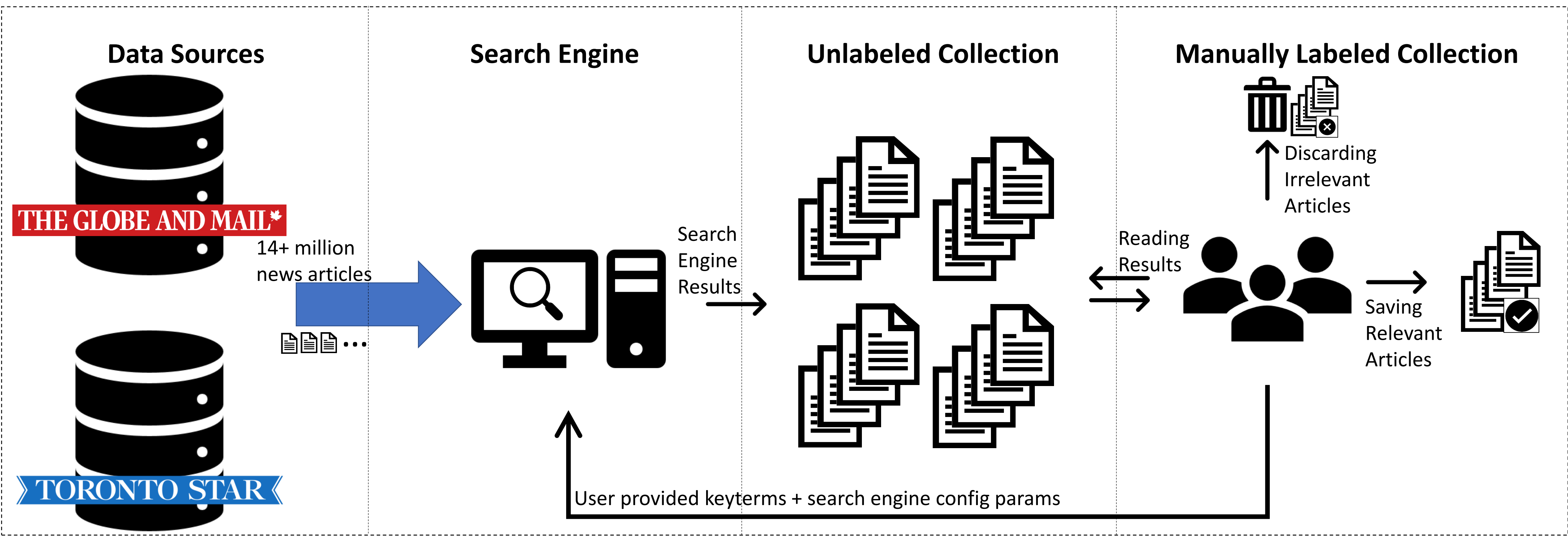
## INTRODUCTION

- MOTIVATION**  
Facilitate text-intensive research in Social Sciences.
- BACKGROUND**  
Large amounts of text need to be retrieved, organized, summarized, and thematized.
- OBJECTIVE**  
Develop Natural Language Processing (NLP) and Visual Analytics (VA) tools for social scientists to manage large amounts of text-based data.
- SIGNIFICANCE**
- Interdisciplinary approach that will move the state-of-the-art in NLP and VA.
  - Enable Social Scientists to tap into the semantics of text corpora that are orders of magnitude larger than those available in the past.

## SOCIAL SCIENCE

- SOCIAL SCIENCE RESEARCH QUESTIONS:**
1. Examine shifting representations of refugees in Canadian mainstream newspapers.
  2. Examine the representation of Displaced Persons (DP) in the media after the Second World War.
- METHOD: Content Analysis: Or Discourse Analysis:**  
To uncover Implicit assumptions, or language and power nexus. **Steps:**
1. Retrieve relevant data within a time frame
  2. Open coding
  3. Focused coding
  4. Organization of resultant themes and their analysis in relation to theoretical framework

