



Canadian Ethnic Studies Association  
Société Canadienne d'Études Ethniques

# *Computer-Assisted Text-Intensive Social Science Research on Immigration*

*Mariano Maisonnave*





Text-intensive research in Social Science relies on the **retrieval**, **organization**, **conceptualization** and **summarization** of large amounts of text.



We propose using existing and developing novel **Computer Science (CS) tools**, to support text-intensive social science research efforts on immigration.

# TEAM



Evangelia  
Tatsoglou

Saint Mary's  
University



Mariano  
Maisonnave

Dalhousie  
University



Eunjeong  
Kwon

Trent  
University



Serperi  
Sevgür

Saint Mary's  
University



Evangelos  
Milios

Dalhousie  
University



Axel  
Soto

Universidad  
Nacional del  
Sur



Ana  
Maguitman

Universidad  
Nacional del  
Sur

# TEAM



Evangelia  
Tatsoglou

Saint Mary's  
University



Mariano  
Maisonnave

Dalhousie  
University



Eunjeong  
Kwon

Trent  
University



Serperi  
Sevgür

Saint Mary's  
University



Evangelos  
Milios

Dalhousie  
University



Axel  
Soto

Universidad  
Nacional del  
Sur



Ana  
Maguitman

Universidad  
Nacional del  
Sur

## Goal

*Contribute to the state-of-the-art in both Computer Science and Social Science.*

# TEAM



Evangelia  
Tatsoglou

Saint Mary's  
University



Mariano  
Maisonnave

Dalhousie  
University



Eunjeong  
Kwon

Trent  
University



Serperi  
Sevgür

Saint Mary's  
University



Evangelos  
Milios

Dalhousie  
University



Axel  
Soto

Universidad  
Nacional del  
Sur



Ana  
Maguitman

Universidad  
Nacional del  
Sur

## Goal

*Contribute to the state-of-the-art in both Computer Science and Social Science.*

## How?

**Enable new research** in Social Science though the **development of novel computer science tools** to assist in the steps of information retrieval, organization, conceptualization and summarization.

## Manual Information Retrieval (IR)

- Costly process.
- Requires keeping track of keywords and relevant articles. Manual process of labeling.
- Bounded by human effort. Does not scale.



## Manual Information Retrieval (IR)

- Costly process.
- Requires keeping track of keywords and relevant articles. Manual process of labeling.
- Bounded by human effort. Does not scale.

### GOAL

*To help the user find nearly all articles relevant to a given research question while minimizing manual effort.*



## Manual Information Retrieval (IR)

- Costly process.
- Requires keeping track of keywords and relevant articles. Manual process of labeling.
- Bounded by human effort. Does not scale.

### GOAL

*To help the user find nearly all articles relevant to a given research question while minimizing manual effort.*

*We are not looking for a representative set, we are looking for all the relevant articles.*



## Manual Information Retrieval (IR)

- Costly process.
- Requires keeping track of keywords and relevant articles. Manual process of labeling.
- Bounded by human effort. Does not scale.

### GOAL

*To help the user find nearly all articles relevant to a given research question while minimizing manual effort.*

*We are not looking for a representative set, we are looking for all the relevant articles.*

*We aim to develop a method independent of the research question to help the user in the IR task.*



# Cost of manual labeling

Source: The Globe and Mail

Topic Description: Displaced Persons in Canada

Search terms: "DP" and "Canada"

Period: 1945 - 1967

Size: 6946 articles

Manually inspected: 2038 articles (521 Relevant / 1517 Irrelevant)

Prevalence\* of relevant articles: 7.50 %

Time spent labeling: ~204 hours (~7 weeks working 6 hours per day)



\* In this context, we refer to prevalence as the proportion of relevant articles in the dataset (in this example,  $521/2038 = 0.075 = 7.50\%$ ).

# Background and terminology

**High recall information retrieval (HRIR).** The goal of HRIR is to find all or nearly all relevant documents for a search topic [1].

# Background and terminology

**High recall information retrieval (HRIR).** The goal of HRIR is to find all or nearly all relevant documents for a search topic [1].

**Technology assisted review (TAR)** involves the iterative retrieval and review of documents from a collection until a substantial majority (or “all”) of the relevant documents have been reviewed [2].

# Background and terminology

**High recall information retrieval (HRIR).** The goal of HRIR is to find all or nearly all relevant documents for a search topic [1].

**Technology assisted review (TAR)** involves the iterative retrieval and review of documents from a collection until a substantial majority (or “all”) of the relevant documents have been reviewed [2].

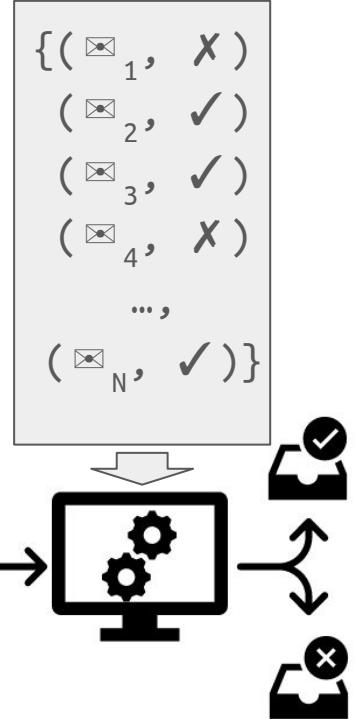
The goal of **Supervised learning algorithms** is to learn a function that maps feature vectors (inputs) to labels (output), based on example input-output pairs [4]. Classification algorithms are a type of supervised learning that we use throughout this work.

# Background and terminology

**High recall information retrieval (HRIR)**. The goal of HRIR is to find all or nearly all relevant documents for a search topic [1].

**Technology assisted review (TAR)** involves the iterative retrieval and review of documents from a collection until a substantial majority (or “all”) of the relevant documents have been reviewed [2].

The goal of **Supervised learning algorithms** is to learn a function that maps feature vectors (inputs) to labels (output), based on example input-output pairs [4]. Classification algorithms are a type of supervised learning that we use throughout this work.



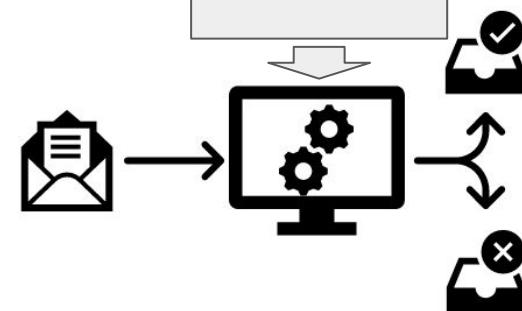
# Background and terminology

**High recall information retrieval (HRIR).** The goal of HRIR is to find all or nearly all relevant documents for a search topic [1].

**Technology assisted review (TAR)** involves the iterative retrieval and review of documents from a collection until a substantial majority (or “all”) of the relevant documents have been reviewed [2].

The goal of **Supervised learning algorithms** is to learn a function that maps feature vectors (inputs) to labels (output), based on example input-output pairs [4]. Classification algorithms are a type of supervised learning that we use throughout this work.

**Active Learning.** A supervised machine learning algorithm can achieve greater accuracy with fewer labeled training instances if it is allowed to choose the data from which it learns [3].

$$\{(\text{✉}_1, \text{✗}), (\text{✉}_2, \checkmark), (\text{✉}_3, \checkmark), (\text{✉}_4, \text{✗}), \dots, (\text{✉}_N, \checkmark)\}$$


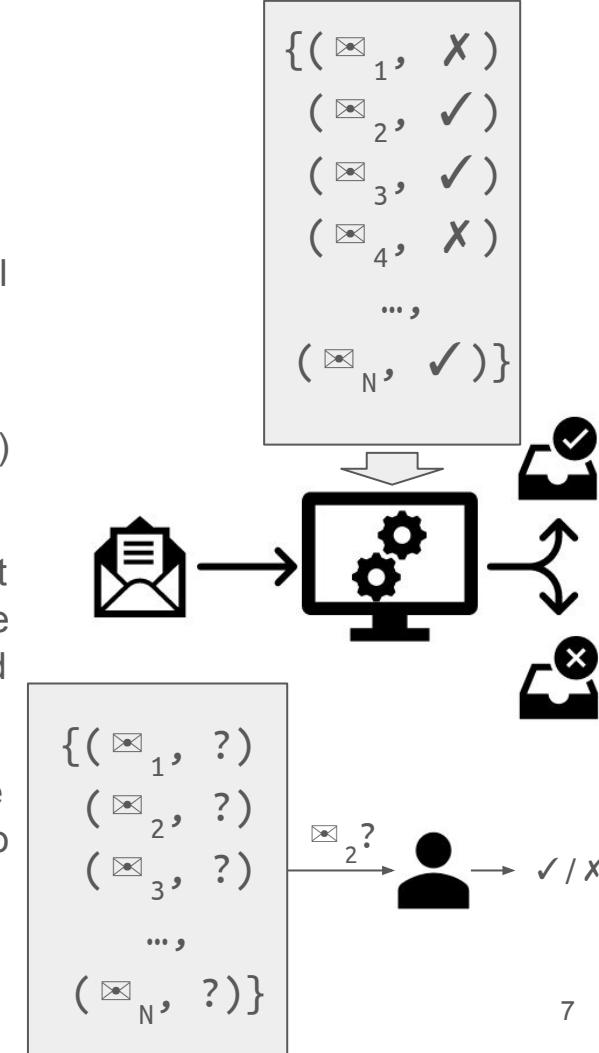
# Background and terminology

**High recall information retrieval (HRIR)**. The goal of HRIR is to find all or nearly all relevant documents for a search topic [1].

**Technology assisted review (TAR)** involves the iterative retrieval and review of documents from a collection until a substantial majority (or “all”) of the relevant documents have been reviewed [2].

The goal of **Supervised learning algorithms** is to learn a function that maps feature vectors (inputs) to labels (output), based on example input-output pairs [4]. Classification algorithms are a type of supervised learning that we use throughout this work.

**Active Learning**. A supervised machine learning algorithm can achieve greater accuracy with fewer labeled training instances if it is allowed to choose the data from which it learns [3].



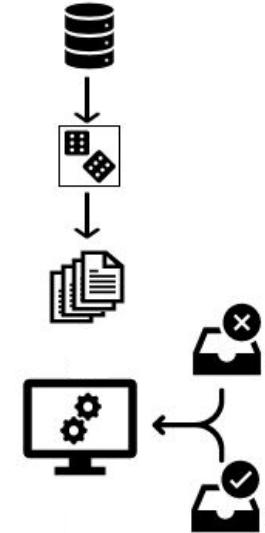
## Proposal - Scalable Continuous Active Learning (SCAL) [5]

1. Ask the user for a relevant example or a topic description.
2. Draw a large uniform random sample of size N.



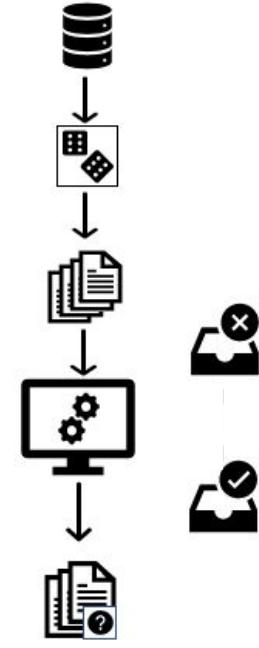
## Proposal - Scalable Continuous Active Learning (SCAL) [5]

1. Ask the user for a relevant example or a topic description.
2. Draw a large uniform random sample of size N.
  - a. Build and train a classifier using current labeled data



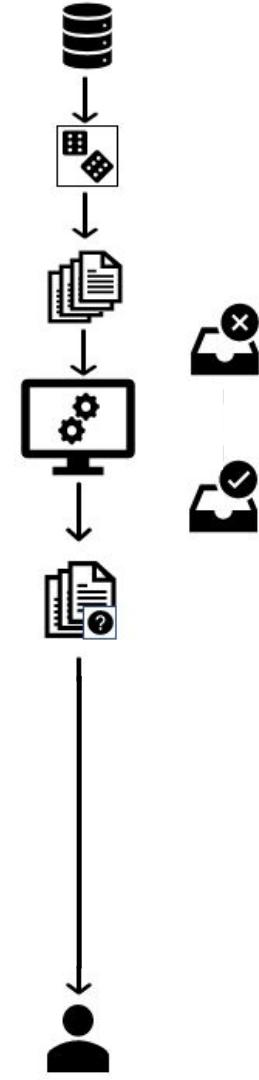
## Proposal - Scalable Continuous Active Learning (SCAL) [5]

1. Ask the user for a relevant example or a topic description.
2. Draw a large uniform random sample of size N.
  - a. Build and train a classifier using current labeled data
  - b. Select batch of highest-scoring documents



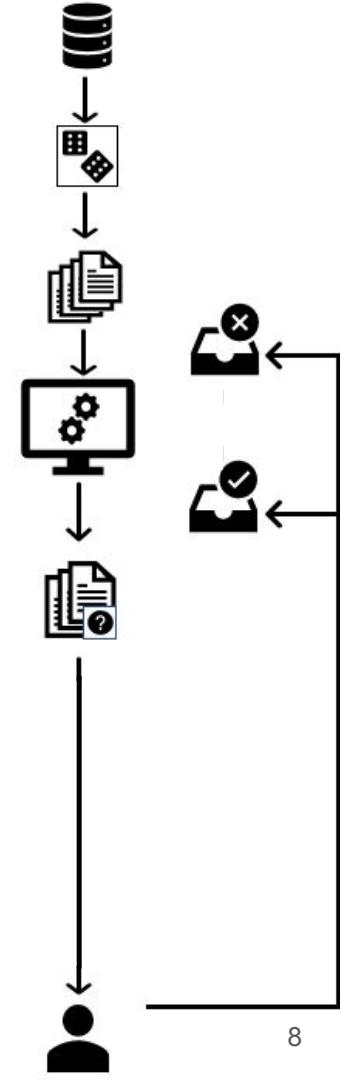
## Proposal - Scalable Continuous Active Learning (SCAL) [5]

1. Ask the user for a relevant example or a topic description.
2. Draw a large uniform random sample of size N.
  - a. Build and train a classifier using current labeled data
  - b. Select batch of highest-scoring documents
  - c. The user reviews and labels the subsample



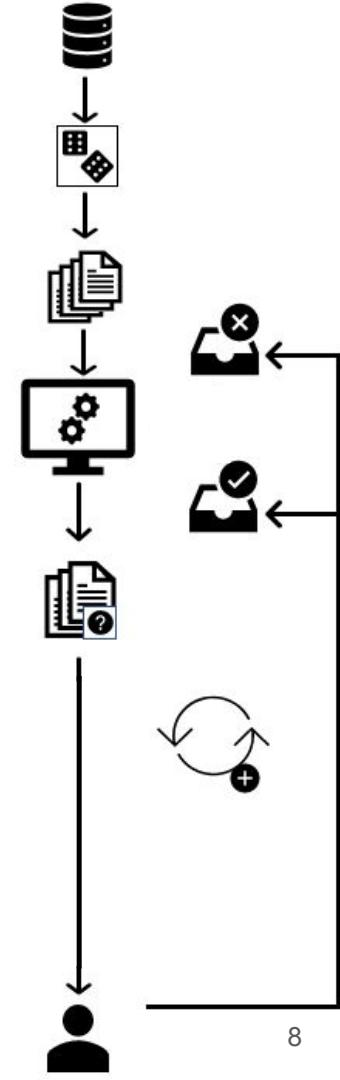
## Proposal - Scalable Continuous Active Learning (SCAL) [5]

1. Ask the user for a relevant example or a topic description.
2. Draw a large uniform random sample of size N.
  - a. Build and train a classifier using current labeled data
  - b. Select batch of highest-scoring documents
  - c. The user reviews and labels the subsample



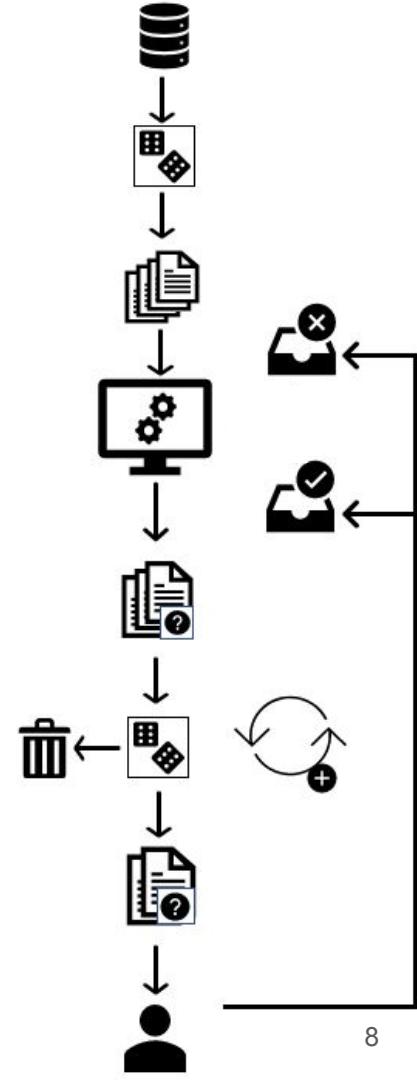
## Proposal - Scalable Continuous Active Learning (SCAL) [5]

1. Ask the user for a relevant example or a topic description.
2. Draw a large uniform random sample of size N.
3. Repeat until random sample is empty
  - a. Build and train a classifier using current labeled data
  - b. Select batch of highest-scoring documents
  - c. The user reviews and labels the subsample