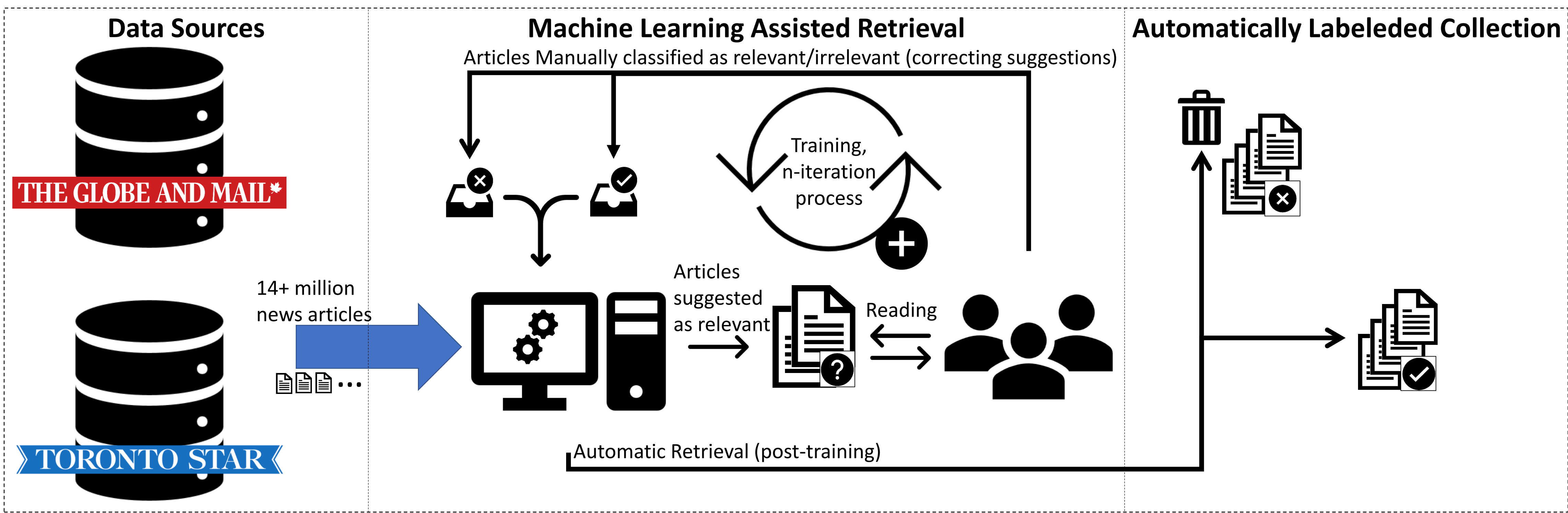
## TRADITIONAL APPROACH FOR RETRIEVAL IN TEXT-INTENSIVE RESEARCH



**Figure 1: Traditional approach.**

1. Feed keywords to a search engine, which returns a huge collection of unlabeled documents from two data sources.
2. The user reads the unlabeled documents to find the relevant ones (discarding the irrelevant ones).

## PROPOSED APPROACH FOR RETRIEVAL IN TEXT-INTENSIVE RESEARCH



**Figure 2: Proposed approach.**

1. The documents are the input to a machine learning system that creates a vector representation for each article.
2. Predictive model suggests relevant articles. User dis/approves the suggestions. Re-training improves the understanding of a relevant article.
3. The system returns a set of articles that are —with a high level of confidence— relevant to the research question.

## RESULTS

- Shortcomings of the traditional approach:**
- I. High numbers of irrelevant articles.
  - II. May not tap into all possible keywords thus limiting the scope of the research process.
  - III. Manual task (time and energy).
  - IV. No learning.

Research Question 2 Period: 1945-1967	
Traditional Approach (keywords: DP and Canada)	Our Approach
Article revised: 2,038 (522 relevant / 1,516 irrelevant) Time: 6 min per article (203.8 hs)	Articles revised: 2+ million Time: 10hs per iteration (6.3% increase in precision for our last labeling round)

- Advantages of our approach:**
- I. Document representation is semantics-based: it detects relevant documents regardless of the vocabulary used in it.
  - II. This set of relevant articles is generated automatically (built), so not bounded in size.
  - III. I & II allow us to explore new research questions and answer those that have not been previously asked.
  - IV. Frequent re-training increases relevance ratio, time efficiency and saves effort.

## FUTURE WORK

- Develop NLP/VA tools to address the coding process, i.e., organize and thematize using:
- topic modeling.
  - clustering.
  - diachronic word representations.

THE TOOL IS RESEARCH-QUESTION INDEPENDENT