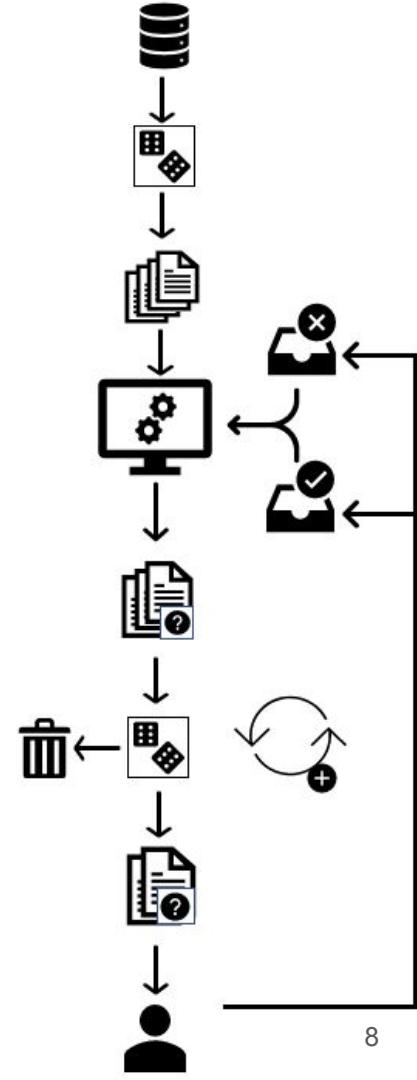
## Proposal - Scalable Continuous Active Learning (SCAL) [5]

1. Ask the user for a relevant example or a topic description.
2. Draw a large uniform random sample of size N.
3. Repeat until random sample is empty
  - a. Build and train a classifier using current labeled data
  - b. Select batch of highest-scoring documents
  - c. **Draw a random subsample**
  - d. The user reviews and labels the subsample



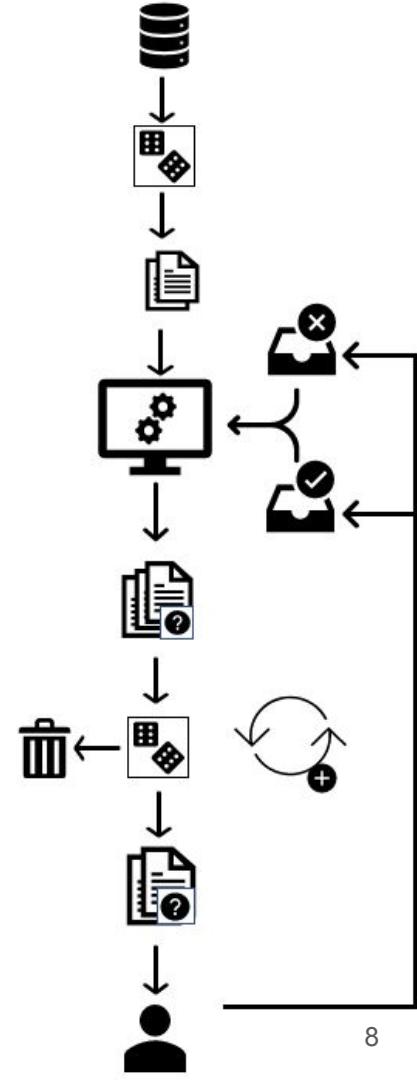
## Proposal - Scalable Continuous Active Learning (SCAL) [5]

1. Ask the user for a relevant example or a topic description.
2. Draw a large uniform random sample of size N.
3. Repeat until random sample is empty
  - a. Build and train a classifier using current labeled data
  - b. Select batch of highest-scoring documents
  - c. Draw a random subsample
  - d. The user reviews and labels the subsample



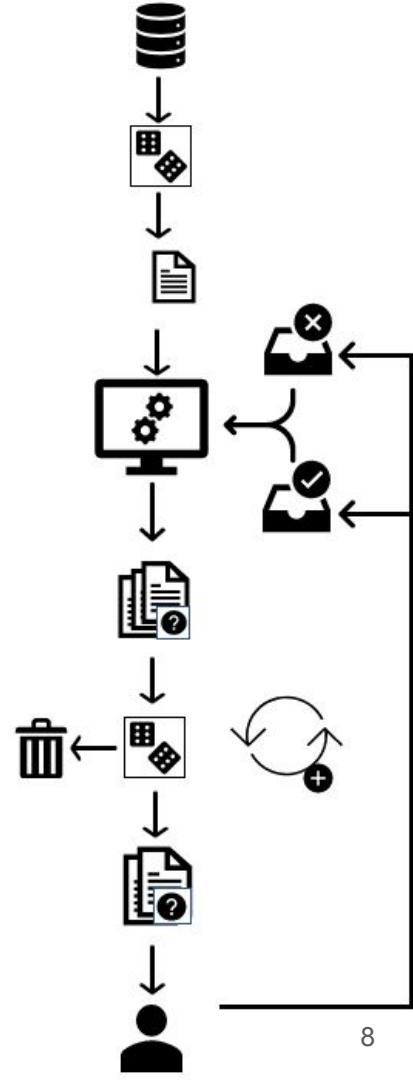
## Proposal - Scalable Continuous Active Learning (SCAL) [5]

1. Ask the user for a relevant example or a topic description.
2. Draw a large uniform random sample of size N.
3. Repeat until random sample is empty
  - a. Build and train a classifier using current labeled data
  - b. Select batch of highest-scoring documents
  - c. Draw a random subsample
  - d. The user reviews and labels the subsample



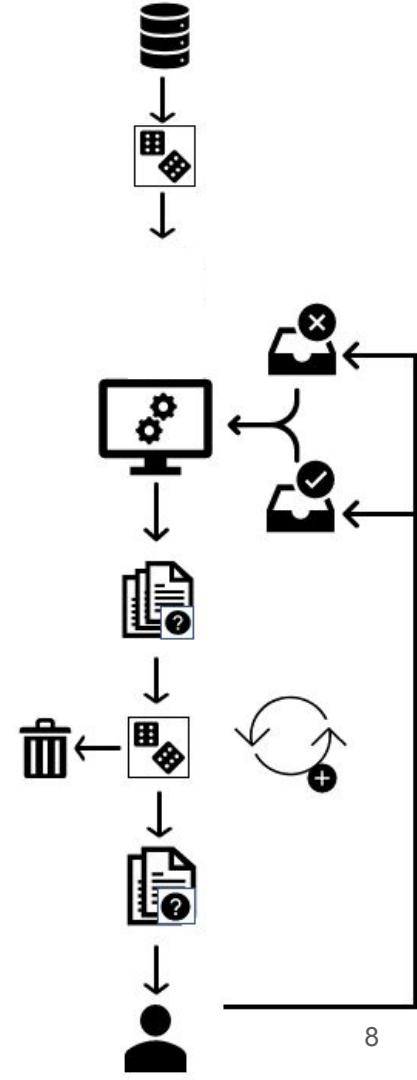
## Proposal - Scalable Continuous Active Learning (SCAL) [5]

1. Ask the user for a relevant example or a topic description.
2. Draw a large uniform random sample of size N.
3. Repeat until random sample is empty
  - a. Build and train a classifier using current labeled data
  - b. Select batch of highest-scoring documents
  - c. Draw a random subsample
  - d. The user reviews and labels the subsample



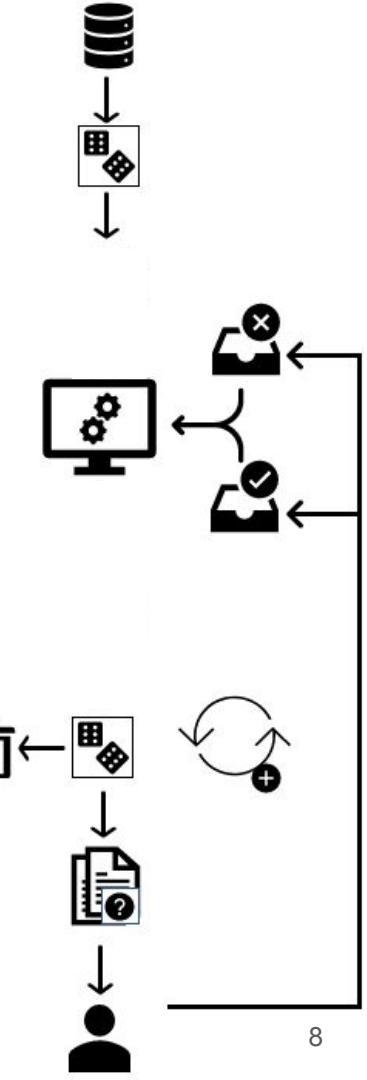
## Proposal - Scalable Continuous Active Learning (SCAL) [5]

1. Ask the user for a relevant example or a topic description.
2. Draw a large uniform random sample of size N.
3. Repeat until random sample is empty
  - a. Build and train a classifier using current labeled data
  - b. Select batch of highest-scoring documents
  - c. Draw a random subsample
  - d. The user reviews and labels the subsample



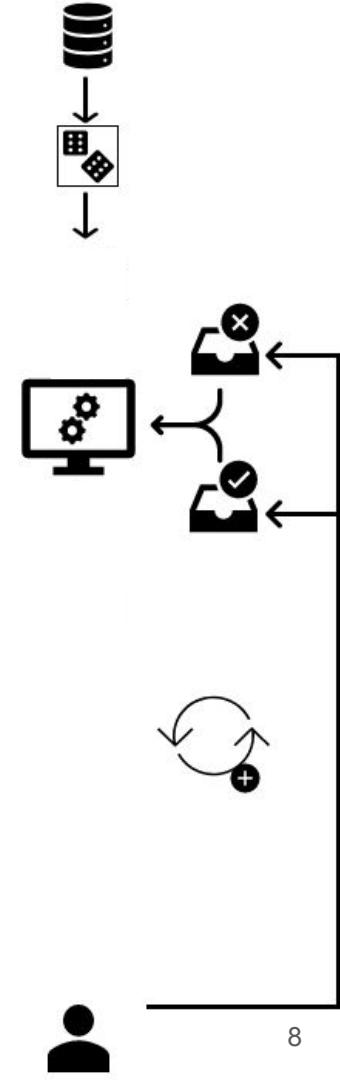
## Proposal - Scalable Continuous Active Learning (SCAL) [5]

1. Ask the user for a relevant example or a topic description.
2. Draw a large uniform random sample of size N.
3. Repeat until random sample is empty
  - a. Build and train a classifier using current labeled data
  - b. Select batch of highest-scoring documents
  - c. Draw a random subsample
  - d. The user reviews and labels the subsample



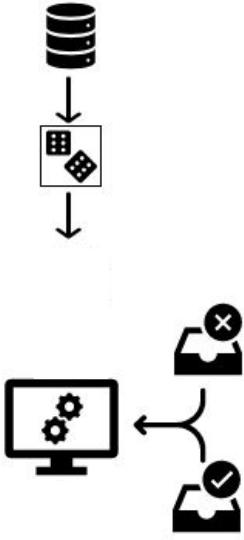
## Proposal - Scalable Continuous Active Learning (SCAL) [5]

1. Ask the user for a relevant example or a topic description.
2. Draw a large uniform random sample of size N.
3. Repeat until random sample is empty
  - a. Build and train a classifier using current labeled data
  - b. Select batch of highest-scoring documents
  - c. Draw a random subsample
  - d. The user reviews and labels the subsample



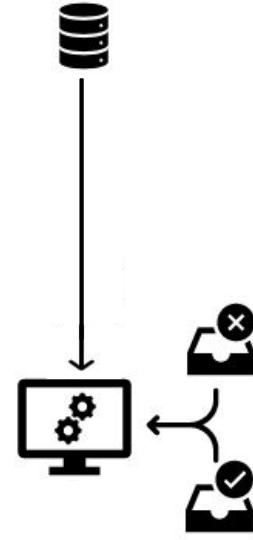
# Proposal - Scalable Continuous Active Learning (SCAL) [5]

1. Ask the user for a relevant example or a topic description.
2. Draw a large uniform random sample of size N.
3. Repeat until random sample is empty
  - a. Build and train a classifier using current labeled data
  - b. Select batch of highest-scoring documents
  - c. Draw a random subsample
  - d. The user reviews and labels the subsample
4. Train the final classifier



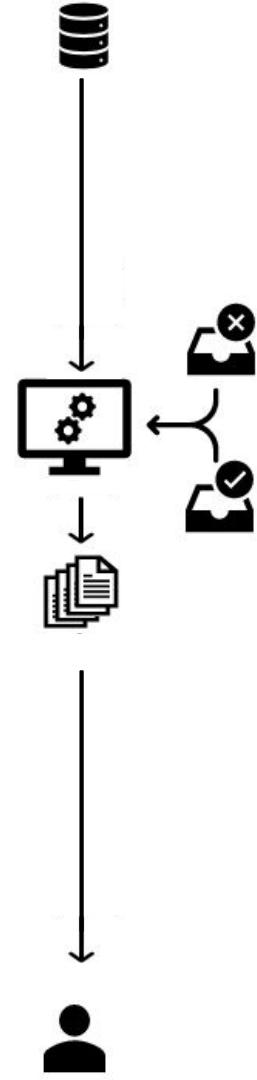
## Proposal - Scalable Continuous Active Learning (SCAL) [5]

1. Ask the user for a relevant example or a topic description.
2. Draw a large uniform random sample of size N.
3. Repeat until random sample is empty
  - a. Build and train a classifier using current labeled data
  - b. Select batch of highest-scoring documents
  - c. Draw a random subsample
  - d. The user reviews and labels the subsample
4. Train the final classifier



## Proposal - Scalable Continuous Active Learning (SCAL) [5]

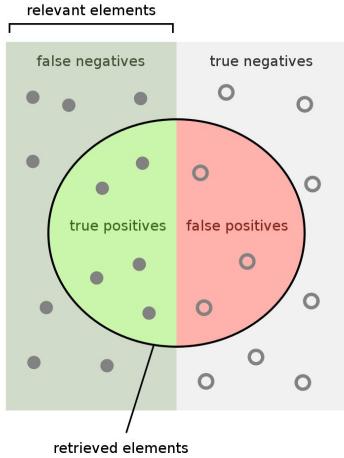
1. Ask the user for a relevant example or a topic description.
2. Draw a large uniform random sample of size N.
3. Repeat until random sample is empty
  - a. Build and train a classifier using current labeled data
  - b. Select batch of highest-scoring documents
  - c. Draw a random subsample
  - d. The user reviews and labels the subsample
4. Train the final classifier
5. Use the final classifier to generate predictions for the entire corpus



# SIMULATIONS

- *Using the "Displaced Person in Canada" dataset (which the relevant/irrelevant labels), and*
- *Using an Oracle that mimics the user behaviour providing the correct labels.*

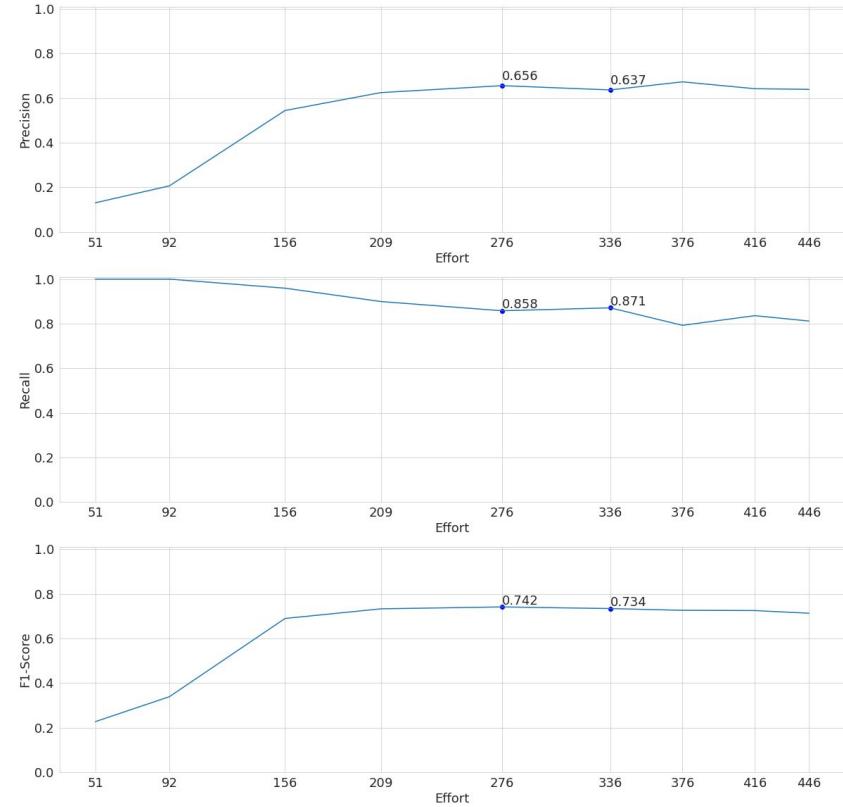
# Simulation Results



$$\text{Precision} = \frac{\text{How many relevant items are retrieved?}}{\text{How many retrieved items are relevant?}}$$
$$\text{Recall} = \frac{\text{How many relevant items are retrieved?}}{\text{How many relevant elements?}}$$

- More labeling budget (effort) means better performance.
- After labeling 150 examples, the performance starts to plateau.
- The method guarantees high recall, and the precision is above 0.6.

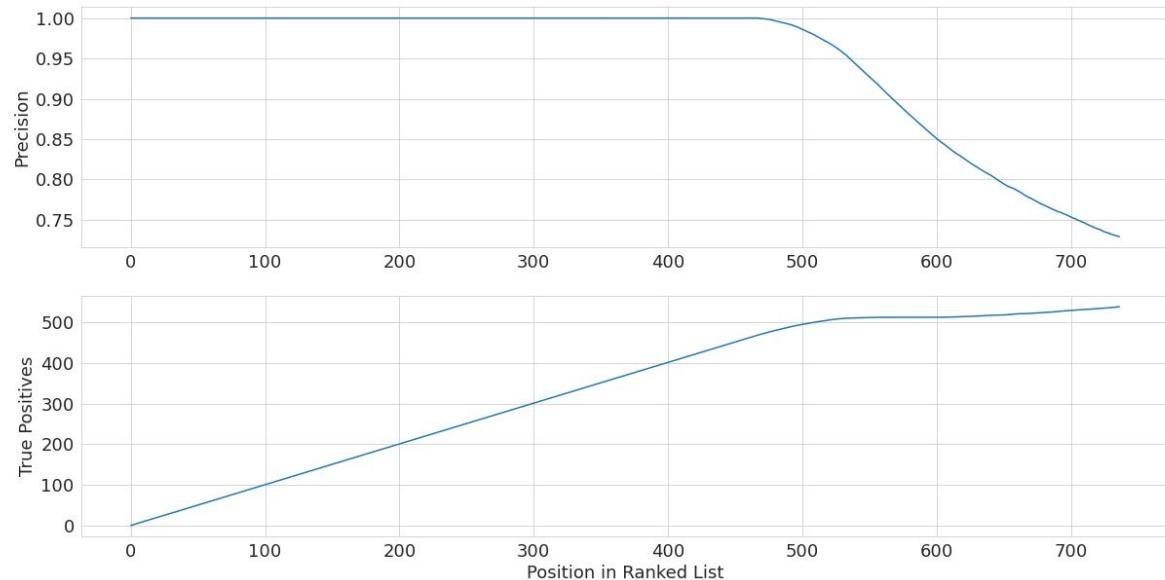
[https://en.wikipedia.org/wiki/Precision\\_and\\_recall](https://en.wikipedia.org/wiki/Precision_and_recall)



# Simulation Results @ effort=276

- The method suggested a little more than 700 articles as relevant, but there are a little more than 500 relevant only.

- Almost all the relevant articles are at the beginning of the list.



# REAL CASE SCENARIO

*New research question with the human in the loop and no prior information about relevant/irrelevant articles*

# Dataset

**Source:** The Globe and Mail

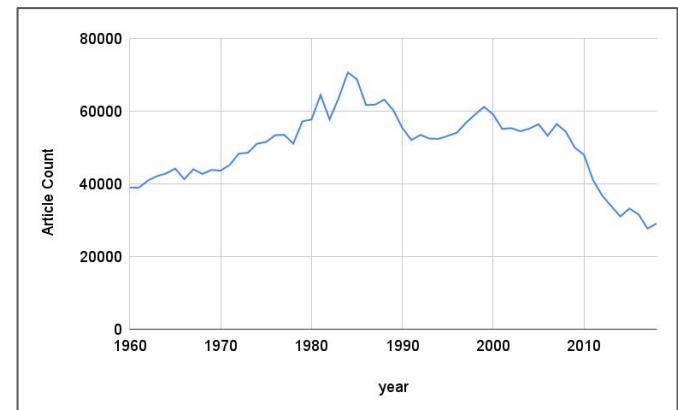
**Topic Description:** Multiculturalism in Canada

**Period:** January 1st, 1960 to December 31st, 2018.

**Size:** 5,502,653 articles

**Filter-out** articles that do not mention Canada or a Canadian province ("Canada", "Canadian", "Nova Scotia", "Nova Scotian", "NS", etc.).

**Size after filtering:** 2,961,906 articles



2,961,906 articles



# Results first step

## Input

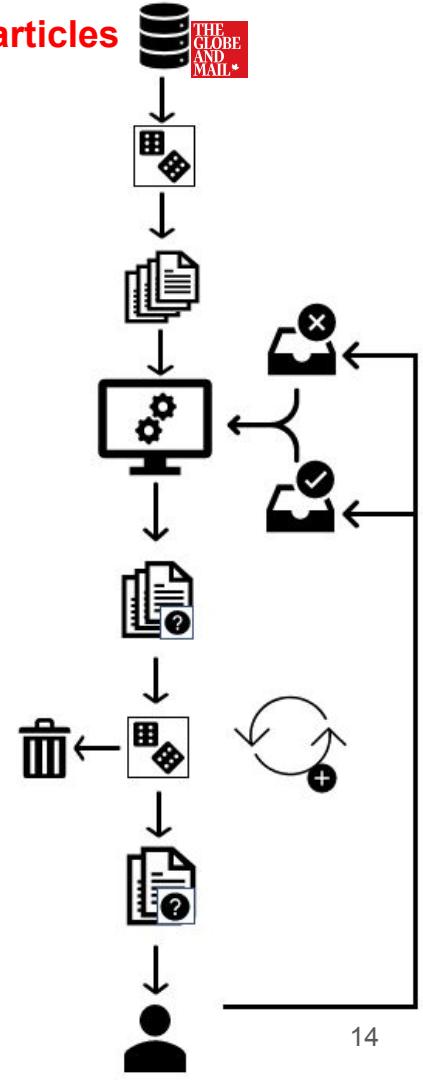
Sample size=30,000

Labeling cap=5

Iterations required=65 (1,2,3,4,5,5,5,...,5)

Labeling budget=315

Target recall=80%



2,961,906 articles



# Results first step

## Input

Sample size=30,000

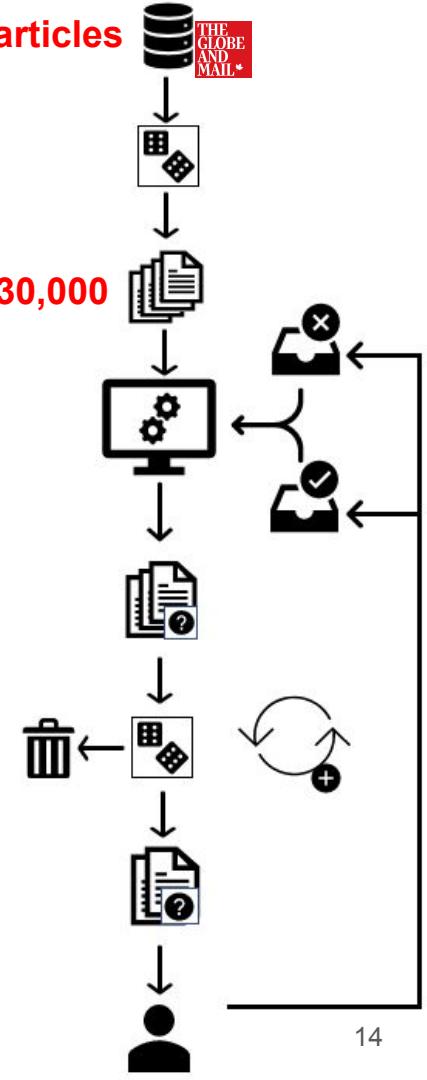
Labeling cap=5

Iterations required=65 (1,2,3,4,5,5,5,...,5)

Labeling budget=315

Target recall=80%

N=30,000



2,961,906 articles



# Results first step

## Input

Sample size=30,000

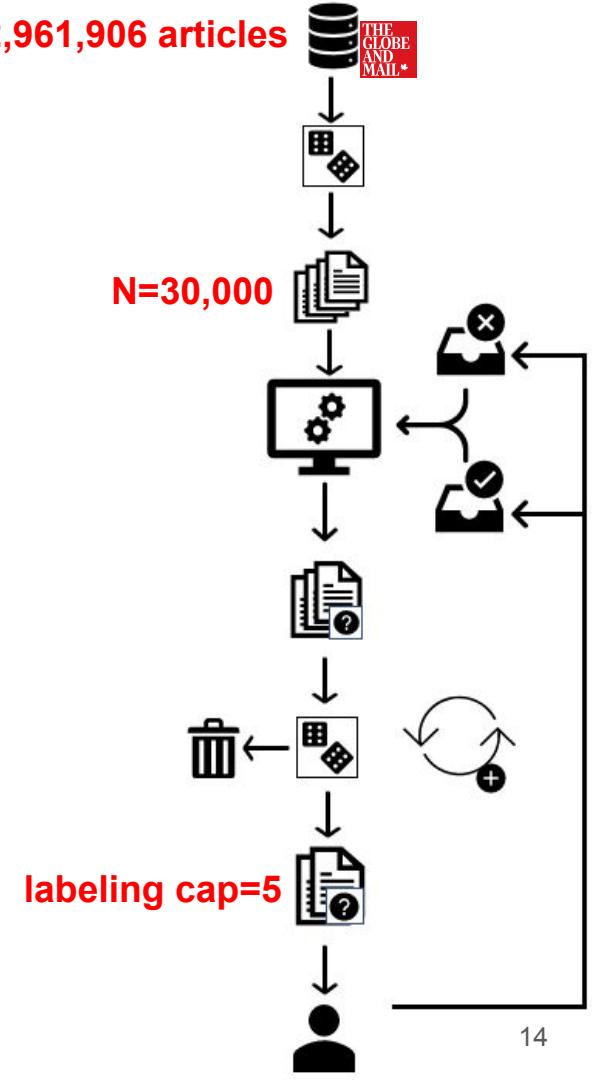
Labeling cap=5

Iterations required=65 (1,2,3,4,5,5,5,...,5)

Labeling budget=315

Target recall=80%

N=30,000



labeling cap=5

**2,961,906 articles**



# Results first step

## Input

Sample size=30,000

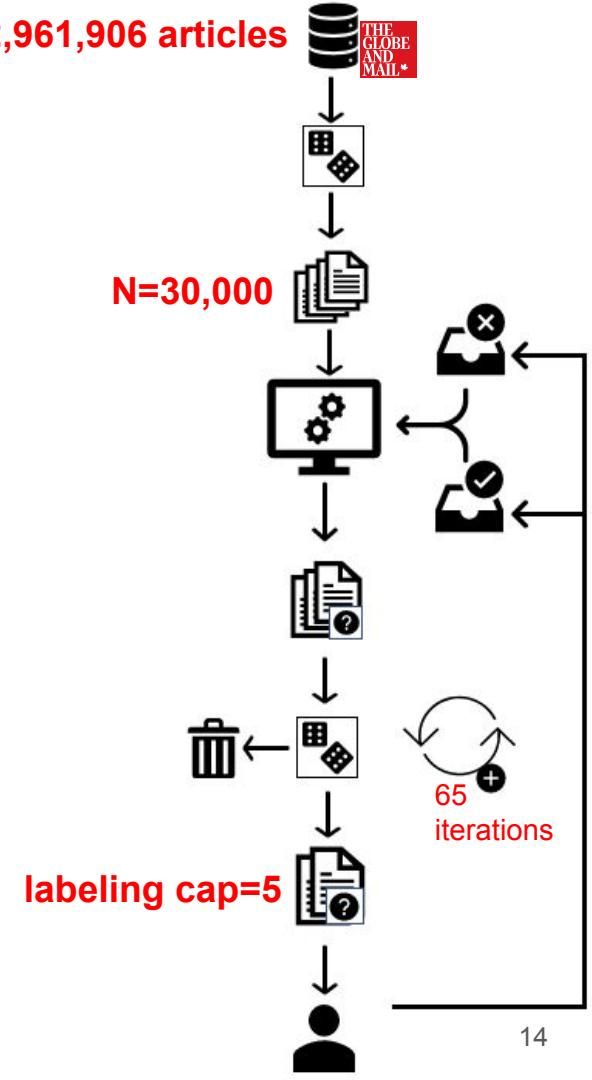
Labeling cap=5

Iterations required=65 (1,2,3,4,5,5,5,...,5)

Labeling budget=315

Target recall=80%

**N=30,000**



# Results first step

## Input

Sample size=30,000

Labeling cap=5

Iterations required=65 (1,2,3,4,5,5,5,...,5)

Labeling budget=315

Target recall=80%

2,961,906 articles



N=30,000

303 articles

Logistic Regression  
with TF-IDF matrix

12 articles

65  
iterations

labeling cap=5

2,961,906 articles



# Results first step

## Input

Sample size=30,000

Labeling cap=5

Iterations required=65 (1,2,3,4,5,5,5,...,5)

Labeling budget=315

Target recall=80%

## Output

Relevant articles manually labeled= 12 articles

Estimated relevant articles in sample= 49 articles

The estimated prevalence of relevant articles= 0.17%

N=30,000

303 articles

Logistic Regression  
with TF-IDF matrix

12 articles

65  
iterations

labeling cap=5

# Results first step

## Input

Sample size=30,000

Labeling cap=5

Iterations required=65 (1,2,3,4,5,5,5,...,5)

Labeling budget=315

Target recall=80%

## Output

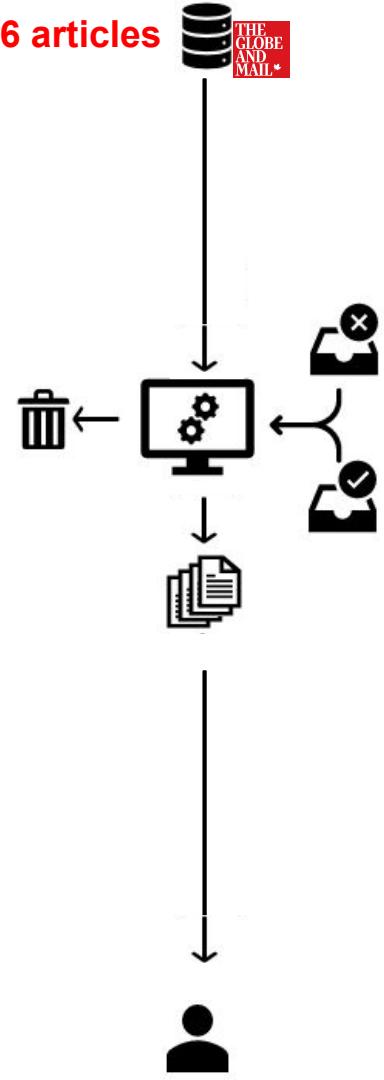
Relevant articles manually labeled= 12 articles

Estimated relevant articles in sample= 49 articles

**The estimated prevalence of relevant articles= 0.17%**

Estimated relevant articles in the whole corpus≈ 5,035 articles

Suggestions made= 17,014



2,961,906 articles



# Results first step

## Input

Sample size=30,000

Labeling cap=5

Iterations required=65 (1,2,3,4,5,5,5,...,5)

Labeling budget=315

Target recall=80%

## Output

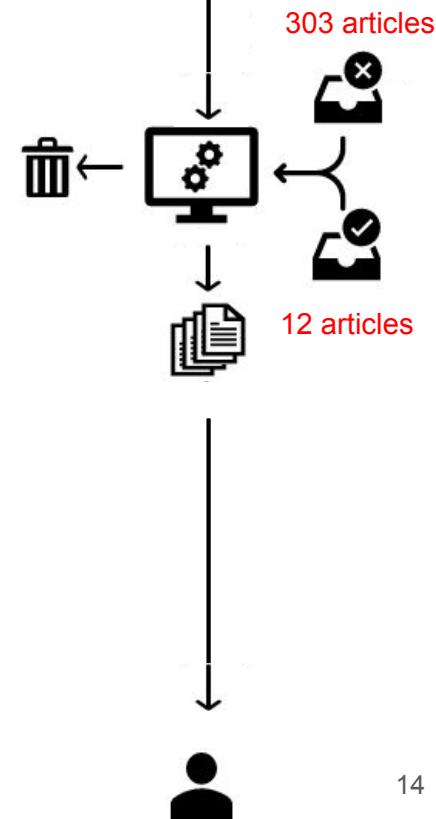
Relevant articles manually labeled= 12 articles

Estimated relevant articles in sample= 49 articles

**The estimated prevalence of relevant articles= 0.17%**

Estimated relevant articles in the whole corpus≈ 5,035 articles

Suggestions made= 17,014



2,961,906 articles



# Results first step

## Input

Sample size=30,000

Labeling cap=5

Iterations required=65 (1,2,3,4,5,5,5,...,5)

Labeling budget=315

Target recall=80%

## Output

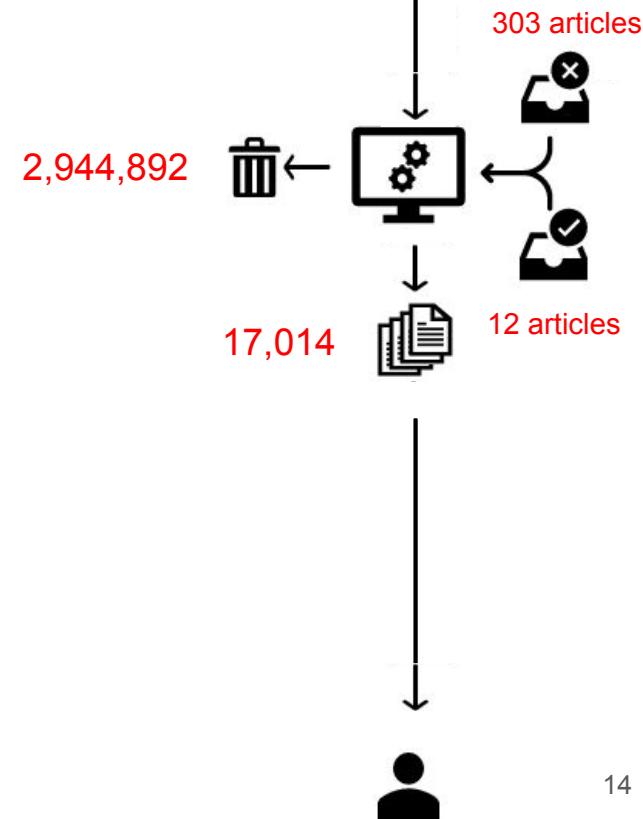
Relevant articles manually labeled= 12 articles

Estimated relevant articles in sample= 49 articles

**The estimated prevalence of relevant articles= 0.17%**

Estimated relevant articles in the whole corpus≈ 5,035 articles

Suggestions made= 17,014



2,961,906 articles



# Results first step

## Input

Sample size=30,000

Labeling cap=5

Iterations required=65 (1,2,3,4,5,5,5,...,5)

Labeling budget=315

Target recall=80%

## Output

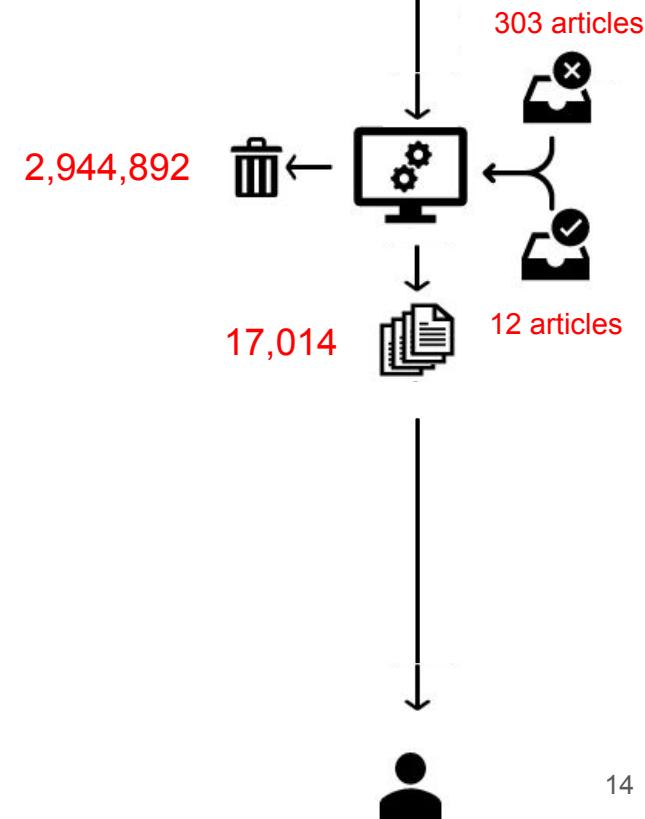
Relevant articles manually labeled= 12 articles

Estimated relevant articles in sample= 49 articles

**The estimated prevalence of relevant articles= 0.17%**

Estimated relevant articles in the whole corpus≈ 5,035 articles

Suggestions made= 17,014



## Timing

Time per iteration: ~32 minutes

Total time: ~35 hours

# Results first step

## Input

Sample size=30,000

Labeling cap=5

Iterations required=65 (1,2,3,4,5,5,5,...,5)

Labeling budget=315

Target recall=80%

## Output

Relevant articles manually labeled= 12 articles

Estimated relevant articles in sample= 49 articles

**The estimated prevalence of relevant articles= 0.17%**

Estimated relevant articles in the whole corpus≈ 5,035 articles

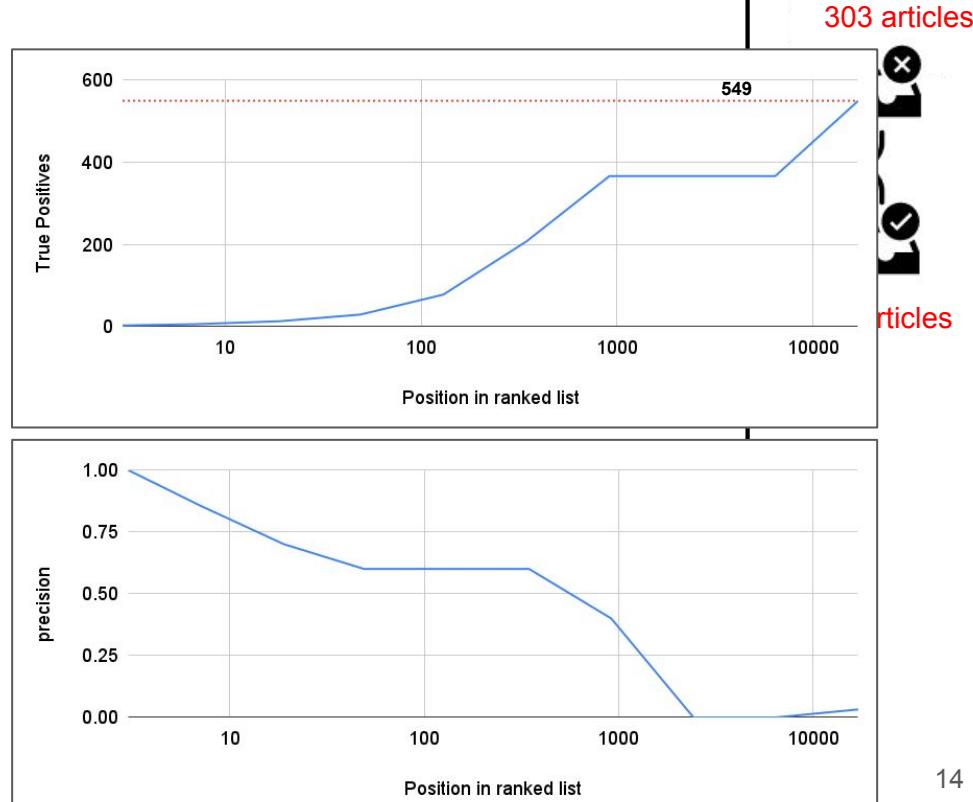
Suggestions made= 17,014

**Estimated relevant after evaluation≈ 549**

## Timing

Time per iteration: ~32 minutes

Total time: ~35 hours



# Conclusions

## Negative aspects

- There is a discrepancy in performance between simulation and real world-scenario. Main difference? Prevalence.

## Positive aspects

- If the only problem is the prevalence of relevant articles, then it can be solved by performing a second iteration of SCAL.
- The proposed approach is still significantly faster than the manual process and scales better.
- The proposed solution can be applied to other research problems and other periods.
- The ranking works.

# Next Steps

# TRIVIR: A Visualization System to Support Document Retrieval with High Recall

Amanda Gonçalves Dias  
gdias.amanda@gmail.com  
University of São Paulo - ICMC  
São Carlos, Brazil

Evangelos E. Milios  
eem@cs.dal.ca  
Dalhousie University  
Halifax, Canada

Maria Cristina Ferreira de Oliveira  
cristina@icmc.usp.br  
University of São Paulo - ICMC  
São Carlos, Brazil

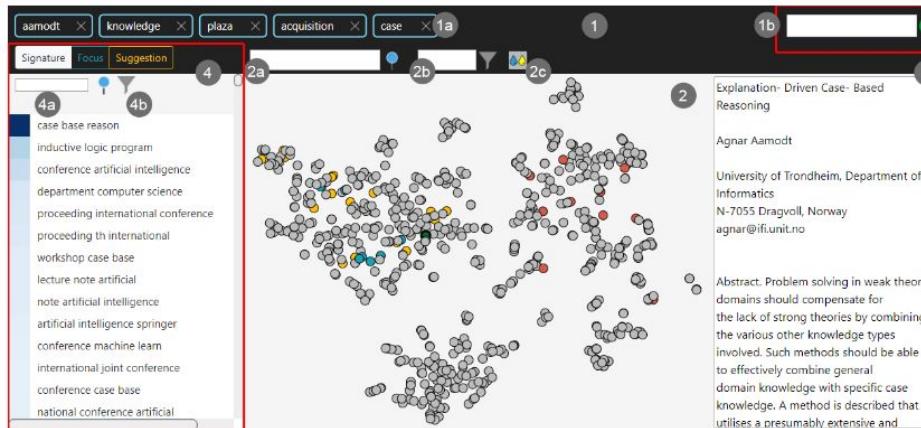


Figure 1: TRIVIR interface. The Scatterplot view in area (2) shows a similarity map depicting a collection of 675 papers in Computer Science (described by title, abstract, author, affiliation and references). The circle colors indicate the current query document (green), and then: the relevant (blue), not relevant (red), suggested as relevant (yellow), and yet unlabeled documents (gray). A user may select documents that include a given term (2a); filter the K documents most similar to the query (2b); and reduce point clutter by displaying only the relevant and suggested documents (2c). Clicking on a circle, the document's content is displayed in the Document view (3) (query document shown). The Terms view (1) shows important terms from the query document, where the user can remove (1a) or add (1b) terms. Area (4) shows the Signature List view, which shows a rank of relevant 3-grams in the corpus. It can be switched with the Focus List or the Suggestion List views by selecting the corresponding buttons. The user can select 3-grams with a particular term (4a) or with terms from the Terms view (4b).

Dias, A.G., Milios, E.E. and de Oliveira, M.C.F., 2019, September. **Trivir: A visualization system to support document retrieval with high recall.** In *Proceedings of the ACM Symposium on Document Engineering 2019* (pp. 1-10).

<https://github.com/amandagdias/TRIVIR>

# RElevant information through Classification in an Active Learning Loop (**RECALL**)

Rocío Hubert

Universidad

Nacional del Sur

Evangelos Miliotis

Dalhousie

University

Refining results

The model will give you some documents, classify them as relevant or irrelevant. This will help to achieve better results :)

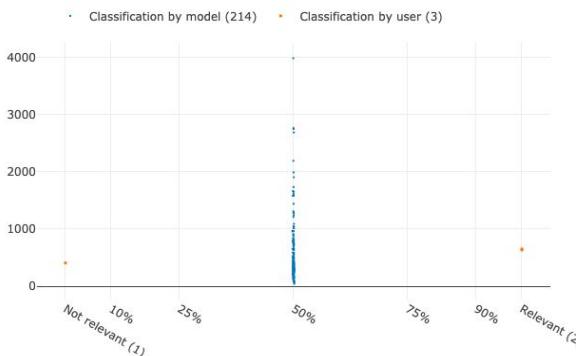
Explore items with 50% to 60% relevance    Show plot

Documents to classify  
that are 50-60% relevants

Items per page: 4 1 – 4 of 90 < >

Document Description	Relevance	Action
Productora de Buffy apoyó y respaldó declaraciones en contra del creador de la serie Joss Whedon (298)	 Relevant	<input type="checkbox"/> Update doc info
La tortura del creador de 'Buffy' a Charisma Carpenter. «Me llamó gorda cuando estaba embarazada y pesaba 57 kilos» (569)	 Relevant	<input type="checkbox"/> Update doc info
Actriz de 'Buffy' reacciona ante acusaciones de abusos de poder por parte del creador de la serie (345)	 Relevant	<input type="checkbox"/> Update doc info
El infierno tras las cámaras de 'Buffy, Cazavampiros': Nuevas acusaciones contra Joss Whedon por su trato "tóxico, hostil y abusivo" (1014)	 Relevant	<input type="checkbox"/> Update doc info

Classification by model (214) • Classification by user (3)



SEE MODEL CLASSIFICATION OF CLASSIFIED DOCUMENTS

COPY RELEVANTS DOCUMENTS COPY NOT-RELEVANTS DOCUMENTS

COPY DOCS TO CLIPBOARD UPDATE CLASSIFICATION

# Interactive Document Clustering Revisited: A Visual Analytics Approach

Ehsan Sherkat  
Dalhousie University  
Halifax, Canada  
ehsansherkat@dal.ca

Seyednaser Nourashrafeddin  
Dalhousie University  
Halifax, Canada  
nourashr@cs.dal.ca

Evangelos E. Milios  
Dalhousie University  
Halifax, Canada  
eem@cs.dal.ca

Rosane Minghim  
Universidade de São Paulo  
São Paulo, Brazil  
rminghim@icmc.usp.br

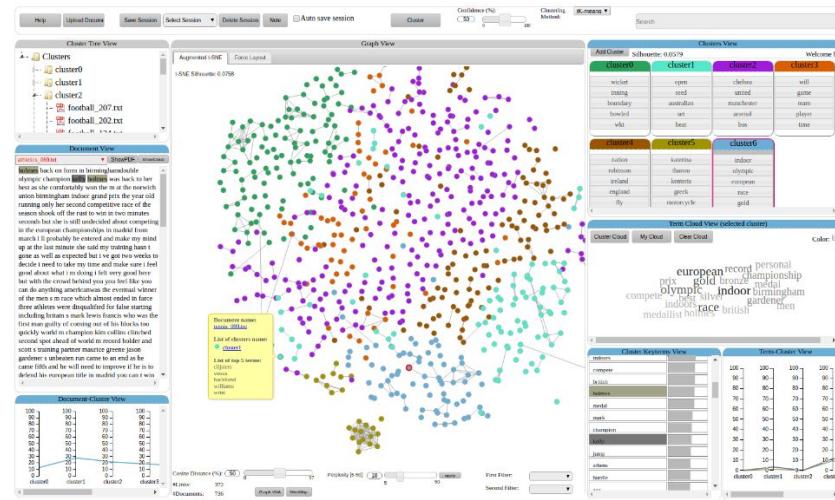


Figure 2. The visual interface of the proposed system. In the middle, the projection of 737 documents of the BBC Sport dataset is depicted (*Graph view*). On the left, we see the *Cluster tree view* for a hierarchical display of clusters and documents, the *Document view* for showing the plain text of documents, and the *Document-cluster view* to depict the relatedness of the selected document to each cluster. The name of each visual component is given in its header. On the right, we see the *Clusters view*, which demonstrates top terms of clusters, the *Term cloud view* for highlighting top terms of a selected cluster or set of documents, the *Cluster key-terms view* for listing top terms of a selected cluster with their level of importance (bar charts), and beside it the *Term-cluster view* to depict relatedness of a selected term(s) in *Cluster key-terms view* to each cluster. The views with colored header are all related to the selected cluster in the *Clusters view*. The selected cluster has a red margin and the same header color. The user can add, remove or recolor a cluster or merge two clusters in *Clusters view*. The feedback to the clustering process by changing the number of clusters or adding/removing terms in *Clusters view*, as well as adding or removing cluster. The user can send changes made by pressing the *cluster* button on the top of the *Graph view*.

Sherkat, E., Nourashrafeddin, S., Milios, E.E. and Minghim, R., 2018, March. Interactive document clustering revisited: A visual analytics approach. In *23rd International Conference on Intelligent User Interfaces* (pp. 281-292).

<https://github.com/ehsansherkat/IDC>

# Interactive clustering and high-recall information retrieval using language models

Sima Rezaeipourfarsangi

Faculty of Computer Science, Dalhousie University  
Halifax, Canada  
sima.rezaei@dal.ca

Ningyuan Pei

Department of Computing Science, University of Alberta  
Edmonton, Canada  
ningyuan@ualberta.ca

Ehsan Sherkat

Faculty of Computer Science, Dalhousie University  
Halifax, Canada  
ehsansherkat@gmail.com

Evangelos Milios

Faculty of Computer Science, Dalhousie University  
Halifax, Canada  
eem@cs.dal.ca

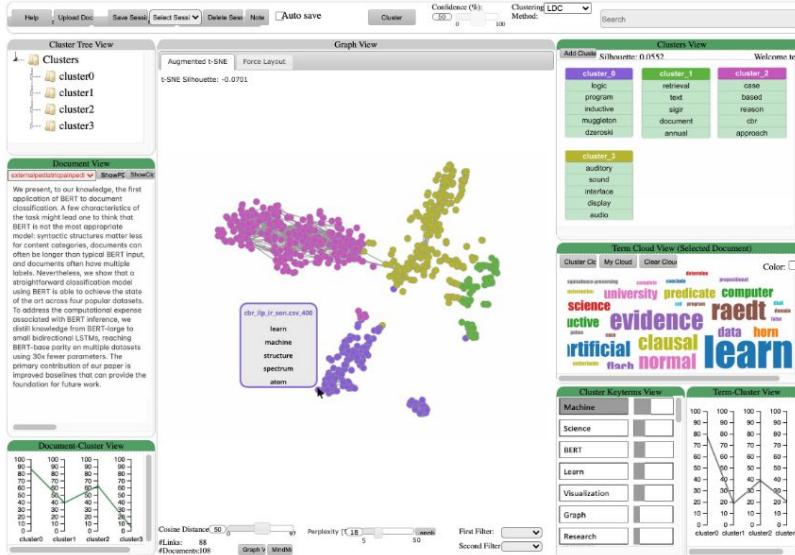


Figure 1: Overall view of the interactive visualization interface. Showing the expert interaction with 675 documents related to computer science. In the figure, Document view: displays the plain text of the focused document; Document-cluster view: shows the relatedness of the selected documents to each other; Term cloud view: displays the word cloud of the given selection (document selection, cluster, or focused document); Graph view: depicts the projection of the documents using different techniques (T-SNE or force-layout graph), a document's title and top key-terms are shown on hovering the corresponding data point; Clusters view: shows the top key-terms of each cluster inside the clustering boxes, where the user may provide further feedback about each clustering; Cluster key-terms view: lists top terms of the selected cluster and depicts their level of importance in the bar charts; Term-cluster view: showing the relatedness of the selected term(s) in the Cluster key-terms view to each cluster.

# Other resources

## Topic Modeling

- Rajendran, M., 2021. TopVis: Visual Text Analytics for Deep Topic Modeling of Reddit Data. <http://hdl.handle.net/10222/80959>

## Summarization

- Ramirez-Orta, J. and Milios, E., 2021, June. Unsupervised document summarization using pre-trained sentence embeddings and graph centrality. In *Proceedings of the Second Workshop on Scholarly Document Processing* (pp. 110-115).

[https://github.com/jarobyte91/auto\\_summ](https://github.com/jarobyte91/auto_summ)

## OCR Correction

- Ramirez-Orta, J.A., Xamena, E., Maguitman, A., Milios, E. and Soto, A.J., 2022, June. Post-OCR Document Correction with Large Ensembles of Character Sequence-to-Sequence Models. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 10, pp. 11192-11199).

[https://github.com/jarobyte91/post\\_ocr\\_correction](https://github.com/jarobyte91/post_ocr_correction)

Thank you!  
Questions?

*We gratefully acknowledge the financial support of New Frontiers in Research Fund (NFRF) (grant # NFRFE-2020-00996).*

[mariano.maisonnave@dal.ca](mailto:mariano.maisonnave@dal.ca)

# References

- [1] Abualsaud, M., Ghelani, N., Zhang, H., Smucker, M.D., Cormack, G.V. and Grossman, M.R., 2018, June. A system for efficient high-recall retrieval. In *The 41st international ACM SIGIR conference on research & development in information retrieval* (pp. 1317-1320).
- [2] Cormack, G.V. and Grossman, M.R., 2015. Autonomy and reliability of continuous active learning for technology-assisted review. arXiv preprint arXiv:1504.06868.
- [3] Settles, B., 2009. Active learning literature survey.
- [4] Russell, S.J., 2010. *Artificial intelligence a modern approach*. Pearson Education, Inc.
- [5] Cormack, G.V. and Grossman, M.R., 2016, October. Scalability of continuous active learning for reliable high-recall text classification. In *Proceedings of the 25th ACM international conference on information and knowledge management* (pp. 1039-1048).